

Evaluation of Education Projects: the Case of the Perry Preschool Program

EDWARD M. GRAMLICH

Department of Economics, The University of Michigan, Ann Arbor, MI 48104, U.S.A.

Abstract — The paper reviews the educational and evaluation experience with the widely known Perry Preschool Program. Basically, the Perry Program generated short-term IQ gains that evaporated within 2 years, followed by long-term gains in attendance, academic achievement, graduation rates and earnings, and reductions in transfer payments and pregnancies. When subjected to a formal benefit-cost analysis, the program returns about \$24,000 per student, with as much as 80% of the benefits going not to participants but to those who benefit from higher participant tax payments, lower transfers and reduced crime. It is at the same time mysterious that a program with such fleeting IQ gains could generate such enormous long-term gains, impressive that preschool programs provide such an attractive human investment opportunity for nonparticipants, and even more impressive that these long-term gains could be identified.

INTRODUCTION

PAST experience with education projects has led to two kinds of pessimism. On the one hand, there is the well-known program pessimism: how can projects be designed that raise student achievement by noticeable amounts? On the other hand, there is what might be called an evaluation pessimism: even if good education projects can be designed, how can the measurement of benefits be improved to tell whether they are working well? Until both education programs and the measurement of their effects are improved, both sides in the debate about education spending will be hamstrung. It will not be possible to argue for or against more spending on education on any basis other than blind faith.

The reasons for both types of pessimism are easy to understand. For the projects, especially those for culturally and socially deprived students, it has proven quite difficult to design programs to teach academic skills, to hold the motivation of students, and to provide for lasting benefits. For the evaluation, the basic problem is that education programs represent an investment in human capital. Costs are incurred now and benefits are realized later on —

much later on. For preschool projects most program benefits will not be realized until twenty years after the project. Given the normal difficulties of measuring anything in the social sciences, it is asking an extraordinary amount to require that benefits this much later in time be measured with any kind of statistical precision.

While these sources of pessimism are well-known, it turns out that there is promising but still somewhat tentative evidence that early preschool education works, and that the benefits can be observed and even measured much later in time. This information comes from programs conducted in such diverse sites as Murfreesboro, TN in 1962, Ypsilanti, MI in 1962, Long Island in 1965, New York City in 1966, Rome, Ga in 1966, Milwaukee in 1968 and New York State in 1975 (the reports on each are listed in the references). All projects were begun long enough ago that long-term evaluation results are now available. All were based on classical experimental principles, either using random assignment of a treatment between experimental (preschool program) and control students or some form of matched assignment. And all seem to work (see the references listed).

One of the earliest and best of these programs took place in the south side of Ypsilanti, MI at the Perry School, from 1962 to 1967. The school was located in the midst of a concentration of low-income housing near some automotive plants. The High Scope Educational Research Foundation, which ran the programs, assigned 58 black students with low IQ scores to a preschool enrichment program operating over this 5-year period. In evaluating the program, they followed these students and a like number of low-IQ black control students up to the present time, when they are all at least 19 years old. The preschool program seems to have worked very well, as based on a wide array of indicators, both formal and informal. While there are the usual number of uncertainties and while the sample sizes are very small, the evaluation is certainly respectable and the program appears to be worthwhile.

In this write-up I briefly summarize experience with this program, some of the results and evaluation quandaries, and the lessons of the experience, both for educational evaluators and educational policy-makers. Most of the information is taken from High Scope's own report on both the project and the evaluation (Berrueta-Clement *et al.*, 1984). This report gives a formal analysis of statistical differences and a full-blown benefit-cost analysis, along with some much more informal (and some would say much more interesting) case studies. Since my basic argument here will be that it is ultimately both possible and necessary to do the formal analysis, I will focus only on that side of the report.

THE PERRY PRESCHOOL PROJECT

The Perry Preschool Project was begun three years before the Office of Economic Opportunity initiated the national Head Start program. As mentioned above, the program was designed for educationally disadvantaged black students, and was set up as a full random assignment experiment from the beginning. The first group of students, in Wave Zero, entered the program as 4-year-olds and participated for 1 year. Then there were four successive waves of 3-year-olds, who participated for 2 years (see Table 1). All preschool programs featured a 2.5-hour morning classroom program each day during the 7.5-month academic year, along with one home visit per week for the whole year.

The sample size for the project is very small, only 8–13 treatment students a year for the 5-year duration of the teaching program. While this small sample permitted High Scope to focus the needed resources on the treatment students (the staff-child ratio was 1 to 5), it also makes it more difficult to observe statistically significant differences from the project. We will see below that many apparent differences between treatment and control students are not statistically significant because of the small sample sizes.

The children for the study all lived in the low-income enclave surrounding the Perry School. Families were screened for socioeconomic level and the students were then given an IQ test. They were selected for the experiment if their IQ was between 60 and 90, with no evidence of a physical learning disability. They were then assigned to either the experimental or the control group by forming matched pairs and then choosing one member of the pair for the experimental group and one for the control group on the basis of a coin flip. At program entry there were no statistically significant differences in IQ, family socioeconomic level, family educational achievement or family size. This absence of differences in family characteristics was maintained 11 years after the start of the teaching program.

RESULTS

The results of the teaching program and subsequent measurements were generally quite promising. The subjects were given a wide variety of tests and measurements, focusing on educational skills in the years immediately following the program and on broader academic and economic variables later on. Through age 19 attrition rates were remarkably small, with 98% of the students still yielding measurements.

Although IQ scores for the preschool students rose immediately following the program, by second grade there were no significant differences between the experimental and the control students. But there were other sustained differences, summarized in Table 2. To begin with, the preschool students attended school more in grades K through six. They averaged 11.9 days absent per year, while the control students averaged 16.3 days absent, a difference that could have occurred by chance with an 8.8% probability. The preschool students were

Table 1. Program Entry Data

Wave	No. in sample	Preschool group	Control group	Age at program entry
Zero	28	13	15	4
One	17	8	9	3
Two	26	12	14	3
Three	27	13	14	3
Four	25	12	13	3
Total	123	58	65	—

All students are black, with IQ ranging from 60 to 90 at program entry.
 Source: Berrueta-Clement *et al.* (1984, Table 2).

less likely to be placed in more costly special education programs, 16 vs 28%. They were less likely to be classified as mentally retarded. They failed fewer grades. They attained a higher grade point average in high school (C as opposed to D+). They were more likely to graduate from high school, and more likely to take postsecondary academic and vocational training. All but the latter differences were statistically significant.

How all these good things could happen as a result of a brief preschool program which only lifted IQ scores for a short time is something of a mystery. High Scope's own rationale focuses on the interaction between a person's capabilities and the environment. In this view, it is critical that students be well-prepared for the highly demanding school experience when it starts. Armed with their preschool-generated competence, students learn to relate to new adults (school teachers) and to display their new-found skills. They begin to like school and to benefit from it to a much greater degree. There is also an important role for the parents and the home visits in bringing about this initial adaptability. Whether this rationale is plausible, a roughly similar experience was found at the other preschool program sites — early IQ gains followed by lasting educational gains.

The attitudinal information shown in Table 2 roughly confirms this story. At age 19 the preschool students consistently show more positive attitudes toward school, though differences on the individual questions are not statistically significant. The preschool graduates also score consistently better on an adult performance survey that tests both academic skills and content (health and legal information) for living in modern society.

The other differences are self-explanatory. As of age 19 the employment experience of preschool graduates is better, earnings are higher and receipt of transfer payments lower. Preschool graduates save more, they have better arrest records and they are less likely to get pregnant or father children by age 19.

BENEFIT-COST ANALYSIS

That a preschool program can raise earnings and lower crime and arrests is well and good. How good? Since the program also costs money, analysts can be forgiven for asking whether the ultimate benefits of the program outweigh the early costs. Benefit-cost analysis has been used to try to answer questions of this sort.

While the logic of benefit-cost analysis is straightforward, its implementation is anything but that. In the case at hand, the main costs are the initial program costs. These can be readily measured by proper accounting at the time of the program. One program benefit involves the fact that fewer numbers of students will be placed in special education programs, which are more costly for the district. These benefits are estimated by combining the change in likelihood of special education with the cost of those programs to the district. Another benefit is in the higher earnings of preschool subjects — these can be measured directly up to age 19 and then extrapolated forward to the post-19 ages. The cost of crime was measured by taking the average cost of particular crimes to the victims, and multiplying by the changes in the probability of committing crimes, again observed up to age 19.

Table 2. Results

Item	Preschool group mean	Control group mean	Probability that difference occurred by chance*
School performance			
Days absent per year in K-6	11.9 (54)	16.3 (58)	0.088
In special education (%)	16 (54)	28 (58)	0.039
Classified as mentally retarded (%)	15 (54)	35 (58)	0.014
Average number of failing grades	0.67 (54)	1.01 (58)	0.073
High school grade point average	2.08 (38)	1.71 (39)	0.018
Graduated from high school (%)	67 (58)	49 (63)	0.034
Postsecondary academic training (%)	19 (58)	13 (63)	—
Postsecondary vocational training (%)	19 (58)	10 (63)	—
Attitudes toward schooling			
Positive feeling about high school (%)	33 (58)	25 (63)	—
Studied more than enough to pass (%)	22 (58)	12 (63)	—
Feeling of belonging at school (%)	46 (58)	35 (63)	—
Importance of high school as a place to learn (%)	52 (58)	43 (63)	—
Felt that teachers cared (%)	36 (58)	30 (63)	—
Mean scale score	21.3 (58)	20.2 (63)	0.083
Adult performance level survey†			
Raw scores (out of 40 max.)	24.6 (52)	21.8 (57)	0.025
Above average (%)	6 (52)	5 (57)	n.a.
Average (%)	55 (52)	33 (57)	n.a.
Below average (%)	39 (52)	62 (57)	n.a.
Employment and earnings			
Employed at time of interview (%)	50 (58)	32 (63)	0.031
Unemployed at time of interview (%)	38 (58)	51 (63)	—
Months without work since leaving school	4.9 (58)	10.3 (63)	0.003
Mean earnings, age 19, calendar year	5386 (58)	4347 (63)	—
Transfer payments			
Has received food stamps (%)	40 (58)	49 (63)	—
Has received AFDC (%)	17 (58)	21 (63)	—
Has received Medicaid (%)	33 (58)	41 (63)	—
Has received General Assistance (%)	19 (58)	41 (63)	0.007
Mean welfare payment, 1981	663 (58)	1509 (63)	0.021
Personal behavior			
Has saved money (%)	62 (58)	48 (62)	0.094
Has been arrested or charged (%)	31 (58)	51 (63)	0.021
Average number of arrests	1.26 (58)	2.30 (63)	0.001
Arrested for a serious crime (%)	24 (58)	38 (63)	0.073
Pregnancy or births per female	0.68 (25)	1.16 (24)	0.076
Children fathered per male	0.30 (33)	0.38 (39)	—

All variables measured at age 19; sample sizes in parentheses.

* In a two-tailed test. Probabilities of greater than 0.1 are not given in the report. The dashes thus signify statistically insignificant differences.

† From a multiple choice test of that name designed to assess skills needed for educational and economic success in modern life. There are differences, often significant, in both the 'content' and 'skill' sections.

Source: Berrueta-Clement *et al.* (1984, Tables 5-8, 10, 12-14 and 19, and data on p. 69).

One benefit that was not included because of difficulties in valuing it was the benefit of reduced teenage pregnancy: in this sense the analysis can be said to understate program benefits.

The outcome of the benefit-cost analysis of the Perry project can best be described by referring to the benefit-cost tally given in Table 3. The table gives program net benefits (costs are negative benefits) to participants, to others (taxpayers and the victims of crime), and the sum, which equals the benefits to society. For each group there is a separate estimation of benefits for the 1-year program and the 2-year program. On the row headings, benefits are given according to whether they are actually observed for the 19-year-old graduates or extrapolated forward on the basis of past relationships. Since benefits and costs occur at various times, and since earlier benefits are in some sense more valuable than later benefits, all entries in the table are standardized in present value terms, as if the programs were conducted in 1981.

To go right to the bottom line, the numbers in the lower right-hand corner indicate that the program is a winner. The 1-year programs yield net benefits of \$28,933 per child, the 2-year programs \$23,769 per child. Either passes a benefit-cost test, but the 1-year program passes by slightly more. Inspection of the columns indicates that the reason is that the 2-year programs have teaching costs roughly twice as

high, as would be expected, but virtually the same benefits. It is tempting to conclude that the second year of the program is not necessary, but I think it is unwise without further data. Remember that the one year program was given only in the first year, to only 13 treatment students. This is one place where the small sample size gets in the way of definitive interpretation: for now we can only say that there is very tentative evidence that many of the program benefits would also come with a shorter duration and cheaper program.

Going more deeply into the numbers, it may seem surprising that earnings increases between experimental and control students of a paltry \$600 up to age 19 translate into such large net benefit figures. The reason involves the nature of this and any other human investment program. Initial costs are borne for 1 or 2 years, but if the program works well, these initial costs permit an increase in earnings or a reduction in crime year after year. When these recurrent flow benefits are cumulated to find the present value of benefits, the sum gets very large relative to the one-time or two-time costs. And net benefits for the program get large as well. But the multiplicative nature of these calculations does lead to a problem. Program net benefits become very sensitive to assumptions made about how long the program benefits will last, and about what rate of time discount to use. Had there been assumed a

Table 3. Benefit-cost analysis: Net present value of benefits per preschool pupil, 1981 \$ (discount rate is 3%)

Net benefits	To participants		To taxpayers and potential crime victims		To society	
	1 year	2 years	1 year	2 years	1 year	2 years
Measured to age 19	226	509	2290	-2389	2515	-1880
Preschool costs	—	—	-4818	-9289	-4818	-9289
Child care	290	572	—	—	290	572
Educational cost saving	—	—	5113	4964	5113	4964
Earnings increase	482	467	161	156	642	623
Welfare reduction	-546	-530	601	583	55	53
Crime reduction	—	—	1233	1197	1233	1197
Predicted from age 19	4856	4715	21562	20933	26418	25649
College costs	—	—	-704	-684	-704	-684
Earnings increase	19233	18674	4580	4446	23813	23121
Welfare reduction	-14377	-13959	15815	15355	1438	1396
Crime reduction	—	—	1871	1816	1871	1816
Total net benefits	5082	5224	23852	18544	28933	23769

Source: Berrueta-Clement *et al.* (1984, Tables 26 and 28).

more rapid decay of benefits over time or a higher discount rate, the overall tally would look much less favorable.

Another source of uncertainty involves the fact, clear from Table 3, that the preponderance of benefits has yet to occur. For the more reliable 2-year program data, net program benefits as the students reach age 19 are still negative. They eventually become positive because of the large earnings increase that is extrapolated on the basis of the initial favorable labor market experience and the fact that more preschool graduates go on to post-secondary training. But this has not happened yet, and in this sense there is still an extra air of uncertainty about the evaluation. In my own view this long payback should not be viewed as a negative aspect of the program — the essence of education is that society is making a very-long-term investment — but it should at least be recognized.

A further interesting aspect of the benefit–cost tally involves the identity of the gainers and losers. Until now I have focused on the ‘To Society’ column, the addition of net gains to subjects and to others. But that is a rather constricting way to show the data — politicians are also concerned about who gains and loses. The table shows the surprising information that the participants themselves gain relatively little from the program. They paid no initial costs, their parents benefited from the implicit child care services (which were included in program costs), and they received a slight earnings increase, but also sacrificed income taxes (that is why there is an earnings increase for taxpayers) and welfare payments. As of now, participants in the 2-year programs are ahead by only \$509, 2% of the ultimate net benefits of the program. Extrapolation of the earnings increase offset by the welfare reduction raises the gain of participants to \$5224, 22% of the ultimate net benefits, but it is still remarkable that over three-quarters of the gains go to others. They pay for the program, but then make it all back and more in the form of reduced special education expenses, reduced crime, reduced welfare payments and higher income tax payments. And the gains from reduced teenage pregnancies are not even being valued. Speaking as one of these others, if the numbers are at all accurate, I personally am quite happy with a human investment of this sort. Even though the payback period is long, the rewards are worth waiting for.

This unexpected finding shows exactly why one

should do a benefit–cost analysis. Many of the estimates are difficult to obtain, to be sure, but doing it one can show what is really at issue in a program decision. Were the benefit–cost analysis not done for the Perry project, that project may or may not have been viewed as a winner, but it certainly would not be viewed as being in the selfish best interests of non-participants. Doing the analysis shows that it is exactly in the selfish interests of non-participants to invest in the education of others. Doing this evaluation also sharpens future evaluations of this type; now that we know where to look for the real action on these tallies, we can sharpen the measurement techniques to make better verification of the important long-term outcomes of the project.

Table 3 gives what can be thought of as average effects across treatment and control subjects. It should be recognized that, as with any average, there is a variance term, too; that is, there are uncertainties in this analysis. These were formally recognized in the comparisons of differences in Table 2, where explicit measures of statistical significance were shown. That cannot as easily be done for a benefit–cost analysis, because in addition to the quantity differences (treatment subjects are less prone to crime), a valuation of this change must be made (the average act of vandalism is worth \$x to the victim). Usually these valuation averages are simply looked up from aggregate statistics, with no readily available measure of variance. So usually a benefit–cost tally does not come equipped with variance figures, and Table 3 is no exception.

There are several ways in which such a shortcoming can be overcome. One is to conduct sensitivity analysis. Some of the uncertain parameters in the analysis are simply varied arbitrarily to see how that changes the results. High Scope did one kind of sensitivity analysis in their computation of present values. For the main results they computed present values by using a 3% time discount rate. I happen to believe that 3% is a good estimate of the real opportunity cost of physical capital in the United States, and hence a good number to use for discounting. But suppose the right discount rate is 5%. High Scope worked all this out (I have not shown the numbers) and the result is that while the ultimate earnings increases, crime reductions and welfare reductions now have a lower present value, it is still high enough to make the program a winner. Exactly the same statements could be made if High

Scope had used a higher rate of decay of labor earnings gains from the program, or assumed a lower valuation of the reduction in crime.

LESSONS FROM THE EXPERIENCE

There are two sets of lessons from all this, one for evaluators and one for decisionmakers. For evaluators there is both good and bad news. The good news is that it seems clearly possible to make very-long-term evaluations, and as in the identity of gainers and losers, there are clear surprises once these long-term calculations are made. There should then be little risk that, after doing all the work, the evaluation results will not be found interesting. The bad news is that evaluators need a great amount of patience. As all of us might have suspected about preschool training, the real gains and losses of these programs occur very far in the future — they are just now happening for the Perry project. That means that careful evaluators must be prepared to stay the course, and they must set up measuring rods that can withstand the passage of time. The simple passage of time, and the enormous changes that will probably be entailed in subjects' lives, lead also to the candid recognition that educational evaluation is never going to be one of the world's highly precise disciplines.

For decisionmakers, the good news seems to outweigh the bad. There are, to be sure, many uncertainties about this project and its evaluation. The sample size is very small, so small that it seems effectively unusable for the 1-year projects. The special education cost reduction benefits seem logical and reliable, but all the others, all the big

ones, are realized long after the program is over, long after IQ differences have evaporated, and, indeed, even after the subjects are examined at age 19. To be fully careful, one must do a follow-up analysis at age 26 or 27 (as, indeed, High Scope plans to do). It may be logical to education specialists, but I also admit to a bit of skepticism that the within-school gains could be so large when the measured IQ gains are so short-lived.

So if one wants to doubt the numbers, it is easy to find grounds. At the same time, there seem to be no obvious biases in favor of the project in the High Scope report. There are uncertainties, but not biases. Indeed, there seems to be one very large bias against the project, in the ignoring of the reduction of teenage pregnancies when the overall tally is made.

In the final analysis we should return to what is really being said here. Preschool education leads to short-term IQ gains, these lead to longer term improvement in students' appreciation of school, attendance and performance, and these in turn lead to higher graduation rates, more postsecondary training, higher paying jobs, less crime and less reliance on welfare. Every one of these results has proven very hard to come by with any program for disadvantaged people in America. The fact that such results can actually be shown for preschool enrichment programs should not be belittled, even with the inevitable interpretation difficulties. Not only are programs like this desirable on grounds of equality of educational opportunity, they even seem to be in the selfish best interests of taxpayers. To me, the results represent a stunning triumph, both for preschool training and evaluation.

REFERENCES

- BERRUETA-CLEMENT, J.R., SCHWEINHART, L.J., BARNETT, W.S., EPSTEIN, A.S. and WEIKART, D.P. (1984) *Changed Lives: the Effects of the Perry Preschool Program on Youths Through Age 19*. Ypsilanti, MI: High Scope Educational Research Foundation.
- GARBER, H.L. and HEBER, R. (1981) The efficacy of early intervention with family rehabilitation. In *Psychosocial Influences in Retarded Performance* (Edited by BEGAB, M.J., HAYWOOD, H.C. and GARBER, H.L.), Vol. II. Baltimore, MD: University Park Press.
- GRAY, S.W., RAMSEY, B.K. and KLAUS, R.A. (1982) *From 3 to 20 — the Early Training Project*. Baltimore, MD: University Park Press.
- IRVINE, D.J. (1982) Evaluation of the New York State experimental prekindergarten program. Paper presented at the annual meeting of the American Educational Research Association.
- LEVENSTEIN, P., O'HARA, J. and MADDEN, J. (1983) The mother-child home program of the verbal interaction project. In *As the Twig is Bent: Lasting Effects of Preschool Programs*. Consortium for Longitudinal Studies. Lawrence Erlbaum Associates.

- MONROE, E. and McDONALD, M.S. (1983) Follow-up study of the 1966 head start program: Rome (Ga.) city schools. Mimeo.
- PALMER, F.H. (1983) The Harlem study: effects by type of training, age of training, and social class. In *As the Twig is Bent: Lasting Effects of Preschool Programs*. Consortium for Longitudinal Studies. Lawrence Erlbaum Associates.