DOCUMENT RESUME

ED 077 568 PS 006 493

AUTHOR Datta, Lois-ellin
TITLE Planned Variation: An Evaluation of an Evaluative Research Study.
INSTITUTION National Inst. of Education (DHEW), Washington, D.C.
PUB DATE Nov 72
NOTE 15p.; Paper presented at the National Association for the Education of Young Children Conference (Atlanta, Georgia, November 15-16, 1972)

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Compensatory Education Programs; *Curriculum Research; Evaluation Criteria; *Evaluation Methods; Preschool Education; *Preschool Programs; Program Evaluation; *Research Methodology; Research Problems; Speeches
IDENTIFIERS *Project Head Start

ABSTRACT

Planned Variation was deisgned as a three-year program to assess the implementation of prominent preschool curricula in Head Start and the immediate effects of the programs. Sites used were those in which the sponsor already had a Follow Through program; the research project lacked the necessary control over site characteristics. Consultants visited the sites monthly. The classroom observation form and observer rating scale were keyed to what the sponsors said distinguished their model. Consultants developed sponsor-specific checklists. Controversy over expected outcomes and selection of tests of cognitive development created additional problems. It was found that statistical analysis could not compensate for the research design. Year 1 saw an emphasis on assessing implementation, the creation of the Classroom Observation instrument, the investment in creating new measures for years 2 and 3, the clinical case history and the consultant as innovations. Year 2 added a review panel for the project and increased the investment in developing new child and family measures. Year 3 added sponsor-specific studies, research for individual sponsors. Year 4 is for phasing out the sites. A summary is made of what was learned about evaluative research administration that may be applicable to similar studies. (KM)

# PLANNED VARIATION: AN EVALUATION OF AN EVALUATIVE RESEARCH STUDY[1]

Lois-ellin Datta
National Institute of Education

This is a report not on the findings or the results of the national Head Start/Follow Through Planned Variation Study, but on the evaluation itself: what was done, how the study was conducted, why did we do what we did, the shortfalls in methodology, approach and evaluation management, and the methodological advances. This last is an important question, since the study is among the most costly pieces of educational research conducted recently, and one 'product" is the learning about evaluation methods and management we claim has occurred.

Some definitions first: Planned Variation is a research study intended to determine which of several outstanding early childhood curricula have the greatest immediate effects in Head Start, and whether participation in well-planned, well-implemented, continuous programs would yield continuous development in the children. By grace of the demands of the Westinghouse Report, the then Bureau of the Budget, and our own concerns at Head Start with program improvement, the study was designed to explain a frequently occurring phenomenon: the curve shown in Figure 1. The curve shows an immediate impact of a preschool intervention, a catch-up by the control group after school entry, and a gradual decline in achievement



[1]Paper presented at the National Association for the Education of Young Children Conference, November, 1972.

of both experimental and control groups after the third or fourth grade. What program would yield the greatest effects or what measures for what children? Maybe the "right" program would have a lasting effect? Or would continuity of experience in any curriculum that was well-planned and supervised have sustained effects?

As previously noted, the study originated both in a concern for Head Start program improvement through incorporation of effective new curricula into the daily program, and the need to "justify" preschool intervention as public policy by the magnitude and durability of its benefits. Such a statement assumes that a happy, healthy, "good" experience for low-income youngsters would popularly justify public investment only if there were long-terms gains in matters which are of public, social concern such as academic achievement.

This is a value-issue that generates considerable heat. The concerns of the 50's and 60's with inequality of educational opportunity stemmed from a belief that education and later economic status were related. The high rate of school failure and low achievement on standard tests of reading, arithmetic, and in older grades, of language and quantitative comprehension and problem-solving, were endemic among the poor, and particularly among poor blacks. Thus public expectancy that preschool programs ought to have a durable effect on academic achievement if public funds are to be spent on income-segregated programs for which the working marginal poor and lower middle class are not eligible, is not an unreasonable expectation. On the other hand, blacks and whites of equal academic

achievement have unequal incomes (which places the blame for economic inequities on other shoulders than the schools per se) and the 1954 Brown decision was predicated on the personal sense of inequality and unworthiness assumed to be prepetuated by segregated public institutions. Thus public expectancy for preschools could well be limited to delivery of services (which most agree are well-provided by Head Start to enthusiastic parents and children), and immediate socialization benefits. Since mostly low-income children attend Federally-supported preschools, however, the in-practice exclusion of working poor and lower-middle class from Head Start has probably reduced the strength of the second argument for many taxpayers who can not afford preschools for their own children. Thus, the academic achievement issue is prominent in decisions on whether public funds support one program or another for children. The Westinghouse Study, and our own smaller-scale longitudinal studies did not show durable academic effects in most circumstances: would a good Follow Through program linked to a good Head Start have the continuity of effects expected when Follow Through was funded? And would the Head Start experience be a necessary experience or could entry into the program be delayed until Follow Through with no apparent irreversible deficit?

It should be noted at this time that the "effects" required are not limited to the IQ by some conspiracy. Motivational changes, social adjustment, positive self-image, sense of hope and self-worth, better use of basic abilities, achievement in school as measured by any appropriate instrument--the responsibility for defining and measuring the outcomes

which are educationally significant to a great extent rest with us, not with some mythical group who are bedazzled by IQs. The policy-makers to whom I have talked are far more interested in achievement and competence than IQ. We, the researchers, haven't delivered evidence on these variables, and we, not Congress or OMB, selected IQ as a reliable, meaningful proxy for other events. It is more an instance of "put up or shut up" than of crucifying children on the cross of IQ. No one I know--parents, teachers, researchers, policy-makers--wants to do this. But, in practice, unfortunately, there are few measures which are reliable, meaningfully interrelated, and feasible except the standardized tests, and this despite prolonged large investments in developing other measures.

A second point of definition: by evaluative research I mean an assessment of (1) what was the treatment or the program, (2) did the treatment or program have the effects it intended to have, and (3) how did different treatments or programs compare in the extent to which they reached their own goals (criterion-referenced evaluation) and transfer to broader goals? The Planned Variation Study was not experimental in the sense of control by the researchers of the treatment and who received it; it was a quasi-experimental evaluative research study with limited ability to control who received treatment or how many replications could be located where.

In discussing Planned Variation as a quasi-experimental study, I will consider first the research design, measures and analytic approach, and then discuss questions of research management and research utilization.

Planned Variation was designed as a three-year program to assess the implementation of prominent preschool curricula, and the immediate effects of the different programs. The curricula were chosen in all but two instances because these programs were installed in Follow Through during 1967-69, and already had extensions downward to the preschool years. A three-year program was planned from the beginning because we expected that sponsors would have to train staff and learn how to operate in Head Start, and we wanted to assess both the ease of implementation of different models, and their effects after a reasonable time for them to become fully operational. This was, and for some studies, still is an innovation in national studies. The performance contracting experiment, for example, gave only one year for the final test of program cost/effectiveness. The Experimental Schools program, on the other hand, began with a five-year implementation period. We now believe that more than one year is essential but also that time per se is likely to be no guarantee of "ideal" implementation in part because of staff turnover everywhere -- (sponsors, trainers, teachers, evaluators) and because programs are affected by many winds of change besides those of the curriculum model (funding hassles, hassles over control, and other demands on the programs). These were independently assessed by site visitors, but clearly, the best we could achieve was measurement of these other factors, and hopefully, covariance. Again, wisdom is to measure and anticipate these uncontrollable influences on the program and to shout that what is being tested isn't the ideal, but what happens in a complex real world. If we want to test the idea in its pure form, we need far more control over these other factors, the kind of

control in fact one obtains through program development effort.

The sites selected for PV were those in which the sponsor already had a Follow Through program. This meant that sponsor and geographic location and site characteristics were confounded since the Follow Through sites had not been selected to begin with to balance child age at entry, ethnicity, SES mix, urbanicity, region, and other factors which can affect entry characteristics, implementation, program acceptability, and outcomes across sponsors. (This variability was not due to the inability of the very competent Follow Through directors to plan a research study. In 1967-68, Follow Through was initiated as a national program to serve all Head Start children. After a cooling of interest in 1968, Follow Through expansion was halted and the program transformed into a national experiment, using the sites where programs had been started and a commitment made to the community and staff.)

What we have learned from grappling with the resulting "design" is that no current statistical technique can compensate for this confounding; future studies which are asking the planned variation questions must have better research control. This is, in fact, a general methodological finding: you can not put the statistical band-aid of regression analysis or post hoc matching on a research design that has a broken leg and come up with much more than hypotheses to be tested on a better day. We will not learn much from early childhood research until we will confront the issue of service vs research, and research needs come first, at least if we want findings that can move programs out of limbo. Our country is littered with programs that are dying from indifference: the data aren't unfavorable enough to justify discarding them, aren't clear enough to show

how to modify them, or unequivocally favorable enough to justify expansion. The sole exception is Sesame Street, which expanded into The Electric Company, and which combined a highly uniform treatment plus measurement by criterion-referenced tests, plus more money invested in Madison Avenue PR than most R&D programs have for development, plus authorization to expand commercially into a self-supporting corporation, plus delivery of service to virtually all homes, more than 95% of which at all income levels have a TV set.

Insofar as possible, Planned Variation required comparison of Head Start children within the sites so the effectiveness of the additional $350 per child costs of Planned Variation over and above regular Head Start costs could be assessed. On-site controls have the research virtue of comparability and the research vice of program dispersion and contamination. In many sites, there were no on-site Head Start comparisons available, and we sought off-site comparisons which were rarely comparable, on-site, we had contamination. Some sponsors accepted the research conditions; others had as their agenda reaching every child they could. Even where sponsors cooperated with the research design, teacher meetings plus teacher-staff turnover meant contamination. How substantial this was we will know when the 1968-69 data are analyzed. In some sites, there was a reverse effect: the experimental programs were not given their usual Head Start services and supplies because they were experimental, or there was rivalry. These design problems are not easily resolved: if one selects only larger sites to reduce contamination and still achieve within site comparisons, then the sample is atypical for Head Start. Also larger sites may have

several delegate agencies so the true comparability of program administration is dubious. There are design options, such as paired sites assigned at random to E and C conditions, but these take time and cooperation.

True non-Head Start controls within sites were politically unacceptable to Head Start national and, I am told, to local staff. In my opinion, this is a research error that can not be compensated for in terms of what we can say about the effects of Head Start and Planned Variation; the nature of the control group, and its incentives are a powerful determinant of "outcomes," and if comparison groups are "equally effective," there is no little danger that "no difference" findings can be interpreted as "programs are equally ineffective."

With regard to measurement, our approach was to invest heavily in describing what was actually happening. We have several techniques. Most innovative were educational consultants who visited the sites monthly. A classroom observation form and observer rating scale keyed to what sponsors said distinguished their models was developed. In 1971-72, a sponsor-specific, structured, carefully developed checklist was completed by site visitor consultants. We had teacher, aide, director, and sponsor ratings of both overall classroom quality as a Head Start program and implementation as an examplar of the model. In retrospect, this investment in description of the treatment was an immensely worthwhile decision; programs were changing and curricula were not monolithic. Implementation is worth studying in its own right and may be essential to analyses of data from a study of this kind. For outcome measures for children and parents, we spent many meetings, workshops, and conferences trying operationally to

define the outcomes anticipated by each sponsor, and to find reliable, feasible indicators for these outcomes. Some sponsors had little difficulty; for others (e.g., EDC), there was no outcome for the child--the message was the medium or process. One moral is that only treatments which begin by being able to describe what they do, and what they expect to have happen to children are suitable for comparative curriculum studies.

Despite these efforts to find good measures, Planned Variation nearly wrecked on the shoal of the Stanford-Binet: there are few reliable tests, and participants in Planned Variation--consultants, sponsors, management, evaluators--held opinions varying from calling the Binet the crime of the century and branding as racist anyone who advocated its use to me, who said then--and still do-- it's the most reliable, sensitive indicator we have of general cognitive development for a longitudinal study. After two years, the Binet was dropped to be replaced by more criterion-referenced measures, and I am hoping that these prove sufficiently reliable to be interpreted. The moral of this, if you will, is my concern that until the state of the art of measurement is improved, comparative curriculum studies may be getting us waist-deep in the Big Muddy. If sponsors have central objectives we can not measure adequately, then we dare not place them in a horserace with sponsors whose objectives can be measured reliably unless the outcome criterion is ease of implementation or treatment drift, rather than child and family development. Comparison of sponsors who share common objectives which we can measure may be the current limit of comparative curriculum studies. Perhaps if early childhood curriculum developers would use formative evaluation as vigorously as Sesame Street did and could

develop in the process criterion-referenced tests, we would make greater progress on the issue of the effectiveness of early intervention--for whom, and for what?

The analyses in Planned Variation are directed primarily to this interactive question: what approaches have what effects on which children? Is there "one best" approach across all outcome measures and for all children? Are there "equally good" approaches? Or do some programs prove effective for some outcomes but not others -- a specificity of effect that seems to me more than hinted at by existing data. Or may some programs have certain effects for some children but not others? From a policy viewpoint, the neatest outcome would be either "equally good" or the "one best" approaches. Finding a specificity of effect will require considerable re-thinking of our curricular models and developing sophistication on the part of program directors, parents and teachers in choosing outcomes wisely. Most complex would be educationally significant child x program x outcome interactions: this finding, which is at the core of "the problem of the match" and much early childhood education belief, would require even more sophistication in individualization of instruction than we now have available, except perhaps in extensions downward of I.P.I.

A different methodological aspect is that the SRI and Huron analyses have identified analytic problems centering around change or gain scores in groups with different baselines to begin with and probably different regression lines; comparison of magnitude of effects against scales which are not standardized to a common unit are equally perplexing for tests of

interactions by outcomes. Among Planned Variation's methodological contributions should be identification of which of our thorniest problems can be solved with current statistical techniques and which represent essentially unnegotiable design requirements: on what can researchers negotiate because alternative solutions are now available which will permit rigorous inference, and what represent unnegotiable demands if the outcome desired is rigorous inference about program effectiveness?

Turning from the What of Planned Variation to the How: we begin with three groups: evaluation, consultants, and case study reports. The evaluation contractor was responsible for designing the study (sample size, etc.), for developing the instruments, for fielding the national data collection effort, for analyzing the data and for writing the reports. The team selected was Stanford Research Institute (SRI), because SRI was the Follow Through contractor. Economy of effort plus continuity seemed an obvious benefit of this arrangement. The second group was the consultants intended both as an extension of the Head Start officers responsible for program implementation (Dr. Jenny Klein and Ms. Juanita Dennis) and as an independent evaluation source of information on implementation. The sense of a team in decision-making evolved during the study and was a creation of it, not a component planned from the beginning. In the second year, sponsors, consultants, OCD staff, and outside researchers formed a review panel which met fairly regularly to discuss the status of the project and policy issues. This review panel approach was adopted for Home Start, with the addition of two parents, a model which when involved from the beginning of Home Start, has greatly strengthened the design. This also is an innovation: to the best of my knowledge, no other Federal agency has an on-going review panel

for national evaluative research which includes researchers and consumers. The panels stay with their program [illegible] final report in new studies in OCD, and, if I can, they will in NIE, too.

The third part was a clinical case study of individual children that was created early one morning when Jenny Klein and I shared a room and insomnia. After a long meeting on the merry-go-round of personal-social measurement, we still weren't happy with assessment and thus couldn't sleep. The idea of a clinical approach came partly from my admiration of the work of Robert Coles and partly from Jenny's background at the University of Maryland Child Study Institute where this was the method of choice. So as an experiment (because no one really knew how to use clinical case data in a national study; it's easy to collect but there are almost no models for data reduction) the clinical case study was in from the beginning.

Year 1 thus saw an emphasis on assessing implementation, the creation of the Classroom Observation instrument, the investment in creating new measures for years 2 and 3, the clinical case history and the consultant as innovations.

Year 2 added the review panel and substantially increased the investment in developing new child and family measures. It also saw the separation of the data collection responsibility from the planning and analysis responsibility. After considerable effort to obtain acceptable reports on time, we concluded that placing the responsibilities of planning, field work, and data analysis on one contractor wasn't do-able. This is a conclusion to which I hold for longitudinal studies with high demands for

new measures and non-standard analytic techniques, and with a demand for yearly or more frequent reports for national release. In Spring 1971, Huron Institute became responsible for the Head Start Planned Variation design and analysis, with SRI continuing responsibility for data collection.

Year 3, the final year of the study, thus began the consultants, with Huron Institute, with SRI, with the University of Maryland, and the review panel as the principle components of the evaluation team. To this was added a new idea: the sponsor-specific study, which was a special set-aside for research which the sponsors might wish to do to augment the other efforts and to present to the public their program, and their accomplishments in their own way. Year 4 is a phase out year for the sites, as planned. Huron, SRI, and the sponsors are analyzing data and preparing reports. In spring, under Huron's guidance, and with the help of the consultants, OCD will collect data on what program elements remain when program support is phased out. We also are concerned with the longitudinal study--with what happens when children enter Follow Through. This is another story, with its own set of design, measurement, and policy issues and one still too much in process to write of.

To summarize what we have learned about evaluative research administration from the Head Start Planned Variation study that may be applicable to similar studies:

- allow two years or more for implementation before a final program evaluation.
- invest as much in studying the process of implementation and establishing the extent of implementation as in studying outcomes.

- select only treatments that are operationally defined, to begin with.
- select treatments where (a) there is agreement to begin with on what outcomes are to be reached (program objectives), and (b) where those outcomes can be reliably, feasibly measured prior to study initiation.
- adopt multiple approaches to data collection: observation, consultant reports, testing, case studies, and others, allowing enough time to test out data reducation and interpretation before a large scale study is launched.
- identify statistical non-negotiables in treatment, site and child selection, and stick to them if the outcome desired is rigorous inference about program effects.
- involvement of a review panel of participants, including parents, from the beginning and throughout the project, is invaluable in preventing premature closure and providing a stability of vision and concern for the study.
- separate data collection and data analysis responsibilities (within a team approach, not sequentially), allowing about two years of data reduction and analysis for every year of data collection.
- set aside funds for sponsor-specific studies and second generation research.
- and, lastly, hope to be as fortunate as we were in the hundreds of dedicated people who are willing to participate in research on behalf of children.

Few who have worked with parents, children, and field data collectors can come away untouched by the intensity of what Head Start as a gateway to a better life for children means to so many people. Far more is involved than job scarcity or protection of narrowly economic self-interest in the hours and energy so many people have given to Planned Variation: consultants trapped in snow storms, researchers who get up at midnight for just one more computer run, community people focusing an almost palpable energy on learning the classroom observation codes, teachers unlearning

the old and trying to learn the new, and most of all, the children themselves, whom I have seen and loved, and whose trust we bear. One NAEYC participant asked, "What does this mean for funding? For the children?" This is a question for which we are answerable with our souls as we report on the Planned Variation data, and learn from PV both methodological and programmatic lessons.