DOCUMENT RESUME

ED 103 494 TM 004 561

AUTHOR Butler, John A.
TITLE Toward a New Cognitive Effects Battery for Project Head Start.
INSTITUTION Rand Corp., Santa Monica, Calif.
SPONS AGENCY Office of Child Development (DHEW), Washington, D.C.
REPORT NO R-1556-HEW
PUB DATE Nov 74
NOTE 93p.; For related document, see PS 007 880

EDRS PRICE MF-$0.76 HC-$4.43 PLUS POSTAGE
DESCRIPTORS Academic Achievement; Behavioral Objectives; Cognitive Ability; Cognitive Development; *Cognitive Measurement; Cognitive Processes; *Disadvantaged Youth; *Early Childhood Education; Evaluation Criteria; Interpersonal Competence; Learning Readiness; *Measurement Techniques; *Program Evaluation; Success Factors; Testing; Test Reliability; Tests; Test Selection; Test Validity
IDENTIFIERS *Project Head Start

ABSTRACT

In past Head Start evaluations, cognitive measures have been weighed heavily. This has not accurately reflected the relative unimportance of cognitive program goals; child performance gains are not an objective with high priority for most Head Start programs. Evaluation planners need to weigh previously encountered measurement problems carefully and decide to adopt either a reliability-based strategy placing emphasis on careful test administration or a validity-based strategy assuming that what is needed is a fundamental reconceptualization of the measurement of cognitive effects, developing new measures. As priorities for cognitive measurement, this study argues that the new evaluation should stress readiness, cognitive process, and social competency and if it is decided to adopt a validity-based strategy, lists of clearly defined behavioral objectives must be drawn up in those realms of stress and then to create or adopt instruments to measure these objectives. What is needed is a battery of face-valid, empirically based, criterion-referenced instruments intended to measure short-term effects. Choice of measures is integrally related to choice of evaluation design. The new evaluation might consider some departure from pre- and post-testing, instead testing three times during the year or only once at the end. (RC)

TOWARD A NEW COGNITIVE EFFECTS BATTERY FOR PROJECT HEAD START

PREPARED FOR THE OFFICE OF CHILD DEVELOPMENT,
DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE

JOHN A. BUTLER R-1556-HEW NOVEMBER 1974

Rand
SANTA MONICA, CA. 90406

The report was prepared for the Office of Child Development, Department of Health, Education, and Welfare, under Grant No. H-9766-A-H-0. Views or conclusions contained herein should not be interpreted as reflecting the official opinion of the sponsoring agency.

Cover: Stefani Relles (age five)

Published by The Rand Corporation

TOWARD A NEW COGNITIVE EFFECTS BATTERY FOR PROJECT HEAD START

PREPARED FOR THE OFFICE OF CHILD DEVELOPMENT,
DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE

JOHN A. BUTLER

R-1556-HEW
NOVEMBER 1974

Rand
SANTA MONICA, CA. 90406

PREFACE

This report arises out of the Rand study to design an evaluation of social competence in Head Start children for the Office of Child Development (OCD), HEW. It was written as the keynote paper for one of four panel meetings of child development experts. The panels were convened by The Rand Corporation and OCD to identify candidate outcomes, measures, and research strategies, and difficulties with them, for a national evaluation of Head Start.

The report was prepared for the cognitive effects panel, held in New York October 17 and 18, 1973. It is intended as policy analysis to help OCD generate evaluation options. It weighs political considerations as well as those related to research design and asks what kind of evaluation would be most *useful* to a number of different audiences--from the Office of Management and Budget (OMB), to the Secretary of HEW, to the Congress, to parents of Head Start children, to the university-based research community.

Mr. John Butler, author of this report, is editor of the *Harvard Educational Review*.

SUMMARY

PURPOSE OF COGNITIVE EFFECTS MEASUREMENT

Decisionmakers judging the value of Project Head Start are likely to use four basic evaluation criteria:

- o Is the program well-implemented?
- o Does it have a political constituency?
- o Does it make sense as a way to support low-income parents and families and to provide them with child care?
- o Does it accomplish anything for children beyond custodial care?

These criteria are probably ranked by many in descending order of importance, suggesting that cognitive effects, included under the fourth criterion, are by no means the only basis for policy decisions. In the future, however, funding decisions may be based more heavily than before on performance outcomes. The program's political support may not be as vocal or well-organized as it once was; now that there are numerous other programs for poor families competing for the same funds, the Congress, the Office of the Secretary of HEW, and the Office of Management and Budget may give the program new scrutiny on grounds of cost-effectiveness.

The primary audience for a new evaluation comprises national legislators and agency personnel. It is not clear what must be demonstrated to convince these groups of Head Start success in the realm of cognitive effects. Five positions seem tenable:

1. In a randomly selected group of Head Start programs *effects must be generalized* and preferably must persist into the elementary grades;
2. *Some kinds of Head Start programs must achieve sizable cognitive effects* with participant children;
3. There must be strong cognitive benefits *for some subgroups of Head Start children*;

4. Cognitive effects need only be demonstrated as a moderator variable--*cognitive goals are not the principal outcomes to be evaluated;*
5. *We do not at present have the measurement technology to assess Head Start's cognitive effects.*

Many policymakers currently espouse the point of view that Head Start must show universal and lasting effects. This raises a fundamental difficulty: It can be predicted that no evaluation design looking only for generalized effects is apt to teach us anything not learned from the two previous national evaluations of the program (Westinghouse-Ohio: Cicirelli et al., 1969; Planned Variation Head Start (PVHS): Smith et al., 1973; Weisberg, 1973); in addition, no evaluation pursuing longitudinal effects with sufficient care is apt to be worth the money. If evaluators adopt position one, they are apt to learn little new about which programs are working and why. It is therefore important to shift the terms of the evaluation away from this position.

Positions two and three have considerable appeal if the evaluation can take the form of a small-scale, well-controlled study of certain program prototypes, perhaps ranked according to cost of delivery. Such a study would be most effective, however, with true randomization of children to programs, clear operational definitions of program prototypes, and sufficient controls. None of these were evident in the Planned Variation Study, which proved at best a preliminary, hypothesis generating venture. To some extent a careful study may involve trading full representativeness of the sample and a large battery of measures for increased depth of analysis on some smaller group of children or programs.

Position four also has definite appeal, especially if the evaluation is to emphasize outcomes in the realms of health and nutrition, social development, or the effects of Head Start on the family. Position five is maintained by certain skeptics, but as a practical matter it must be rejected because some evaluation, however imperfect, is required.

ROLE OF COGNITIVE EFFECTS MEASURES IN THE HEAD START BATTERY

In past Head Start evaluations, cognitive measures have been weighed heavily. This has not accurately reflected the relative unimportance of

cognitive program goals; child performance gains are not an objective with high priority for most Head Start programs. In general, past evaluations also have been plagued by three measurement problems: low quality of the field operation for test administration and other aspects of data collection; poor theoretical rationale for the individually administered cognitive tests in the Head Start battery; and equally poor theoretical rationale for observational and interview measures, with the additional problem of low reliability for such measures.

Evaluation planners need to weigh these problems carefully and decide whether to adopt a *reliability-based* strategy in devising the new evaluation, or a *validity-based* one. A reliability-based strategy would accept as given a limited number of the best available instruments and place emphasis on careful test administration. New data need not be different in kind from past data, only of better quality. A validity-based strategy would make a different assumption: that we need a fundamental reconceptualization of the measurement of cognitive effects, developing new measures. This strategy would require more time to fulfill.

Cognitive effects can be loosely divided into five realms: (1) norm-based kindergarten or first-grade readiness; (2) theory-based developmental shifts; (3) changes in cognitive process; (4) social competency and awareness; and, (5) general knowledge. As priorities for cognitive measurement, this study argues that the new evaluation should stress readiness, cognitive process, and social competency.

THE NEW COGNITIVE EFFECTS BATTERY

It may be unwise to spend additional funds on the development of new instruments. There is ample evidence from laboratory-school studies as well as from the two national evaluations of Head Start (Westinghouse-Ohio and PVHS) and the Educational Testing Service (ETS) Longitudinal Study that good Head Start programs show consistent short-term effects on a variety of measures. A new set of instruments might show only the same pattern again.

If evaluation planners do decide to adopt a validity-based strategy and devise new measures, they need to begin by making lists of clearly defined behavioral objectives in the realms of readiness, cognitive

process, and social competency and then create or adopt instruments to measure these objectives. What is needed is a battery of face-valid, empirically based, criterion-referenced instruments intended to measure short-term effects. At present there is a paucity of good measures of this type.

A related issue is the appropriate balance of individually administered tests to observational measures or rating scales. In general, individually administered tests are more reliable but tend to measure too small a slice of the child's world. Other instruments are higher in risk but also higher in potential gain: They are more likely to be unreliable or of low validity, but if they successfully overcome these obstacles they stand to be more persuasive than other instruments.

Choice of measures is integrally related to choice of evaluation design. Past evaluations have tried to investigate too much at once, throwing even the most elementary conclusions into doubt. One persistent problem, as an example, is that the same tests may not be appropriate for both four and five year olds.

The new evaluation also might consider some departure from pre- and post-testing, instead testing three times during the year or only once at the end.

ACKNOWLEDGMENTS

The author is indebted to various experts in child development, testing and measurement, and program evaluation for their ideas and wisdom about past experience. Among them are Joan Bissell, Jerome Kagan, Gerald Lesser, Richard Light, Frederick Mosteller, David Mundel, Vicki Shipman, Marshall Smith, Sheldon White, and Susan Woolsey.

CONTENTS

I. INTRODUCTION

Each of the three major sections of this report is a discrete unit, but each follows from the previous section in its logic and increasing level of specificity. Section II is a consideration of the purposes of the measurement of cognitive effects. Why are we looking at cognitive effects of Head Start at all? What should an evaluation of cognitive effects set out to demonstrate or explore? Some of the issues raised are generic to Head Start evaluation, applying as readily to other measurement domains as to the measurement of cognitive performance. Other issues surround the role of cognitive measures in particular, what they have meant in past Head Start evaluations, and what they should be designed to accomplish in the next.

The remaining sections deal more directly with practical and technical questions of measurement. Section III discusses the role of cognitive effects measures within the Head Start battery. What has been done in previous Head Start evaluations to assess dimensions of cognitive performance, what are some of the problems in measurement, and what can be done to improve the test battery itself and the quality of the data generated in a new evaluation? Section IV builds on the conclusions of the previous section and asks what kinds of instruments should be included in the new cognitive effects battery. In the domain of cognitive competence, what are appropriate behavioral objectives? Categories of measures are listed, ranging from individually administered pre- and post-tests, to classroom observation instruments, to interviews and rating scales. "Best bets" are considered among established measures and promising new ones. The report concludes with a brief discussion of the relation between choice of cognitive effects instruments and choice of overall experimental design.

II. PURPOSE OF THE COGNITIVE EFFECTS EVALUATION

Most would agree that a program evaluation should be decision-related--designed to enable policymakers, researchers, parents, or others to make rational choices. Too often researchers have applied an analysis of variance model to the world without first asking why they were doing it and what kinds of information it is apt to generate. Who are the primary audiences for the evaluation? What are the minimal sufficient data that can tell us what we need to know? And within budget constraints, which evaluation strategy will yield the highest return in valid and useful information given the dollars it costs to implement?

THE FUNCTIONS OF PREVIOUS EVALUATIONS

In designing a new national evaluation of Project Head Start, it is helpful to begin by recalling the purposes of past evaluations and considering how their results have been used. The first national evaluation of Head Start (Westinghouse-Ohio, Cicirelli et al., 1969) was an impact study, intended to find out whether Head Start programs in the aggregate were having any effect. Although at that time the program was still too young for conclusive judgments about its success, questions o cost-effectiveness were in the minds of many: Was Head Start a wise expenditure of federal funds or could comparable sums of money better be spent on children in some other way? The Office of Economic Opportunity, then sponsor of the program, was to provide an initial estimate of the program's effectiveness as preliminary data for a rationally based, go no-go decision on Head Start for coming fiscal years.

The Westinghouse-Ohio evaluation placed heavy emphasis on measures of children's cognitive performance. It tried to answer one basic question: Are pre- to post-changes in the performance of children in a randomly selected group of Head Start programs higher than those experienced by comparable children without any program? The design looked for effects generalized across the entire Head Start population, regardless of particular center, location, or child subgroup. Head Start children were compared [illegible] with children without any preschool experience.

The methodological pros and cons of the study are amply discussed in an exchange between the principal investigator, Victor Cicirelli, and Smith and Bissell in a 1969 issue of the *Harvard Educational Review*. The actual *use* of the evaluation by policymakers, however, has never been formally analyzed. Several hypotheses can be ventured, based on conventional wisdom about the influence of the report. First, its principal finding--only very slight effects across programs on the most reliable measures, not enough to impress anyone with Head Start outcomes--probably served to dampen the enthusiasm of many liberals and policy researchers about the prospects for an early childhood "cognitive inoculation" against the ravages of poverty. There were no apparent quantum jump in the cognitive competence of Head Start children compared with other children. Also, and more disturbingly, the evaluation probably reconfirmed the belief among many conservatives that Head Start efforts were a fool's errand. Results could be interpreted in support of the view that environmentalists had been too sanguine about the malleability of early intelligence and cognitive performance.

Second and equally important, however, was a political groundswell supporting Head Start and believing that the terms on which it had been evaluated did not accurately reflect the goals envisioned by its architects or community participants. In some cases, this opposition took the form of scholarly rebuttal of the Westinghouse-Ohio Report, but scholarly response was probably less important than the fact of a continued, strong political constituency for the program in the field--Head Start parents, teachers, and other supporters--who believed that Head Start *did* make a difference, that it was worth the money, and that it would be a mistake to end the program. In general, liberal support for Head Start at various levels sustained itself despite lukewarm evaluation results. Although the program was closely scrutinized from then on, no decision to curtail the program was made on the basis of the findings.

The recent Planned Variation Head Start Evaluation (PVHS 1969-1971: see Stanford Research Institute, 1971; Smith et al., 1973; Weisberg, 1973) was both more sophisticated in research design and more astute in its anticipation of political implications. It did not directly address

the issue of generalized effects--the go no-go question--instead it asked another question: Among various Head Start prototype programs developed at laboratory schools, which ones were most effective when replicated in a field situation? Also, what were the *differential* effects of the various programs when compared with each other and with traditional, non-sponsored Head Start programs? The evaluation asked not whether Head Start was succeeding on the whole, but rather *which* Head Start programs were succeeding or achieving unique results. The Office of Child Development (OCD), now sponsor of the program, had in mind an incrementalist strategy: Discover which programs are most successful and then build on them for the future (see Light and Smith, 1970). Findings were intended to inform two kinds of decisions, those by the agency itself about which programs to support most heavily in the future and those by parents and communities about what kind of prototype curriculum best suited their needs. All children in the study were attending either sponsored Head Start programs, based on a prototype, or traditional programs. Aggregate comparisons of Head Start and non-Head Start programs could be made only by pooling all of the data and using prescores of older children in the sample to simulate a non-Head Start control group.

PVHS was one of the most ambitious natural experiments yet attempted in education at any level, and its full implications have yet to be fully sorted out. But it is clear that many of the problems of the Westinghouse-Ohio evaluation recurred in the attempt to extract policy implications from the results, and some new problems arose. Except in the case of a few sponsored programs, effects in PVHS as assessed by traditional measures of cognitive performance continue to be slight. Proponents of some programs continue to say that evaluation instruments did not measure what their programs were setting out to accomplish. Detractors continue to say that most Head Start programs do not have sizable effects and are not worth the money. Differential effects, the main area of exploration, apparently have not as yet been the basis of any policy decisions by OCD about which prototype programs to support for the future or any decisions by community groups about which program

configurations are apt to best serve their needs. (For an ample discussion of problems in making policy inferences from studies of differential effects, see Stodolsky, 1972.)

The only other national study of Project Head Start, the ETS Longitudinal Study (see Shipman, 1973), still has not been fully analyzed. Its intent is less explicitly policy-related than either the Westinghouse-Ohio or PVHS evaluations, and its architects do not expect fundamental policy decisions about the future of the program to be based on their findings.

There is much still to be learned about the relation between evaluation results and program-related policy decisions (see, for instance, Cohen, 1973). But in the case of Project Head Start this much is clear: Budget decisions from year to year have reflected little of the direct influence of evaluation results. Inflation and extension of program services to new realms have necessitated many program cutbacks, but as yet there has been no dramatic dismissal of the program by the Congress, the public, the Office of the Secretary of HEW, or the Office of Management and Budget. Head Start's budget has risen since 1965 despite evaluation outcomes.

What should this past experience tell us about the measurement of cognitive performance in a new national evaluation? First it should make us reexamine the significance of cognitive effects as they relate to decisions about overall funding. In general, evaluation results are only one of many indicators in a complex political equation determining whether the program is sustained, curtailed, or subsumed under another program. Decisionmakers are sensitive to program popularity as well as to program effects, and four basic criteria are likely to influence their opinion of Head Start's value:

1. *Is the program well-implemented?*

As an input consideration, do Head Start centers look in the field as they should according to written descriptions? Is there an efficient delivery system and management structure? Are the program's various components functioning well?

2. *Does the program have a political constituency?*

Are there sufficient numbers of parents, community members, and agency employees who like the program and what it tries to accomplish? Are these people powerful enough in their numbers and lobbying finesse to push for budget increases? Prevent budget cuts?

3. *Does the program make sense as a way to support low-income parents and families?*

Is Head Start the best mode of child care delivery? How does Head Start articulate with other federal programs for the poor, such as AFDC, child care under Title IVa of the Social Security Act, and Medicaid? Should it compete with them for funds?

4. *Does the program accomplish anything for children beyond giving them basic custodial care?*

Are there measurable developmental or educational benefits of this program not experienced by non-Head Start children or children in custodial care programs?

Policymakers probably rank this rough and ready set of criteria in descending order of importance. If so, it is not surprising that past evaluation data on child performance in Head Start has been used selectively, often to rationalize decisions made for other reasons or to bolster a preconceived view of the program's value.

A second conslusion from past experience, as a corollary to the first, is that in general political support for Head Start is not dependent on evaluation results and, conversely, probably must be sustained independent of such results. OCD should not assume it can defend Head Start by evaluating it. If the program needs more friends in influential places, OCD should consider creating advisory panels, talking again to congressmen, and establishing a broader base for the coalition supporting the program, including businessmen and others not usually included. There may even be a need for a new and full-scale public relations effort simply to remind the nation that children and

parents are enthusiastic about the program, that centers are clean and well-organized, that teaching is the best available, and that Head Start offers numerous indirect community benefits.

To say that evaluation results will not relate simply and directly to funding decisions, however, is not to say that such results are inconsequential. Strong positive or negative findings, in particular, might be weighed more heavily now than in the past. At a time of federal budget cutbacks the government is more serious than it once was about criteria of cost-effectiveness. This is especially true if liberal supporters who backed the program in the 1960s and early 1970s, despite ambiguous evaluation results, are not so enthusiastic as they once were.

If the primary audience for a new evaluation is the decisionmakers who determine future funding of the program, a third conclusion can be drawn: We still do not have any clear decision rule in the domain of cognitive effects for what would constitute program success. In the past, results have been a Rorschach blot of sorts, open to varying *post hoc* interpretations. There has been no operational definition to tell us when Head Start is "working" or "not working." This problem is especially thorny in that there are numerous reasonable, competing conceptions of success. The next section discusses the issue of success criteria in some depth.

CRITERIA OF COGNITIVE SUCCESS

Five basic positions have been taken regarding a sufficient demonstration of Head Start's cognitive effects. Each leads to a rather different evaluation design and a different role for cognitive measures within that design. The positions are presented here in order of descending stringency:

- *Position* 1. In a randomly selected group of Head Start programs, there must be demonstrable short-term cognitive effects for participant children that are not enjoyed by non-Head Start children. These effects must be generalized

across centers and preferably should last into the elementary school grades.

The logic of "go no-go," which dictated the Westinghouse-Ohio design in 1967-68, is reflected in Position 1. Many policymakers wish to establish whether Head Start programs *in the aggregate* have effects on participant children--whether there are generalized Head Start effects not enjoyed by children outside the program. In the Westinghouse-Ohio Study, Cicirelli and his colleagues were responding to this question, anticipating that overall conclusions were more important than fine-grained analysis of whether Head Start worked better for some children than for others.

Arguably any positive evaluation must show effects for the aggregate Head Start population. It may not be enough to select a group of the best Head Start centers and compare them with each other and with traditional centers, as was done in PVHS, or to compare exemplary centers with non-Head Start controls. If effects generalized across all centers are the fairest estimator of what the government is getting for its investment, then the evaluation design must involve a random sampling procedure or at least a representative stratified sampling procedure including centers of all types.

A second aspect of evaluation, which did not play a role in the design of the Westinghouse-Ohio evaluation but has occupied researchers based at lab schools and those doing follow-up studies for the past several years, is trying to measure whether effects last over time. This has also been a question of great interest to policymakers, some of whom believe it is necessary to demonstrate lasting effects in order to justify continuation of the program. Program success can be demonstrated only by showing such effects and would be conclusive if effects for all Head Start children, or for a significant proportion of the children, were maintained well into elementary school.

Although Position 1 is the dominant view of many, I will argue that it would be a serious mistake to design a new evaluation that assumes these are the most valuable criteria of success and failure. Let us pursue further the logic of designing an evaluation to demonstrate

generalized effects, then lasting effects, to see the pitfalls that await if we adopt this position.

First, it may not be feasible or desirable within the OCD evaluation budget to administer a full battery of tests to a nationally representative group of children. The major issue here is the cost of administering various kinds of tests at acceptable levels of reliability. Curiously this is not something that has ever received systematic study in the context of Head Start evaluation. It is clearly better to do *well* an evaluation of modest proportions that can inspire the faith of policymakers, professional researchers, and community participants because it is well-executed and its results are reliable, than to do something overblown and unconvincing. In the trade between quality and scope, reliability of measurement has to be emphasized in the first instance, even if it means considerably reducing the number of Head Start centers or children in the study.[1] This in itself, given budget constraints, may rule out a new national impact study.

Another problem of any generalized effects study is that all kinds of programs must be represented, or at least have an equal chance of being represented, and variations now abound. With the advent of the Improvement and Innovation program, involving substantially different time and place options within Head Start, it is not clear that there is any longer much reason to consider Head Start a single program. Perhaps there never was. From the standpoint of Position 1 the level of analysis that most interests us is the highest one, pooling every

---

[1]Along with other information on test standardization it would be interesting to know (1) how much it costs to train and pay those administering a given test to do so at an acceptable level of reliability in the field, (2) how much it costs to mount a field operation of size x that would yield acceptable data on the test, (3) the tradeoff between reductions in price of administering a test and the resultant marginal reduction in reliability, and (4) the trades between different kinds of tests (e.g., individually administered pre- and post-tests as against classroom observation instruments) in cost and reliability of the data collected. Cost per test for any kind of measure in the battery could be plotted as a function of testing procedure, length of test, training necessary for its administration, acceptable level of reliability, and other variables. This kind of function would enable rational decisions in an arena where to date decisions have been made impressionistically.

program and every child. This not only assumes a homogeneity of offerings, which may not be accurate, but also that effective programs are best considered side by side with ineffective ones--that the grand mean is more important than means for particular kinds of programs and groups of children. Neither assumption seems wise.

Following the logic of Position 1, we can estimate what in Bayesian statistics is called a "prior"--a preliminary guess about the likely magnitude of effects in the evaluation. The effects of interest would be differences between test gains for Head Start children and test gains for non-Head Start children over the Head Start year, as in the Westinghouse-Ohio evaluation, or simple differences at post-test on criterion-referenced measures if at the outset there were trun random assignment of children to treatment groups. The magnitude of the difference between gains of Head Start and non-Head Start children on many cognitive tests probably can be estimated with reasonable accuracy simply by looking at past evaluations that compare children in traditional or non-sponsored programs with children not attending Head Start at all. Accumulating evidence from a variety of studies (Light and Smith, 1971), the evaluation designers could establish overlapping distributions as in Figure 1, one for gains of children who experienced Head Start and the other for those who did not. The difference between

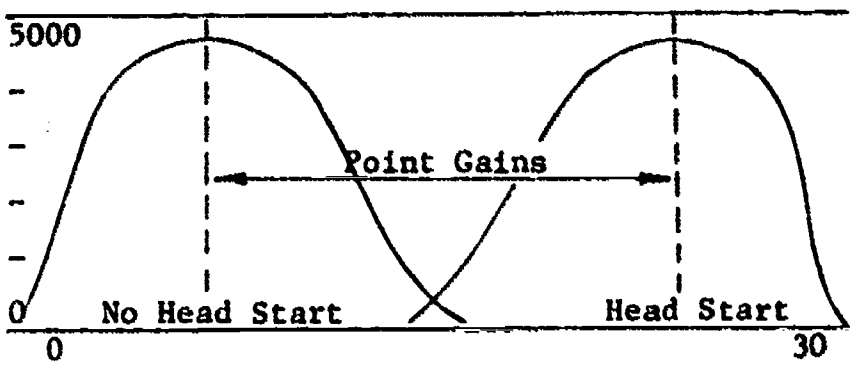

Fig. 1 - Distribution for children who did and did not experience Head Start.

the means of the two distributions, with an appropriate confidence interval around it, would be the prior expectation of Head Start effects. Knowing laws of sampling error and prior estimate of effects, the evaluation designers can next establish sample size, if need be selecting a large enough sample to give a reasonable assurance of significant results.

All of this sounds reasonable and rational. But it leads to problems when applied to a design based on Position 1. First, a prior estimate of differences between the aggregate Head Start and non-Head Start groups on most reliable, individually administered tests might turn out to be as small as a quarter of a standard deviation. This suggests immediately that if a sample is going to be representative, avoiding obvious problems of non-sampling error, it is also going to have to be very large. We also must ask whether it is enough to show psychometrically reliable differences between groups, or whether differences of such a size, however well-measured, will still be seen as trivial by policymakers and others. There is no good evidence about how large a gain has to be before it is taken seriously, but a quarter of a standard deviation, whether or not it is statistically significant, probably is not enough to excite anyone greatly about Head Start's short-term cognitive effects. Carl Bereiter (personal communication, July 1973), for instance, once asked his introductory psychology class how large a Stanford-Binet IQ gain would have to be before they were impressed that it "made a difference." He did not attempt to define what this phrase meant. Almost everyone gave an answer in the vicinity of eight points--or half a standard deviation.

The question of what is a *sufficient* gain to interest policymakers, apart from statistical significance, is not unimportant. Planners may be forced to make estimates of the face validity of change scores. If *no* reasonable estimate of Head Start effects that could be anticipated from a design matching Head Start children with non-Head Start children would yield very sizable overall effects, even in the short term, then this is an important reason to question the logic of Position 1. Once this position has been adopted there may be no way to impress policymakers favorably, because a demonstration of statistically significant

differences, which no doubt could be engineered for a price, simply would not yield sizable enough results. In addition, of course, such an evaluation strategy would not in the first instance tell us how various Head Start programs *differ* in their effects, an issue that PVHS was intended to explore and about which there are a number of interesting hypotheses.

There is also a problem in trying to show that effects last over time. We know that for non-sponsored programs--traditional programs not based at any lab school or involving any specially developed curriculum--short-term effects of preschool tend to wash out soon. We also know that even for the best lab school program, effects are difficult to sustain beyond the first three grades of school (see, for instance, Stearns, 1971, and S. White et al., 1973.)

There are two camps regarding the further exploration of longitudinal effects. The first group maintains that such effects could be demonstrated, or at least explored, if we were willing to invest the money to mount a research effort sophisticated enough to find them. Such a design has been weakly approximated in the national evaluation of Project Follow Through, but because of cohort attrition, non-comparable treatments, non-comparable child populations receiving different treatments, and other design problems, there has been no pretense that this is an adequate study to tell us what we would like to know. The only longitudinal studies to date that have approached sufficient methodological rigor have been those tracing small, lab preschool groups into elementary school. Even these studies have often been suspect, with inadequate controls and blinds in follow-up assessment procedures. For those who advocate exploration of longitudinal effects the issue is not whether such studies are feasible and valuable, but whether as a practical matter we are willing to spend the money and perhaps exert the necessary persuasion and social control to follow well-matched groups of children through the elementary school years, despite the difficulties created by high rates of geographic mobility, the need to orchestrate treatments in the elementary schools, and the enormous sample size that would be required for valid inference about effects, which are apt to be marginal.

A second group does not believe the enterprise of exploring longevity of effects is at all valuable. Along with certain experts in research design and methods, this group includes economists concerned about the misuse of cost-accounting procedures and a number of educational sociologists pursuing the logic of the findings contained in Inequality, by Christopher Jencks and his colleagues at Harvard's Center for Educational Policy Research (1972). The principal argument put forth is that although such a long-term research effort may be feasible, any effects would be small indeed, explaining only a negligible portion of the variance in subsequent school grades or other later outcomes of interest. No effects of any curricular intervention, however well organized, can be expected to last five years; and to formalize this criterion as a measure of Head Start program success is absurd. The cost-benefit economists in opposition add that any intervention probably will explain so little of the variance in subsequent school performance that even if it could be demonstrated that [illegible] Head Start had a greater effect or more explanatory power, no policymaker or cost-accounting expert would be impressed enough with the marginal differences to act on the basis of it. If Head Start is to be viewed as an investment, then the real question is one of "best bets" about maximum social return for each dollar invested in the child; unfortunately at present we have no systematic way of comparing the relative benefits of programs for older and younger children.

The sociologists make a related point. If we look at the Jencks et al. work, no school-related input in the lives of young children at any socioeconomic level currently seems to have much effect on sixth and twelfth grade achievement. Moreover, as a second missing link in the chain, these later achievement scores do not seem to predict strongly to various adult success criteria of interest, notably adult income. School effects in general do not seem to have much to do with social mobility or with aspects of adult success that really matter. If nothing related to schooling at any level predicts strongly to important outcomes later, why should we expect this will be any different for Head Start, and why should we make lasting effects a criterion

of program success? We do not discontinue schooling on these grounds.

In weighing the merits of arguments for and against a longitudinal effects evaluation, we must give the first camp its due. No one has ever attempted a careful follow-up design in the field, and such a study no doubt could be performed if money were forthcoming. In addition, as a political fact, many policymakers for better or worse have become wedded to the notion that Head Start must demonstrate lasting effects in order to justify itself. We as researchers have trained policymakers to think in such terms, and now we may find it difficult to reverse this line of reasoning.

But it clearly makes most sense to side with the second group, rejecting the predictive validity of gains as a success criterion. First, to do a longitudinal effects study in the field would be too expensive to be worth the money. In itself it would not withstand cost-benefit analysis: Results would probably be meagre and their policy implications unclear. Second, to accept this success criterion is to impose an unfair burden on the program. In other federal program evaluations it is almost always sufficient to demonstrate success in the short term only. Thus, for instance, Medicaid expenditures are not typically justified according to how they affect the life expectancy of the patient or how they reduce the probability of his returning to the hospital with some new problem four years hence. The aggregate effects of the Medicaid program for the price can be compared to the effects of previous programs, but since there is little similarity between a program like Head Start and previous schemes for the poor, such an approach is not very useful.

There is a need to shift the terms of the debate about accountability, proposing reasonable competing conceptions of accountability and reasonable ways of justifying Head Start's existence without requiring that it demonstrate longitudinally stable effects. To the extent we equate "giving OMB and the Secretary of HEW's office what they want" with mounting an evaluation based on Position 1, we have made a serious mistake.

A third design aspect, which escaped consideration in the Westinghouse study and subsequent Head Start evaluations but is wholly in

keeping with the logic of cost-effectiveness and go no-go, is trying to assess whether Head Start is a better investment than some other program for the same target population. It can be argued that decisions about the future of Head Start are more dependent on assessment of its merit in comparison with federal expenditures for other kinds of children's programs than on assessment of program effects in themselves. No evaluation of Head Start has attempted to compare its costs and effects with those of another program, such as Sesame Street. But there has been impetus from various sectors of the agency structure for just such comparisons. Certainly this was one of the reasons for OCD's recent g[illegible]al report on compensatory programs (S. White et al., 1973).

It is hard to think how this third design feature could be built into Head Start evaluation without being artificial or raising a hue and cry from many who would rightfully feel that Head Start was once again being made to conform to a procrustean set of evaluation criteria having little to do with its intended aims. Thus, for instance, if it were suggested that the Sesame Street test battery be applied to Head Start and that comparisons of children's performances in both programs on letter and number recognition, understanding of relational terms, and [illegible] subsequently be made using this single set of tests no one at Head Start would be pleased, and even the most sanguine advocates of cost-effectiveness probably would agree that such an approach is simplistic.

[illegible] 2. It is sufficient to demonstrate that a new Head Start program achieve sizable cognitive effects with participant children.

The most often-mentioned alternative to a simple comparison of Head Start and non-Head Start children is a differential effects study comparing various prototype programs. This is what PVHS was, and many feel that it remains the most promising approach for assessing Head Start effects. Any study that looks in the first instance at aggregate effects tends to overlook differential effects. Since we would expect such effects, with some programs being more successful

than others, it makes little sense to focus primarily on the aggregated programs.

A differential effects study focuses on interactions of program type and other independent variables. The strategy was first discussed in the aftermath of the Westinghouse-Ohio study (Light and Smith, 1970; Smith and Bissell, 1970). Theoretically it would allow incremental selection of good programs, an approach appealing to policymakers. If some programs seem to be working and others do not, then it makes sense to fund the ones that are. In addition, predictable interactions of program type and child population, if any, have possible policy implications. Such a study may offer communities information for coming to a decision about which kinds of programs they would prefer. If they know something about the predictable effects of a given program type, they may be able to make better choices about what seems best for their situation and their children.

Some programs also may turn out to be more robust than others in their effects regardless of child population, and policymakers may prefer to back such programs. For purposes of aggregated comparisons in such a study the data subsequently may be pooled, enabling judgments similar to those made in a general effects evaluation. In the main, however, the question of "whether Head Start works" is finessed; any conclusion says that it works in some programs under some conditions and not in others.

These are clear advantages, but such a design also has several potential disadvantages. The first arises from lack of clarity about what a coherent educational "treatment" looks like. This is a lesson we have learned from PVHS. In general, there are two ways to sort programs into typology. One is to begin empirically, first going to the field, observing the full range of *natural* variations among programs, and then constructing a matrix of dimensions on which programs differ. These dimensions become the treatment variables or one set of independent variables in a subsequent study. This approach makes it very difficult to evolve an agreed-upon list of program difference dimensions with adequate face-validity of adequate salience to explain much of the variance in effects. We are forced to choose among

various partially face-valid grouping schemes, focusing on widely differing aspects of classroom process, teaching style, instructional materials, and teacher-child interaction. Some sense of how hard it is to derive "natural variations" of classroom process can be gained by reading Jackson's Life in Classrooms (1968). No equivalent work is avaliable for preschools.

If we forgo the empirical approach, then the alternative is to turn, as PVHS did, to *planned* variations--"treatments" in the form of theoretically different programs, each representing the best efforts of an individual or research team at a university-based laboratory school. Template programs are generalized to the field situation. In PVHS, a number of promising programs were selected, including among them prototypes with widely differing aims and teaching strategies. These programs could be grouped according to their differing philosophies regarding teaching materials and techniques, degree of teacher-initiated activity, hours spent in didactic exercises as against free play, and so forth. In PVHS this approach generally supported a weak dimensionalization, with certain gross and face-valid differences between programs on a dimension called "structure." But regrettably it did not support much more. Many of the purported differences among PVHS sponsors were not readily apparent when programs were visited in the field. Even within programs of a single sponsor there often was wide variation in different sites, so that a Bankstreet College program in site X might look more like an Educational Development Center program in site Y than it did like another Bankstreet program in site Z.

This confusion has given rise to an entirely new field of inquiry, as is often the case when there are unanticipated complications in an evaluation and when the research community senses the logic of the situation. The new field is called "implementation research"; the object is to determine how well the sponsor template--the original program configuration created in the lab school setting--is replicated in the field. It has been discovered, and is still being discovered, that in this new area of research all the problems of an empiricall dimensionalization reassert themselves at one remove. Criteria are needed to decide how well a program in the field matches its template

program and other second generation programs in other sites. This means that the goals of each sponsor must be operationalized and we are once again in difficulty.

Any evaluation scheme studying the differential effects of natural or planned variations is committed to looking at such effects in the context of a weak dimensionalization. Even the staunchest proponents of Position 2 are humble about the problem of categorizing programs as coherent treatments and understanding how they differ. Accepting this limitation, it may nonetheless be valuable to group programs along the kinds of dimensions proposed by Bissell (1970) and Mayer (1971), or those used in the PVHS evaluation (Featherstone, 1972; Smith et al., 1973). It might even be enough to separate programs on only one or two face-valid dimensions, perhaps the ones with greatest consequences for program cost or the ones with most promising consequences for a theory of pedagogy.

By comparison, Sesame Street is in an enviable position. It is a coherent treatment that does not differ from site to site and suffers minimally from "noise" as it is disseminated. It is also reasonably modest about what it purports to teach. For those who have never considered problems of program dimensionalization and implementation, it is instructive to think of Sesame Street as an analogue, or better an opposite, to a Head Start treatment.

There is another major problem with a differential effects study. Perhaps all programs cannot be evaluated with the same instruments. The logic of non-comparable treatments can lead quickly to the position that non-comparable outcome measures are required. This problem is especially evident in planned variation studies and laboratory-school studies comparing more and less "structured" programs. To make matters worse, there simply are *no* trusted measures in many domains, especially those of affective development and self-concept, that might enable assessment of the goals certain sponsors say they are trying to achieve (Walker, 1973). The choice among current instruments is a harsh one: Either we must include measures with extremely low reliability and validity in the battery or we must exclude them and risk a biased evaluation. This problem can never be fully resolved until there are

equally valid and reliable measures for *all* relevant domains of Head Start process and outcome. One compromise solution in the meantime might be to have a basic battery of tests on which all programs are compared, and then allow each program to select one or more additional measures that it alone will use, or that all programs will have to use at its request.

A final problem of any differential effects study, pointed out by Stodolsky (1972), is that if differences in program effects are found, their policy implications often are unclear. If we discover, for instance, that the Weikart Hi/Scope program results in large gains on measures of general intelligence but that some other program results in happier parents, how does this readily translate itself into educational policy? Certainly such findings are useful information for community groups choosing a new curriculum, but they do not in any obvious way inform agency decisions about which programs to support in the future and which to terminate. The government does not get the kind of information that would enable it to distill the best configurations from the initial group of prototypes by successive approximations, and the incrementalist strategy envisioned by Light and Smith (1970) is not readily fulfilled.

Despite these complications and disadvantages, Position 2 may be the most reasonable to espouse in a new Head Start evaluation. It remains attractive for three reasons. First, more from a political than a measurement standpoint, a differential effects design justifies looking closely at a subset of the *best* or most clearly defined programs. If we are interested in differential effects, we must be as clear as possible about the treatments compared. This probably leads to selecting program types that have had some identifiable dimensionality and some measurable effects in the past. We have just finished investing ten million dollars over five years in PVHS to learn something about various sponsored Head Start programs and their effects on different group of children. From that study facts were gathered about which programs were most readily implemented, which achieved the best effects on a range of outcome measures, and which differed most from each other. We might, as Fred Mosteller has suggested

(personal communication, September, 1973), construe the PVHS study as a preliminary, hypothesis-generating field venture, not high enough in its standards of scientific rigor to be called an experiment but leading to various initial ideas that should now receive more careful study in a controlled field experiment with true randomization of subjects to treatments. PVHS results should not be thrown away; the next evaluation should build on what has been learned. PVHS data should be used to generate a much more limited and careful design, asking more fine-grained questions. This approach no doubt would appeal to OMB and other sectors concerned about cost-effective use of evaluation results.

Second, and equally important, it seems only fair to assess Head Start in terms of what a *good* program can accomplish, on the grounds that once this baseline is established, dissemination can follow. In few areas of the federal government are programs justified on the basis of performance estimates taken from samples of performance under average or randomly sampled conditions. It should be sufficient to demonstrate that within certain budget constraints it is possible for some programs to have good results in field testing. These programs are apt to be those that have received most care in their design and formulation and are most ready to be implemented.

Third, it can be argued that a differential effects design enables pooling and therefore also enables us at a second level to answer the question of aggregate effects. This is in contrast to a general effects design, where if there are only slight aggregate gains we are never certain there were not strong selective gains in certain programs. This research wisdom can be combined with a parallel bit of political wisdom: If an evaluation has slight generalized effects as its primary finding, chances are it will have a negative influence on attitudes about the program. If it has large selective effects, policymakers may regard the entire program positively. It is important to consider political effects and make prior estimates about where findings are apt to be sizable.

Finally, two other practical questions can be asked in a differential effects study that cannot in a general effects study, both of them of interest to the policymaker: Which programs have fairly robust effects across different child groups, and in each program, which clusters of

Head Start objectives can be attained jointly? The latter question has never been explored. Every Head Start program has a list of goals in the domains of cognitive development, health and nutrition, parent participation, and so forth. It would be interesting to attempt a factor analysis of sorts, trying to figure out which goals tended to be attained independent of each other, which went hand in hand, and which were mutually exclusive.

What might a new differential effects study look like? Whatever the design it would need to be small, careful, and unpretentious. Two approaches suggest themselves, one of which would be interesting if OCD wanted to stress cognitive effects and their relation to program cost, the other if cognitive effects were a secondary consideration in the evaluation. The first study can be described as an analogue to a crop fertilization experiment, in deference to R. A. Fisher, with the hope that equivalence between Head Start treatment and fertilizer treatment will not be misunderstood. The study would ask the same question the agricultural agent asks when he plants a field with a single strain of wheat and then fertilizes each third of it differently. The first third receives no fertilizer at all, the second receives an average dosage (low-cost), and the third receives an intensive dosage (higher-cost). Does the intensive dosage merit the additional money, and in general does dosage seem to matter? By analogue the Head Start research design would have equal numbers of sponsored programs, randomly selected traditional Head Start programs, and non-Head Start controls. Questions would be those of cost and value added.

Such a design has a number of merits. First, it speaks directly to one question policymakers want to ask. They do not really want to know *whether* Head Start is succeeding, because if they look at the data they know that a rather good prima facie case can be made--as good as in most other national evaluations--that some Head Start programs are succeeding and others are not. Instead, they want to know what it takes to put a good program in the field and what is the magnitude of predictable difference between a well-executed but more expensive program and an average, lower-cost program. These questions are natural ones for the economist or cost-benefit analyst. A three-part design

comparing a sponsored program with traditional Head Start with non-Head Start controls would enable us to begin to answer them.

But there is also a problem with this idea. We must choose a single sponsor, or a very few sponsors, to represent the "high-cost" treatment. This means first of all that difficult judgments must be made about what is going to be called a coherent treatment--a single set of programs whose phenotypic variation in the field is not so great that they are no longer identifiably based on the same parent program. This could mean relying on a limited set of sponsors without all program types represented. One approach would be to explore in more detail the effects of the or or two programs that looked most promising, or had the most pronounced effects, in the PVHS study. The new study might attempt to learn more about success-related aspects of these programs that could be generalized or exported to other programs; it might also compare the programs with less expensive programs to gather baseline data on cost and quality. In such a study there would be no need to further dimensionalize centers, since the level of analysis would be the sponsor and not the type of program. OCD might choose as high-cost variations one structured program (e.g., Weikert Hi/Scope) that showed cognitive gains in PVHS, and one good program emphasizing social-emotional development (e.g., Bankstreet) with effects that probably were not given a fair chance by the PVHS battery. This kind of study would be a "mini-planned variation study," but with a new emphasis on the relation of costs and effects.

The other kind of experiment that comes to mind assumes much more limited interest in cognitive effects, using them perhaps as one of a number of variables in a design of managerial program variations. The OCD has recently initiated the Improvement and Innovation program, according to which all Head Start programs in the field must choose one of the following configurations:

1. Standard Head Start, center-based, five days per week.
2. Variations in center attendance for individual children--varying hours of the day and days of the week.
3. Home-based model.

4. Double sessions; two classes per day in a center.
5. Various locally designed options.

In addition there are a number of experimental programs or demonstrations--Home Start, Parent-Child Centers, Child and Family Resource Centers, programs for the handicapped, and a proposed demonstration in the area of "developmental continuity"--that will explore the articulation of preschool with elementary school programs.

Assessment of these program variations according to managerial criteria might be the most sensible evaluation strategy, with cognitive effects a secondary consideration. Children's attainment of minimal sufficient cognitive benefits might be compared, for instance, in centers with regular attendance, centers with variable attendance, and home-based programs.

- *Position* 3. It is sufficient to demonstrate strong cognitive benefits for *some Head Start children.*

Another kind of evaluation strategy would try to determine which child subgroups were benefitting most from Head Start experience, either in randomly selected programs or in certain sponsored programs. There is, for instance, a line of evidence in the preschool research literature (Bissell, 1970, Karnes, 1973, Weikart, 1967, 1972) suggestive that the principal benefits of preschool experience may be for children with Stanford-Binet IQs of 80 to 90, a full standard deviation below average. Many studies suggesting highest effects for this group are thrown into question because of inadequate procedures for controlling regression to the mean from pre-testing to post-testing, but in a least one analysis involving the PVHS data (Smith, personal communication, September, 1973) it looks as though such effects may be real even with proper statistical adjustments. In addition, preschool advocates like Weikart point out that many children enter their programs "at risk"; without preschool experience they would be likely to end up assigned to classes for the mentally retarded (MR) in elementary school. After preschool these children may show a lower rate of MR class assignment. It would

be impressive indeed to demonstrate that Head Start children of low IQs were less apt to require costly attention in elementary school than children of low IQs without Head Start. There may also be other subgroups for whom Head Start offers special benefits, such as physically handicapped children or children below a certain level in Standard English fluency.

One obvious question comes to mind: Why not combine a program effects and child-group effects study and do an evaluation principally intended to explore the interactions of program type and child subgroup? Helen Featherstone (1972) explored interactions in the PVHS data, and her work leads to a number of tempting hypotheses for further investigation. The answer to this question, I believe, is that even though interactions will be important to explore in any evaluation, it is probably not advisable to attempt an evaluation focusing in the first instance on them. This would require a sample of Head Start children differing in its subgroup proportions from the actual Head Start population, and it would necessitate a fully crossed design, which might prove impossible or greatly at variance with naturally occurring combinations of programs and child subgroups.

Another important variant of Position 3 has been espoused principally by B. White (et al., 1972; 1973) and others concerned about sensitive periods and optimal times to intervene in the child's early development. The central question for this group of researchers, and the one they maintain should be central for policymakers and well, is *when* to involve the child in a preschool program. Assuming a fixed amount of federal money available for preschool programs and elementary school programs, it may be the case, for instance, that the most important period to reach the child is not during the Head Start years at all but from 12 to 18 months. Burton White believes that by the time a child is four, when most children enter Head Start, it is too late to have much effect on cognitive and language development. It is also too late, he feels, for cost effective identification and treatment of basic deficiencies in sight and hearing, and other screenable developmental problems. Other maintain that early infancy is the most important time to intervene, and still others (e.g., Bereiter, 1972) have come

to feel there is nothing done for a child's cognitive development in a preschool program that could not be done as efficiently or better in the first year of elementary school. Many issues surrounding the relative costs and benefits of federal intervention at various age levels have yet to be resolved.

For present purposes we might limit the question somewhat, asking whether the current Head Start program is as effective with four year olds as with five year olds, or more effective. Within the two year age span of the Head Start child population, which children are benefitting most? There is ample evidence from the PVHS study that four year olds show fairly high gains in the more successful programs but do not continue to gain so dramatically in their second Head Start year if they remain in the program. This kind of information might be weighed, along with information about effects for five year olds entering for the first time.

It is appealing to argue that the government is committed to spending X dollars on educational programs for children and that the real question is the cost-beneficial one of when that money should be spent. But this argument has one glaring problem: To compare the effects of programs at different age levels we have to be able to compare assessments across years, which, as all psychologists appreciate, is extremely difficult. Imagine the nightmare of trying to compare average gains on the Shaefer or the Bailey in an infant program with average Stanford-Binet gains for the same group or a comparable group at age five. This recalls a point underscored years ago by Kagan and Moss (1962)--there is often little comparability of phenotypic behavior patterns from one age level to the next on a given dimension of personality, cognitive ability, or achievement. We might add that at different age levels there is apt to be even less comparability of program-related *changes* in behavior. Competence in a particular domain is reflected differently at different ages, and a theory of mental development and mental process is needed to link earlier behavior patterns to later ones through some enduring dimension of mental ability or some latent trait.

Theories of development can differ greatly in their implications for policy decisions. As an example, if a Piagetian or Montessorian theoretical framework is adopted, then motor acts at one age level are believed to instruct verbal and perceptual ones at a higher level. This means that we would seek some equivalence between ability with an embedded figures task at age four, for instance, and perceptual form discrimination at a later age. Most of us would agree that such equivalences or links probably do exist, but it would be impossible without a more sophisticated knowledge of mental process to devise a test battery for four year olds that taps the same latent dimension as for seven year olds. Even on the Stanford-Binet, which has various forms for various age levels, there are serious problems. First, the test is largely unreliable for children younger than five or six, and its predictive validity is notably lower for this group than for older children. Second, it is generally acknowledged to measure different factors of cognitive performance at different age levels, and consistency of score is more the result of a heuristic process of item refinement over the years than an indication that the same dimension of mental ability is being tapped across test levels.

Some statements probably can be made about the best times for economically identifying and curing such gross neurological impairments as poor eyesight and hearing. This is the area of intervention in which an age-related evaluation looks the best. For instance, it would be interesting to know the intersect of two curves plotted on an x axis of age and a y of cost, one presumably descending and the other ascending, the former being for cost of diagnosis and the latter for cost of cure. Bur primary concern in the Head Start evaluation has to be with effects of educational programs, not early detection of physical disabilities, and in this domain age comparisons of treatment effects are apt to be too difficult to pursue.

- *Position* 4. Cognitive effects should be a moderator variable. It is necessary to demonstrate cognitive effects for some Head Start programs and children, but such effects are not among the most important evaluation outcomes.

This position assumes that in the forthcoming evaluation there should be a more modest role for cognitive effects measures. The coming evaluation is of social competency, broadly defined, and consumers of the research are not going to expect a return to major emphasis on cognitive instruments. Instead, cognitive outcomes will serve as moderator variables of sorts, necessary but not sufficient to demonstrate that programs are accomplishing something. They will be allowed to fade into the background to the extent that other measures in other domains can be presented with high enough reliability and validity to command respect. By this logic, for instance, evaluation might administer one or two well-established cognitive instruments to the entire sample or to a randomly selected or stratified subsample. This would replicate what had been done before, indicating whether programs were doing as well as previous ones according to traditional criteria, but it would leave major emphasis and the burden of proof of positive effects on other kinds of instruments. This is a "maximum" strategy: Guarantee that a trustworthy baseline of moderator data is provided by a modest cognitive battery, and then offer a set of more high-risk, high-gain measures as the principal evidence of program effects.

The wisdom of this strategy cannot be assessed without knowing how much we are going to be able to trust new instruments in non-cognitive areas, and whether OCD will have sufficient time to develop new instruments. These issues need to be clarified.

- *Position* 5. We do not have the measurement technology at present to assess Head Start's cognitive effects.

A fifth position is mentioned briefly here because it is a kind of null-hypothesis, representing the views of certain skeptics in the educational research community and the community of psychologists. This position holds that it is folly to mount any new Head Start evaluation at all right now. Instead, we should return to basic research in program dimensionalization, test design, and quasi-experimental methods. We still do not know how programs differ from one another; we do not know how to measure their effects very

accurately in any domain and can't measure them at all in some; and we have not yet been able to mount a full-scale natural experiment without many confounded variables and violations of the rudimentary canons of good research. Perhaps the evaluations conducted so far have done nothing but squander the taxpayers' money, resulting in mistaken or confused inference about program effects, not helping policymakers at all in deciding whether to continue the program, terminate it, or work to strengthen certain parts.

These issues deserve a paper by some ardent psychometrician, policy researcher, or planner. But for practical purposes the position is not very appealing; it seems extreme to maintain that Head Start cannot be evaluated because of a lack of adequate evaluation technology. The predictable reply is that we must do the best we can with a difficult assignment, forging a new evaluation mindful of the vicissitudes of field-based educational research. The point of view that "more basic research is needed" is find for scholarly journals, but it does not help decisionmakers unless we are honestly prepared to say that we can learn nothing from an evaluation. Most of us would stop short of saying that.

WHICH POSITION SHOULD BE ADOPTED?

Choices about a sufficient demonstration of cognitive effects and the weight placed on the cognitive effects battery in the overall design of the evaluation deserve careful consideration. Which position shall be espoused? Much of the subsequent discussion of cognitive effects measurement is contingent on the choice

Current orthodoxy and the need for formal evaluation criteria have led many policymakers to espouse what I have called Position 1. It is safe to assume that this position remains the dominant preconception: (1) Head Start must demonstrate generalized effects and (2) Head Start is on much firmer ground if it can demonstrate effects that maintain themselves over time. It would be impossible to overlook these two success criteria completely in any forthcoming evaluation. Despite this, it also would clearly be a mistake to design an evaluation with either of these questions as the primary one addressed. To do so

would almost surely guarantee that Head Start did not have a fair chance in the evaluation. We know that overall effects are apt to be minimal even if statistically significant, regardless of measures employed. We also know from the heterogeneity of Head Start programs that any really interesting or systematic effects probably will not be universal. Finally, longitudinal effects are apt to be slight and expensive to trace even in the best Head Start programs.

The concern of policymakers with these two criteria should be acknowledged, putting the criteria forward candidly in the new evaluation proposal. But generalized effects and longitudinal effects should be shown to be of secondary interest. Primary interest should be elsewhere, perhaps in examining closely and more systematically than before the effects of limited subsets of programs or the effects of programs on limited subsets of participant children. The evaluation could adopt as its principal strategy some variant of either Position 2 or Position 3. This is in keeping with the proposition that PVHS was a preliminary exercise in hypothesis generation, and now we are ready for one or more carefully executed, smaller scale social experiments--real experiments, perhaps in the sense that children are randomized to program types, or at least in the sense that basic aspects of field research design are not overlooked and basic care in test administration is not forgotten. The design of such a study, whatever it examines, needs to be small and well-controlled without a host of sponsors, enormous sample size before inferences can be made, and confounded independent variables and uncrossed levels of the design. Keeping the study small and elegant will enhance its credibility immeasurably.

If one accepts instead some version of Position 4, then in the cognitive domain we need only resurrect some of the most reliable individually administered measures from past Head Start evaluations, make a judgment about how many children should be tested, and try to do the testing more carefully than before. Position 4 implies an emphasis on new non-cognitive measures and measures outside the domain of cognitive performance.

III. THE ROLE OF COGNITIVE EFFECTS MEASURES IN THE HEAD START BATTERY

In past Head Start evaluations, heavy emphasis on individually administered pre-tests and post-tests of cognitive performance has left many observers with the impression that the tail was wagging the dog. Evaluators, looking for *any* measures for young children with enough validity and reliability to be respectable in traditional psychometric terms, have returned time and again to the Revised Stanford Binet, subtests of the Illinois Test of Psycholinguistic Abilities (ITPA), and the Peabody Picture Vocabulary Test (PPVT) as measures of cognitive ability, along with certain achievement tests more directly assessing short-term Head Start learning (e.g., the Wide Range Achievement Test (WRAT), the Preschool Inventory (PSI), and the Deutsch NYU Booklets 4A and 3D).[1] Criticisms have been made of the weight such tests have received in all major Head Start evaluations; many feel that they do not fairly tap what Head Start programs are trying to accomplish, even within the cognitive domain. Some programs are more concerned with motivation and cognitive process than with cognitive performance. Others are presumably slighted because their curriculum does not "teach to the tests" or teach to the specific domains of competence tapped by the tests.

The choice to weigh cognitive effects heavily has been made largely by default, not because researchers thought cognitive instruments were more important or had higher validity in any absolute or theoretical sense, but because other measures, including those in the areas of affective growth, motivational or attitudinal change, and classroom behavior are usually so poor and of such low reliability that they cannot be taken seriously. This fact has not changed much in the past

[1] Tests mentioned in this and the subsequent section will not be referenced separately as long as they appear in Walker, Bane, and Bryk's PVHS test summary (1972). Copies of the tests and manuals in the form used in the PVHS evaluation can be obtained from the ERIC Clearinghouse for tests, measurement and evaluation, Educational Testing Service, Princeton, N.J. 08540.

eight years despite the heroic efforts of various test developers to devise measures of non-cognitive effects and observational schemes enabling a departure from paper and pencil or individually administered, clinical testing. For an excellent review of non-cognitive measures, readers are referred to Walker's 1973 book on the subject.

Two individually administered tests of cognitive effects have dominated the major evaluations of Head Start: the Revised Stanford-Binet (Form L-M) and the Preschool Inventory (PSI), developed by Betty Caldwell in 1965 especially for Head Start as a criterion-referenced measure of school readiness. The PSI was reduced from 64 to 32 items in 1969 to facilitate Head Start testing by paraprofessionals in abbreviated testing sessions. A third instrument included in the PVHS evaluation was the Deutsch NYU Test Booklets, two of which, the 4A and 3D, are straightforward achievement measures with fairly high reliability in Head Start measurement situations.

It is valuable to look at these tests to gain a notion of the "state of the art" in measuring Head Start's cognitive effects. Here they will be considered exemplary and typical. A fuller summary of cognitive measures can be found in excellent ETS and Huron Institute volumes (Educational Testing Service, 1968; Walker, Bane and Bryk, 1973).

Three general propositions should be considered. First, the evaluation of cognitive effects may *not* be the most important goal of a new Head Start evaluation. This issue relates to Position 4 in the previous section. It is deeply engrained in the tradition of Head Start that cognitive gains are a good basis for policy decisions, but few of those who care most about Head Start care principally that a child gain five points on an IQ test during the Head Start year. Most are far more concerned that children have a socially exciting experience, that they get involved with other children of comparable age, that they prepare themselves emotionally for school, that they be proud of themselves and their own capabilities, that they be aware of various facts and entities in their physical and social surroundings, and so forth. Indeed, for the most part those who have found cognitive gains a predominant criterion of program success have been researchers

worried about reliability and policymakers worried that the program is not going to lift children out of poverty by giving them a boost toward average or better school achievement and the later benefits they assume will flow from this. Arguably the evidence of the past several years points to how silly the researchers and policymakers have been in this emphasis and how right the parents, providers, and other taxpayers have been to ignore or protest it. Cognitive gains may simply *not* be that important. The coming evaluation may need to assert this and tell policymakers why the evaluation needs to be recast.

Just as in basic psychology, where a theory is replaced only when a better one comes along, there is a need to offer measures in other domains--notably health and nutrition, social competence, and influence on the family--that match the cognitive measures in credibility while supplanting them in importance. This credibility need not involve levels of psychometric validity as high as would be demanded in a laboratory setting. Face validity is sufficient. But they do require that measures be reliable and that measurement be feasible in field-testing situations. In addition, if new instruments are to be used, the team finally responsible for analyzing the data--those who will write the final report--must be involved with the evaluation early enough to agree to the idea of weighing these measures heavily. In other words, it should not be allowed to happen as has been the case in the past that those conceiving the evaluation are an entirely different group from those analyzing the data, with a different conception of which instruments to stress.

In selecting instruments, it is also important to realize that there is a difference between *psychometric* validity and *political* validity. This difference helps explain, for instance, why the Stanford-Binet has repeatedly been chosen as a Head Start outcome measure. No psychologist who knows the Binet and also knows what Head Start is trying to accomplish feels that using this test to evaluate the program makes much sense. The instrument was designed to measure a unitary, stable trait of general intelligence, not to measure program-related achievement or increases in performance in specific realms of cognition. Items are not samples from larger pools representing theoretically

coherent dimensions of mental ability, there is no subscale structure, and the predictive validity of gains on the test, as against measurement in the one-shot testing situation, is unknown.

A measure of Head Start's cognitive effects ideally would be quite different, telling us (1) which dimensions of cognitive performance Head Start was able to influence, (2) whether gains on these dimensions had face-valid importance or predicted to better than expected outcomes in later schooling and later life, and (3) whether these "leverageable" dimensions, on which Head Start could have some effect, could be linked causally to specific curricular components of the Head Start program. This means, as a fanciful example, that it would be nice if we knew gain in the area of digit-span memory was one of the effects that Head Start often had; that such gain resulted in some benefits during kindergarten or for the Head Start child's immediate life before kindergarten (for example, it generalized, enabling the child to expand short-term memory in a number of other realms by a new chunking strategy); that a gain in this area predicted to greater competency for the Head Start child in later schooling; and finally, that one area of the Head Start curriculum, in this case a specific set of structured drills, taught this particular skill. Knowing that much, we would indeed be on the track toward a theory of instruction. We would have some idea of how to appraise cognitive gains. As a less ambitious goal, even if we knew nothing about generalizability of acquired skills or about how they predicted to desiderata in later schooling and later life, and even if we did not know exactly which aspects of the Head Start program caused shifts in cognitive performance, it would be enough to show that *some* face-valid gains in areas of obvious practical interest could be effected and then to characterize these areas.

The Binet does not enable us to talk about any of these things. It certainly is not a good measure of Head Start effects in the ambitious sense outlined first. We know nothing from it about dimensions of cognitive gain, since items on each test level are not representative of various domains of cognitive performance. Nor does the test tell us anything about the predictive validity of gains. In fact, since the instrument was designed to measure a stable latent trait, any gain

arguably must be interpreted as a reflection of low test reliability.

The Binet also is not good in the more circumscribed sense of a criterion-referenced achievement measure. Its items are not intended to tap skills that Head Start teachers feel are the most important ones to teach, and they often have little apparent connection with actual kindergarten-related skills or first grade skills. Indeed, they are chosen to measure something that is *not* teachable--a permanent characteristic of general intelligence--rather than skills that can be readily acquired. In addition, of course, the Binet is culturally biased--it was designed for a middle-class white population and normed on this population. To use it on Head Start children--and to be oblivious of its differential validity for different groups by geographic region, ethnicity, and so on--is to ignore test aspects that Terman and the other designers of the test would never have overlooked themselves.

Why, then, have preschool evaluations persisted in using the Binet? The answer, I think, is that the test has political validity; that is, it has a certain credibility among researchers and policymakers simply because it is known (by name, if not by psychometric pedigree) and has been used traditionally in assessing the intelligence of young children. An IQ "gain" has a mystique about it--it suggests that one fixed level of intelligence, or "g," has been replaced by another. As we know, this interpretation is largely spurious. But it is the public notion, and it is firmly enough entrenched that many researchers and others turn to the Binet almost reflexively rather than fight the more difficult battle of trying to explain to an audience of non-psychologists why the test is inappropriate. In addition, of course, the Binet is administered by trained testers, many of whom exist around the country already, and although it is more expensive to administer than most other tests in past Head Start batteries (the PVHS evaluation could afford to give it to only half the sample), its level of reliability in very short-interval test-retest situations and its inter-rater reliability have probably been higher than for other tests (Walker, Bane and Bryk, 1973). This too has tended to give it credibility.

Political validity is important and should not be ignored altogether. But in a future evaluation it is probably important to educate policymakers and others about the inappropriateness of certain time-honored instruments when these instruments are applied in the context of Head Start evaluation, rather than cater to predispositions about the tests that "really matter."

PROBLEMS WITH THE PAST MEASUREMENT OF HEAD START'S COGNITIVE EFFECTS

It is useful to summarize certain recurrent shortcomings of past cognitive effects batteries. None of the problems mentioned here are easy to remedy; perhaps some of them are inevitable, given the limitations of current instrument development. But all of them are likely to recur if no special efforts are made to avoid them.

1. *Low quality of the field operation for test administration and other aspects of data collection.* This problem is first on the list. It cannot again be overlooked without serious consequences. In past Head Start evaluations, even the most rudimentary aspects of data collection have gone wrong. We have tried to collect too much data for too little money, and the results have been appalling. Those who have worked with the data have never been sure which test results could be trusted, even among those that should be most reliable and reliably administered. Examples of oversights abound; it may be useful to mention a few:

- Test-retest and inter-rater reliabilities for administration in the Head Start setting generally have not been reported, and in two instances where such information has been gathered, the ETS Longitudinal Study and the Huron Institute reliability study conducted as part of 1971-72 data collection (see Walker, Bane and Bryk, 1973), results have been unacceptably poor on many cognitive measures.
- In the PVHS study, some tests were administered by trained and specialized testers (e.g., the Stanford-Binet, the 8-block Sorting Task), but most were administered by community paraprofessionals who received only a short briefing in how to give them (e.g., the PSI). Many of the testers were not

uniform in techniques for establishing rapport with children, presenting test materials, or scoring children's responses.

- On the PSI and most other cognitive measures, identical forms were administered at pre-test and post-test, with no alternate forms or item sampling procedures. Practice effects seem likely, and it was fairly easy to teach to the test.
- In some PVHS sites, there was a single tester pre and post for certain children and different testers pre and post for other children. For instance, in one site on the Binet, control children had the same tester pre and post and experimental children did not. It was also noted that experimental children had unusually low pre-test scores in comparison with controls. Since the children were from the same preschool population, this suggests that perhaps there were selection effects or unreliable pretestings. How should such data be interpreted?
- Until quite recently there has been no scoring of individually administered tests for response style. Now at least the Hertzig-Birch scoring scheme has become part of the PVHS and ETS studies. But most testers in the field, especially those administering tests other than the Binet, have never been trained to code response style according to the Hertzig-Birch scheme or any other, and without sufficient training for testers this kind of coding is apt to be of low inter-rater reliability. Much of the data collected so far, while suggestive, may not be worth analyzing because it is of poor quality.

Problems like these are apt to occur for a very simple reason. At the outset, those conducting the evaluation have the best of intentions; a limited test battery is selected with some efforts made to ensure reliable test administration. Then for political or other reasons, as the evaluation progresses a new measure simply must be included in the battery, or a new site simply must be added, or a deadline for initial test administration simply cannot be met with enough time for training of testers. The integrity of the field operation is gradually undermined by decisions subsequent to the original plan for data collection.

One does not have to be an organizational theorist or have any experience with large-scale data collection efforts to know this much:

If an evaluation does not collect good data no one will believe whatever it concludes. This means that we should look carefully at the Sesame Street effort and other field assessment efforts of high quality, trying to emulate them. Probably we should estimate initial sample size on the basis of higher than anticipated cost estimates for test administration, and then multiply that estimate by two or so to get a fair approximation of real cost!

2. *Poor theoretical rationale for the individually administered cognitive tests in the Head Start battery.* Any adequate treatment of this problem could fill a good-sized book. I will only try to spell out some unresolved issues. In general, psychologists put all of these issues under the rubric of validity questions. In the present context we are not concerned, as we were above, with tactical questions about sufficient magnitude and duration of effects. Instead we are concerned with "truth" questions about what we as researchers have actually demonstrated when we find an effect, usually in the form of a transition from time 1 to time 2 in children's performance on an individually administered test. I will list some persisting confusions in the measurement of Head Start effects that can be traced, I think, to confusions surrounding the theoretical rationale for the instruments themselves as they are applied in the context of Head Start.

The first problem is what might be called *the assumption of initial incompetence.* This is the notion that Head Start children begin at a level of cognitive functioning that is somehow inadequate, "deprived," or ignorant and progress to some level of competency. Robert Hess (1969) has created an interesting taxonomy of "models of deprivation," arguing that implicit in most people's thinking about compensatory programs is some notion of a mental state pre and post--some conception of what deprivation and non-deprivation look like. Most such conceptions are based on an operational definition of competence, publicly defined, often related to performance expectations in the schools. Hess points out that these models may *all* be wrong or bigoted, a point also made convincingly by Cole and Bruner (1972), Labov (1972), and others. These researchers suggest that we often err because we do not begin as anthropologists, assuming a position of cultural

relativism. The child's mental state pre may be just as sophisticated as his mental state post; the only change brought about by Head Start may be to introduce him to a set of role expectations, norms, patterns of acceptable verbal conduct, and so forth that prove adaptive for him in getting from his own cultural context to that of the school and the so-called dominant culture. His versatility is increased, not his capacity.

Most of us are familiar with this point and I will not belabor it, except to ask that we consider its implications for the cognitive effects battery. It suggests that perhaps we need a new and pluralist conception of the appropriate end-point for Head Start activities; for different cultural groups, different sets of goals may be appropriate. In the past evaluations we have avoided this issue because it has seemed to lead rapidly to a test battery comprising culturally unique and non-comparable measures. This remains a danger, but in the past we have gone too far in the other direction. The selection of tests and the development of special Head Start tests have made the assumption of a uniform initial competence. None of the tests has been able to tap culturally relative patterns of mental performance at either pre-test or post-test. The Binet is notorious for cultural bias, the PSI was developed explicitly to be culturally biased on the theory that this was the fairest way to assess level of preparation for middle-class school situations, and other measures also show no particular ability to tap skills that a child may bring to Head Start. There is much to say for choosing measures--or developing new ones--that ask how skills the child brings to Head Start become transmuted over the year into skills he can use at school or in cultural contexts outside his own.

If we want to base tests on the cultural relativist model of cognitive development and program effects, one way to proceed would simply be to look at interactions of child group, test, and program type, as Featherstone (1972) and others have done. This is certainly a step in the right direction, but it does not attack the problem at the level of the tests themselves, exploring whether magnitude of shifts in score on a single test can ever be a fair yardstick of

program success for various cultural groups.

Another confusion surrounding theoretical rationale is *lack of clarity about the relative importance of cognitive-developmental as against behavioral Head Start goals.* A particularly interesting exchange on the question of appropriate cognitive goals for preschool programs took place in Interchange in 1970. It was between Lawrence Kohlberg and Carl Bereiter, with Kohlberg arguing the position of the stage-sequential Piagetian and Bereiter the position of the behaviorist. The discussion has direct bearing on the question of theoretical rationale in choice of tests and test construction. Bereiter tried to make the case that there was no point in measuring anything but face-valid changes in skill levels and other readily perceptible dimensions of cognitive performance that would be adaptive in school, because we simply did not have an adequate theory of intellective functioning or intellectual development to allow us to see other kinds of changes, in this case the attainment of concrete operations or the extension of concrete operations into some new domain, as an important achievement. Bereiter maintained that on theoretical grounds the Kohlberg point of view was suspect because the child presumably would attain concrete operations anyway sooner or later and there was no point in hastening the process, even assuming it could be hastened. He also maintained on empirical grounds that we had no valid or reliable measures to tell us when a child has successfully extended his capacity for concrete operational thinking into new realms.

Kohlberg responded that from the standpoint of cognitive development, the kinds of "gains" Bereiter was left with as a residue were trivial after he eliminated all that he believed could not be discussed because of inadequate theoretical rationale. To teach a child numbers and letters, for instance, is not an important enough task to merit the efforts of a federal program (especially, he might have added in hindsight, since Sesame Street seems to be doing it so much more cheaply). This is a specific skill, like learning to swim, and the fact that children learn it says nothing about enhanced or generalizable cognitive functioning in other domains, or in the future. For Kohlberg, face-valid and directly school related skill acquisition is

not a sufficient goal for a preschool program.

Without doing justice to all aspects of the Bereiter-Kohlberg debate, I only want to suggest that reasonable men differ about whether we should have cognitive-developmental or strictly behavioral goals for Head Start, and that these differences depend on their own theories of development or their judgment that some theories are worthy of influencing choice of goals while others are not. The measures included in the Head Start battery reflect such theoretical or atheoretical predispositions, whether implicitly or explicitly. The point is fundamental: Every Head Start measurement strategy is based on a theory of cognitive growth. Thus the educational policy researcher finds his own measurement strategy no stronger or weaker than the basic developmental theory on which it is founded.

In the past there have been two main implicit biases reflected in the measures selected. The first has been theoretical but curiously inappropriate and unlike Kohlberg's--researchers have chosen tests like the Binet designed to measure a stable trait of general intelligence. The second has been wholly atheoretical--researchers have chosen criterion-referenced measures of skills directly involved in kindergarten and first grade competence. Neither approach has been satisfactory, the former because it does not tap any growth function of the sort that Kohlberg would emphasize and the latter because readiness tests have been too sparse, too culturally biased, and too little able to demonstrate concurrent validity in correlating with other areas of competence even in the short term.

There are two directions we should consider in advancing to a clearer theoretical rationale for the tests in the Head Start cognitive effects battery. The first is toward theoretically oriented tests, which focus on patterns of cognitive growth instead of cognitive stasis. Some of the Piagetian clinical assessment techniques and Kohlberg techniques for assessing stage-sequential development and horizontal *décalage* may be worth exploring (see, for instance, Green, Ford, and Flamer (1971), and Marcus Lieberman's (1970) thesis on a maximum likelihood estimation of stage-assignment for children according to performance on various Piagetian tasks). It would also be valuable

to develop or select more thoughtful criterion-referenced achievement measures.

A third problem surrounding the theoretical rationale for test selection is *persisting confusion about whether we should use criterion-referenced or norm-referenced tests.* This point has two aspects, the first related to the question of what we are trying to measure, and the second concerning when it is appropriate as a matter of testing theory to use each kind of measure. Norm-referenced tests are designed to show where an individual child's performance stands in relation to the distribution of performances for all individuals in some appropriate referenced group. Scores are reported, therefore, as they relate to the mean performance of all children at a given age or grade level or in terms of a percentile rank. Such test are developed by choosing items from a larger pool of face-valid items according to intermediate item difficulty, high item-scale correlation, and theoretical coherence in the dimension they measure. Items that are too easy or too hard are excluded because they do contribute to the variance that can be explained by the test. Criterion-referenced tests, in contrast, try to compare an individual's performance to some set standard--hence "criterion"--rather than to the performance of a reference group. The basic idea is to reach agreement on what constitutes acceptable performance in some area and then to select items from an item pool that either are highly correlated with some other direct measure of such performance or somehow themselves represent an agreed test of such performance.

In the case of color recognition, as an example, it would no doubt turn out that if a child could identify the colors of four crayons chosen at random from a box, this would correlate highly with his ability to identify colors other than the ones actually selected, and in various objects other than crayons. In some cases, passing the test itself might be sufficient demonstration of attaining the criterion. We might ask the child to interact with peers in a classroom, for instance, which in itself is identical to the competency expected of the child later. This literal achievement of a criterion is what Kohlberg (1970) means when he refers to the "industrial

psychology" approach in testing. By analogue, if an adult has to operate a particular machine, it goes without saying that it is a sufficient test of his ability to sit him down with it and watch him perform. In either case, the one where criteria correlate highly with competencies or the one where they *are* the competencies to be domenstrated, criterion-referenced items are selected according to which ones and how many of them have to be passed before it can be reliably predicted that the individual will be able to meet the criterion, or perform acceptably. The test is designed with a threshold in mind, above which adequate performance can be expected.

In principle, of course, there can be a rank ordering of criterion performance of children from worst to best, and the criterion-referenced test can be converted into a normed one. But the idea of an absolute confidence level rather than a relative one is quite different in the first instance, especially in its implications for item selection. Norm-referenced items are chosen first on the grounds of intermediate difficulty and scalability, along with face validity. Criterion-referenced items are chosen with external validity as the prime consideration. According to a non-referenced item selection strategy, for instance, a particular embedded figures task might have been included on the Stanford-Binet because it is of average difficulty for a particular mental age group and it is highly correlated with other items on its six-item age scale. It might also correlate well with later composite IQ score and load on the same factor as a secondary consideration. But there would be no theoretical reason why it need correlate highly, for example, with increases in understanding teacher requests in kindergarten or first grade, with specific knowledge of useful school-related facts, or with other practical aspects of cognitive attainment.

If we want to find out about short-term achievement, perhaps controlling for IQ, then we choose items initially on rather different grounds. We select them first for external validity--what measures directly predict best in the short term or to the competencies preschools teach? Items selected with this practical purpose in mind can then be scaled, but what is desired is a test with high item

difficulty or so le-difficult at pre-test and variable item difficulty at post-test, such that there is maximum homogeneity of scores at pre-test and maximum variance at post-test. This guarantees that our "criterion score" is sensitive and that fine gradations can be made among children regarding whether or not they attain it.

The PSI is close to being a criterion-referenced test even though it has not been interpreted or standardized as such in the Head Start evaluation. But it is culturally biased, it is intended only as a measure of readiness for middle-class kindergarten, and it does not even have any great face validity as a measure of school readiness. The items are too few and too arbitrary. Curiously, its value has supposedly been legitimated by demonstrating its "concurrent validity"--how highly it correlates with the Binet and other tests in the Head Start cognitive effects battery. This is precisely how it should *not* be validated if it is tapping different things from those tapped in a general intelligence measure. We should, I think, follow the lead of the Sesame Street test developers and carefully consider a foray into unabashed criterion-referenced testing (Ball and Bogatz, 1970; ETS, 1974).

A fourth problem is that there are *persisting technical difficulties in the statistical analysis of gains*. During the analysis of the PVHS study Marshall Smith and his co-workers filled two fat notebooks with articles on change scores and how they should be treated, many of them contradicting or rebutting each other. The evaluation group never did feel confident enough of the issues to select a single technique, instead looking for consistencies among outcomes using a number of different techniques. This made sense from the heuristic standpoint, but it is hardly reassuring for those who would like precise estimates of effects. We need a better statistical technology for analyzing gains with appropriate covariate adjustments, or else we need tests explicitly designed to measure changes in cognitive skill attainment. Problems of analysis are clearly related, of course, to the earlier-mentioned problems of inadequate developmental theory.

3. *Poor theoretical rationale for the observation and interview measures in the Head Start battery, with the additional difficulty*

[illegible] If the picture is gloomy among the individually administered tests of cognitive effects, it is even worse among classroom observation and interview techniques. Among the classroom observation measures, there has been a tendency for researchers to spread their nets wide, attempting to capture all that is going on in the classroom rather than exploring particular and limited aspects of classroom process. In so doing they usually have captured very little. Classroom observation techniques need to be more closely orchestrated with individually administered tests to explore the performance of cognitive competencies whose capacity is assessed in the individual testing situation. It is confusing and difficult to dredge [illegible] hypotheses from data generated by the current measures, and analysts of the PVHS data have tended to ignore them. Classroom observers also have often not met acceptable levels of inter-judge reliability in field testing.

Teacher and parent interviews, while more reliable, are of lower validity as measures of Head Start success, since neither parents nor Head Start teachers are unbiased observers. Blind situations have not been engineered with independent observers in the home or the center, which might enable competency ratings or convincing interview [illegible] Head Start effects.

The observational instruments with high reliability either from individually administered tests but led to a greater emphasis on the individual tests than is desirable. We need to consider whether this emphasis is inevitable. Some feel that it is, arguing that there is no point in trying to obtain high reliability in anything but careful, one-to-one tests with individual children. The recent evaluations of [illegible] most of their conclusions on a few individually administered cognitive measures--the only ones the investigators felt could be trusted enough to interpret.

But this reduction of measures tends to sacrifice, for the sake of reliability, much of what Head Start personnel think should be explored. Unforeseen cognitive effects will have to be more amply examined in the forthcoming evaluation. Even if there are trades to be made between emphasis on individually administered tests and less

of reliability or added expense, it is still not clear that we should opt for individual measures to the exclusion of others.

WHAT IS TO BE DONE?

As a practical matter, the next Head Start evaluation cannot radically revamp all current measures and techniques for assessing child performance. Many issues of test development are long-range ones, for which solutions are likely to emerge only after years of careful basic research and incremental test development. In particular, it is unlikely that Rand or anyone else, given a six to twelve month period, will be able to devise entirely new item banks for individually administered tests and entirely new observation schemes. Instead it is far more likely that the new cognitive effects battery will involve imaginative scavenging from parts of tests already available and resourceful application of various tests now being developed. It is important to decide what work to cut out in the creation of a new test battery. If we honestly feel that no measures now exist that are appropriate for Head Start evaluation, then perhaps it is best to deal with that now rather than later, rejecting the notion that a new evaluation can start within a year. If however, there are certain measures currently available, others that are appropriate in part, and others that could be developed without too much effort, then we can proceed within the anticipated time frame. As a third possibility, of course, some may feel that the current PVHS battery is perfectly adequate and that the problems I have cited are occupational hazards that any federal evaluation must undergo, without major policy consequences.

I would like to state biases regarding these options. I think two strategies deserve consideration in planning the forthcoming evaluation, one I will call *validity based* and the other *reliability based*. A validity-based strategy follows from the point of view that the principal problem with past evaluations has been a validity problem: We were not measuring what Head Start was actually doing. Those who would support this strategy feel that any new cognitive effects battery for Head Start must represent a significant departure from past batteries. If it does not have new conceptual foundations, however well it might

be administered, it will show us little that we do not already know. This strategy anticipates considerable time for the development of new instruments, perhaps as much as two years before the end of field testing and the beginning of the national evaluation.

A reliability-based strategy assumes a different point of view: In the past we have failed not so much because instruments were invalid but because they were not administered carefully enough. The task at present, therefore, is less one of devising new measures and more one of designing and administrating the new evaluation so that the integrity of the child performance data can be assured. We can begin a new evaluation soon if we can accurately estimate the cost of reliable test administration and design a study that permits reliability within known budget constraints.

Each of the strategies leads to a different conception of appropriate next steps in planning the evaluation. The validity-based strategy suggests we should begin by recasting our conceptions of cognitive effects, letting contracts for the development of new measures, and arranging field tests for instruments as they are devised. The reliability-based strategy suggests we should start by identifying the best currently available measures, estimating the cost of administering them in the field, and considering various designs for an evaluation in which they would be used.

Predispositions about the best sequence of steps in designing the study itself also depend in part on whether planners are validity-oriented or reliability-oriented. A validity-oriented group is apt to recommend the following steps: (1) Isolate areas of potential program effects; (2) devise instruments that assess change among Head Start children in these areas; (3) estimate a sample size and design large enough to enable valid inference about cognitive effects on these instruments for the total population and important subgroups; and (4) compute the cost of the study, cutting back the design in certain areas if cost is too high. One implication of this sequence of steps is that the validity-oriented planner will end up with longer phases of planning and preparation and will tend to think about overall budget only after he has decided what needs to be measured, what instruments must be

developed to measure it, and how large a sample it will take to assess cognitive effects for various groups of Head Start children.

With a reliability-orientation, the sequence of steps in planning the evaluation would be quite different: (1) Estimate the total budget available for the evaluation and select a minimal set of the best currently available instruments; (2) compute the per-child cost of administering the battery with sufficiently high reliability in the field; (3) divide total budget available by per-child cost of test administration to arrive at sample size; (4) with sample size as a constraint, figure out what design is both feasible and sufficient to answer important policy questions. Unlike the validity-based approach, this one is conservative, beginning with present testing technology and budget expectations. These are seen as prior constraints in deciding sample size and selecting questions the evaluation can afford to ask.

The validity-based and reliability-based strategies as I have sketched them are archetypes, primarily useful as schematic ways of thinking about planning choices. No doubt the actual planning of the evaluation will reflect both strategies. But it is also likely that in the planning process one of the two modes will predominate, exerting marginally more influence than the other. It would be useful for the OCD to decide in advance which of the two makes more sense as a primary strategy, given current administrative and political realities.

If money and time are to be spent on developing new instruments, this is a major commitment. Probably a certain amount of new instrument development is important, but I believe it would be wise for the evaluation design team to devote most of its energies to thinking about how to execute the evaluation well, with a first-rate field operation. I would be happy if the new evaluation administered only four to six cognitive effects measures of various kinds, plus a few measures of related outcomes or processes to enable concurrent validity estimates. In general, reliability of test administration and high face-validity are less ambiguous and more realistic as goals than high predictive validity; or, as one policy-analyst phrased it, "observational power" is more salient than "predictive power."

LIKELY REALMS OF COGNITIVE EFFECTS

This section offers a list of cognitive effect domains, discussing each and considering whether it is an area where Head Start effects are likely to be found. The five domains in the present typology were originally proposed by Sheldon White (personal communication, September, 1973).

1. *Norm-based kindergarten or first grade readiness.* The cognitive dimensions of first grade readiness have received much attention in the past, not only in Head Start evaluations but in other school-related testing. The Metropolitan Readiness Test, like the PSI, is a well-known standardized achievement measure designed to assess preparation for school. There also are such tests as the Meeting Street Inventory (Hainsworth and Siqueland, 1969), intended to screen "high risk" children who, because of some minor physical or behavioral disability, may require special attention when they enter school. Some of these tests predict reasonably well to kindergarten and first grade achievement scores, although partial correlation with achievement in kindergarten and first grade is likely to be less impressive when IQ is introduced as a control. Another problem is that these tests are designed to be administered only once; their characteristics in the pre-test to post-test gain situation, especially where alternate forms are not available, are not so clear. On a test like the PSI, with only one form, practice effects seem inevitable.

In general, these tests have the one major advantage of face validity. Cognitive performance items and behavioral objectives that relate to first grade readiness are easier to gain consensus about than items and objectives in other domains. This is especially true if test developers do not anticipate a diversity of kindergarten or first grade situations, instead contenting themselves with measuring what is required for competence or adaptability in an average, white middle-class kindergarten (see, for instance, Caldwell, 1967). We have seen that this is a questionable assumption.

If there is any concern with diversity of child populations, geographic regions, and kinds of kindergartens or first grades, the task of developing face-valid measures in this domain becomes more difficult;

but it is by no means impossible. It makes sense to be concerned with such diversity, looking for tests and test items with high face validity for the actual child groups considered. This may mean developing various equivalent measures of first grade readiness for different ethnic or geographic subpopulations.

One of the most interesting areas for school-related measurement is a time-honored one: reading readiness and readiness in numeric skills. These need to be tested in a number of ways, not only with letter and number recognition tasks but also with techniques borrowed from the psychology laboratory. To measure decoding skills we might, for instance, ask children to distinguish between letters of the alphabet and Gibson's (et al., 1962) experimental stimuli. It would also be valuable to consult with the Children's Television Workshop team assessing effects of the Electric Company. In addition, observation of increased interest in reading in the Head Start classroom or at home might be an important face-valid indicator.

In the area of numeric skills, it seems wise to consult with the group working at MIT and the Educational Development Center in Boston on a new television program to teach math skills, analogous to the Electric Company in the area of reading skills. This group is devoting its efforts to discovering teachable components of numeric reasoning in young children. In addition, Piagetian measures should be carefully explored (Green, Ford, and Flamer, 1971). Piaget is especially convincing in talking about the shift from pre-operational thinking to concrete operational thinking, and the implications of this shift for the child's notion of reversibility, class inclusion, and other aspects of logic and inference. Many feel this is the kind of "math" Head Start should be teaching.

Another approach to readiness assessment involves having elementary school teachers observe Head Start children in actual kindergarten or first grade classrooms. This procedure already exists in many school districts, where each child spends a trial half-day at elementary school the spring before entering the school. Perhaps this technique could be used with appropriate blinds to see if teachers could distinguish between the readiness of Head Start children and that of others,

using a teacher checklist or rating scale.

In general, school readiness remains a *good* area in which to evaluate Head Start's cognitive effects. Face-valid measures are easier to develop than in other areas, and short-term effects should have an unambiguous meaning that policymakers and other nonresearchers can appreciate.

2. *Theory-based developmental shifts.* If school readiness goals are fairly clear, theory-based developmental goals are ambiguous and difficult to rationalize. They are tempting to explore, because we know that from five to seven the child undergoes a dramatic transition in many dimensions of cognitive process, emerging a qualitatively different thinker at the end of this period than he was at the beginning. But theories of development do not tell us much about how and where to teach. Even if we could adjudicate among them, differentiating, for instance, between the claims of the Piagetians and the claims of learning theorists, we still would not fully understand their implications for pedagogy. This point was made by John Dewey years ago (1900), and it has cropped up again in the debate about the fallacy of trying to "accelerate" Piagetian stage-sequential development. Even if we have a clearly articulated, norm-based theory of development, we know little from it about those specific teaching interventions that will enhance development or predict to a fuller development. It is almost as though a norm-based theory of development is one kind of predictive entity and an intervention-based theory of short- and long-term effects is another. The latter does not follow automatically from the former.

The Bereiter-Kohlberg Interchange debate (1970) again is instructive, where each theorist feels the other's goals for preschool are trivial--not deserving of a major federal program. Kohlberg believes specific skill acquisition is easy to effect, but it is reversible and not of enduring importance. In any event, it could be done without all the educational trappings of a preschool program. Bereiter feels that stage-sequential development is an elusive notion. We do not know whether we can influence it with an educational program. Even if we can it is not clear we should bother to do so, since the

child sooner or later attains concrete operations regardless of early intervention. Moreover, stage-sequential goals cannot ever be satisfactorily translated into behavioral objectives.

Each of the theorists also make concessions to the other, however. Bereiter admits, and has increasingly been on record as saying, that any specific skill worth teaching in the preschool could as easily or more easily be taught in the first grade. In this sense he believes preschool is not cost-effective. Kohlberg agrees that the goal of trying to accelerate stage onset is not worthy; he feels the real effort should be in trying to avoid inexcusably *late* stage onset in some children, and to bring about wider horizontal *décalage* of the child's present stage.

An evaluation using theory-based developmental criteria could adopt one of three strategies. The first is to *rely heavily on Piagetian measures* and make Kohlberg's preschool goals pre-eminent. There are certain areas of development where this strategy would be wise. The transition from egocentric to sociocentric activity on the part of the child, for instance, is clearly important in school, home, and neighborhood situations. Such a shift would have obvious face validity. Piagetian measures also would be useful in the area of numerical skill development, another link between theory-based and school-readiness criteria of program success.

The second approach is to *sample competence in a number of domains of basic cognition* as indicators of developing thought processes in the child. This procedure might enable us to explore certain five to seven growth dimensions that, although largely maturational in their etiology, establish a backdrop for achievement gains in the school-readiness domain. If such processes are monitored, it is important that some of the instruments measuring them be non-verbal. It is also important that such measures use stimuli familiar to children of different cultural groups. Sampled competence domains might include:

- Short-term and long-term memory, with special attention to memory *span* and *transformations*, while retaining information in short-term memory. The child might work on a problem while being required to keep two other things in mind.

- Perceptual detection, tapped by embedded figure tasks, reorganization of familiar objects, upside down transformations.
- Using a code. The child might be asked to learn a six digit glyph code and then *apply* it in some familiar situation.
- Conceptual and perceptual equivalence tasks. Here there are lots of examples on current tests, some of them bad because they involve unfamiliar objects or are confounded with verbal response requirements. One good beginning is the ETS enumeration task; the second half of this test combines a recognition task (similarities and differences) with a Piagetian perceptual inference task involving mental transformation of a picture.
- Simple problem solving and other inferential tasks.

In general, it is probably a mistake to sample dimensions of basic cognition except as a means of acquiring limited baseline data about maturation-related changes. These dimensions are important, but Head Start cannot reasonably be expected to have much effect on them. If Head Start children experience gains on basic cognition items, any improvement beyond the purely maturational is apt to stem from better rapport with the tester, motivation in the testing situation, or other incidental factors.

A third theory-based approach might be based on *some theory of sensitive periods*. The Montessori approach, for instance, espouses the theory that earlier motor training and training in perceptual discrimination is a necessary prerequisite to later competencies, perhaps not in the sense that it represents an irreversible critical period, but at least in the sense that a motor substrate can be laid down more easily at an earlier age than a later one and has to be present before some subsequent capacity can develop. If the hand instructs the eye, according to this line of reasoning, then let us instruct the hand at the right moment.

This point of view undoubtedly has some truth to it. The idea of prior motor schemata emerging to become conceptual schemata later is somehow right, although far too impressionistic even in the most fully articulated theories to do more than satisfy our yearning for aesthetically pleasing constructs or whet our appetite for more concrete and

testable ones. We certainly do not know at present how to link earlier instruction with later cognitive benefits--how to massage the black box in a certain way and a year later have it reward us with some desirable, newly established competency. According to some theories we are not even sure about how to verify the existence of the later capacity. Without a more convincing theory of sensitive periods, and one that can be easily operationalized, we are probably ill-advised to consider measures intended to assess prerequisite early learning unless that learning has face validity as well as theory-based significance.

3. *Changes in cognitive process.* Head Start may have its most dramatic effects in the area of cognitive *process.* This realm is a promising one for exploration in the next evaluation. Investigation of cognitive process shifts also overlaps conveniently with consideration of social competency, which the OCD has recommended as a principal focus of the new evaluation.

There are three facets of cognitive process; each merits attention. One aspect is quite narrow, having only to do with *response style and response coding in the individual testing situation.* Individually administered tests can be coded for response style as well as correctness. Little is learned by a coding of correct-incorrect as compared with some coding scheme that can register shifts in the child's approach to the task and means of solving it. Some aspects of cognitive style are closely related to the thinking involved in solving the problem itself, such as the search strategy the child uses in trying to recall something he was asked to retain in short-term memory. Others are related to impulsivity or reflectivity in solving the problem, the child's technique in probing the tester to elicit clues, and the child's global reaction to the testing situation. Although the non-task-related response style factors may not generalize beyond the testing situation, chances are that cognitive process factors related to problem solving itself will. These are the ones we should be attentive to in the individual assessment situation.

A second aspect of cognitive process is the *transfer and actual performance in a larger behavioral context of new skills or capabilities that have been demonstrated in the individual testing situation.*

We are interested that children know the alphabet, for instance, but we also are interested in how and when they use it in the classroom and the home. Observational schemes are required to study this facet of process.

A third sense of cognitive process is *linked to global notions of social-cognitive competency* and cannot be reduced to skills generalized or transferred from the ones measured in the individual test situation. Changes may take place in learning to use adults as resources, learning such attentional techniques as dual-focus monitoring, learning to select attainable and satisfying activities and goals in the classroom and elsewhere, learning appropriate *tempo* of play (at what pace, how long, and with what duration of sustained involvement in particular aspects of the activity), learning to *seek* good problems. This third sense of cognitive process has been largely overlooked in past Head Start evaluations. Empirically and naturalistically defined, it is a high-risk, high-gain area for measurement in the next Head Start evaluation, with much to be measured but few current instruments to do the job. New measures might be based on instruments for ethological observation in the neighborhood and home (Barker, 1968; Schoggen and Schoggen, 1971; Watts et al., 1972; White and Watts, 1973; Wright, 1967).

4. *Social competency and awareness.* This category has a sizable overlap with the process category. But here we are concerned with the child's instrumental knowledge of his or her immediate environment--knowledge about people, rules, etiquette, institutions (What does a policeman do?). It would be valuable to have a paper from Irving Goffman on "children's relations in public," trying to map some of the strictly kinesic dimensions of awareness about the social world of the neighborhood and the school, about older children and what to do and not to do around them, and about appropriate and inappropriate strategies for getting what you want as a child. This kind of knowledge could be tapped by various kinds of measures, but it does not lend itself to assessment in one-to-one testing situations. It is possible to ask the child a number of simple, direct, true and false questions about what to do and what not to do in his or her neighborhood, but attempts at this kind of individually administered item often are without much

validity. They are detached from particular local circumstances and surroundings or only fragmentary in what they measure. In this domain we should probably place more trust in observational measures--samplings of child behavior in the classroom, the home, or the neighborhood.

Sheldon White (personal communication, September, 1973) has called this type of cognitive competence "ability to use community-accepted metaphor." This is apt. Head Start children probably grow in their awareness of cultural norms--usually the norms of two or more cultures. If the program helps children learn to mediate the discontinuity between the culture of their homes and that of school and workplace, then we should be trying to assess this increased sophistication directly. Some of the measures Cole is developing (personal communication, July, 1973) in connection with his school-based research in New York City deserve to be considered for adaptation to Head Start.

5. *General knowledge.* The general knowledge category refers to public knowledge any child, regardless of "ecological niche," might be expected to know about the world. The category includes general information about history, government, current events, and other areas regarded as important but without any immediate practical significance to the child. In many tests, items tapping general knowledge have been included both as indicators of school readiness (a dubious purpose) and as general intelligence items. The predictive validity of such questions in one-shot testing usually proves as high when correlated with later school performance scores as any other intelligence or achievement item. But they have been controversial when used in Head Start evaluation because it is not clear that general knowledge is useful or necessary for a preschooler, or that because a child does not know some specific fact he will suffer later in school or in his day-to-day life. It is also hard in general to make the case that there is any single corpus of knowledge and facts that all children should know.

Certain Head Start programs may feel that mastery of a particular realm of general knowledge is an important program goal. If knowledge-based testing is performed, it should be seen as largely a test of language comprehension or vocabulary. Language comprehension or

vocabulary items need to follow the Berko-Brown (1960) format, whereby a picture is shown and then two alternative sentences are presented, one of which is a correct description of the picture and the other not. Such questions are not intended to test language production by the child or syntactic understanding, but purely semantic understanding.

PRIORITIES FOR MEASUREMENT

Among the five realms of cognitive effects, I feel priorities for evaluators lie in the areas of *school readiness*, *cognitive process*, and *social awareness and competency*. Before we understand the pedagogical implications of cognitive developmental theory, it seems unwise to orient an evaluation to theory-based changes. It also seems wrong to stress general knowledge, since this realm is so hard to stake out unambiguously and harder still to make a virtue of mastering.

The three priority areas have a common advantage: They are amenable to assessment with a theoretical, empirically based, and criterion-referenced measures. For the most part issues of latent-trait shifts and predictive validity are finessed; the emphasis of the evaluation is with short-term observable changes in the child's performance and general behavior, as tapped by individual tests, observational schemes, and rating scales and interviews. These three sets of assessment criteria weight the evaluation primarily toward school readiness and the child's growth as an *effector* of his environment, as a manipulator of the immediate physical and social surroundings. This is a modest but practical orientation.

IV. THE NEW COGNITIVE EFFECTS BATTERY

It is tempting to recommend that Head Start discard all instruments used in past evaluations and develop an entirely new cognitive effects battery. Before taking such a position, however, we should seriously consider one variant of what I have earlier called Position 4. This is the hypothesis that there is no point in spending a lot of money to develop new measures because we have already shown what we will show again: A good program can get gains on *any* measures and a bad one probably cannot. There are already at least 20 good lab school studies, and now the Planned Variation study, indicating that some programs do achieve short-term effects on a variety of cognitive measures. If short-term effects need to be demonstrated again, perhaps we should not spend money on new instruments, instead simply pointing to the growing list of studies showing short-term gains and choosing a limited number of tried and true measures to show that these gains can be replicated by a good Head Start program.

If we shuffle our measures and succeed only in demonstrating effects on the same order of magnitude and reflecting the same order of program differences shown in the PVHS study, the average policymaker may not see that we have told him anything new. In fact, such a shift could make matters worse. In the search for new measures evaluators might well trade away reliability for a presumed increase in face-validity, risking distrust of results.

I believe we should invest in instrument development only if it will result in a limited but excellent battery of cognitive effects measures that will give us both more trustworthy evidence of results in the field and substantial increases in validity, especially in talking about the differential effects of high cost programs and low cost programs. In other words, if we can have a tight design, a limited number of program prototypes to explore, and a limited number of cost-related questions to ask, then I am satisfied it is worth investing a considerable portion of the evaluation budget to develop new instruments. If, however, there is no indication that

[illegible] of these conditions will be fulfilled, then perhaps it makes most sense to administer the Binet, the PSI, the WRAT, the MRT, or some other familiar measure, showing effects for certain groups of children or showing higher effects for some programs, and letting it go at that.

Let us proceed on the assumption that we are interested in reconceptualizing the measurement of cognitive effects, accepting a validity-based strategy and devising a good new battery with a limited but promising repertoire of instruments. In each of the three areas emphasized in Section III--school readiness, cognitive process, and social competency--we need to begin by listing specific behavioral objectives relevant to all programs. Ideally, although it is outside the scope of this study, we also need to list behavioral objectives that differ from program to program according to the goals of a particular staff or sponsor.

It is difficult to select a set of measures that will show all programs to equal advantage. Some programs have goals that translate themselves more readily into measurable behavioral objectives. When the Bereiter program, for instance, says that it wants children to be able to perform certain rote manipulations of letters and numbers, it is not difficult to figure out how to assess this goal. When the [illegible] program says that it wants to teach children to "learn how to learn," however, things become much more difficult. We still do not know how to tell a child who has learned how to learn from a child who has not.

Despite all these difficulties, it is probably necessary to be somewhat impatient toward programs that cannot operationalize what they are trying to do, or give examples of specific behavioral objectives demonstrating goal attainment. If program designers cannot be specific about at least some behavioral objectives, then we can justifiably wonder whether they know what they are doing.

Whatever the instruments finally included in the final battery, they will probably assess the goals of one program somewhat more accurately than those of another, but it is probably [illegible] [illegible]

Instead we should simply make sure that at least one instrument is included in the battery that each program can accept as a measure of its cognitive effects.

The following list of behavioral objectives in each of the three high-priority measurement realms is far from complete:

[illegible]: Does the child know his numbers and letters?

Can he keep three bits of information in short-term memory and work on one of them?

Can he detect the difference between the Gibson stimuli and letters?

Can he sustain attention on some school-related task for five minutes?

Can he comprehend sentences presented in the Berko-Brown format?

Can he exhibit advanced pre-operational thinking on certain clinical, Piagetian measures of quantitative reasoning?

Does he exhibit a sufficient level of socio-centric awareness when playing with his peers? Does he fight with them?

Does he use relational terms in carrying out a series of commands?

Can he express himself clearly enough in standard English to make various requests of his teacher and other adults?

[illegible]: Does the child progress toward greater reflectivity in problem solution?

Is his tempo of play well modulated? How long is each sustained involvement? How does this differ for various activities?

Can the child use adults as resources? Does he have a number of different strategies for doing so in the classroom, at home, or in the neighborhood?

Can the child monitor one activity in the classroom while doing another?

Can the child select something he wants to do and see it through to completion, in the classroom or from day to day in the neighborhood (sustained, goal-directed activity)?

Can the child invent alternate strategies for solving a problem in the test situation or attaining some goal in the classroom or the neighborhood?

Can the child monitor, relate to, and manipulate the desires of his peers?

Can the child apply to a new problem a strategy he has been taught in several previous structured, problem-solving situations?

Can the child *seek* good questions and does he routinely do so?

Is the child more verbal, in the simple sense of gross production of coherent sentences?

Is spontaneous verbal elaboration of answers more pronounced in the individual testing situation? In the classroom?

Is the child more observant of his older siblings as role models around the house and neighborhood?

Social Competency: Does the child understand the functions of various community institutions and officials (in a culturally valid sense, not a textbook sense)? Does he know what his older brothers and sisters think of the police? The mayor? School? And why they feel this way?

Does the child know his neighborhood--its geographical layout, various points of interest (e.g., the library, the community center) and various people in these places who can be of use to him?

Does the child know his *rights* in the community? Where to go if something happens to him, whose business it is to protect him if something goes wrong (e.g., the doctor in the local hospital, the family counselor at the welfare agency, etc.)?

Does the child know and understand the attitudes of his parents toward him and his siblings? And how these might differ from the attitudes of other parents?

Does the child know where his father and mother work? Has he ever visited them there and does he know what they do?

Has the child ever visited the school he will attend in the coming year and met teachers there?

Does the child know certain things around the neighborhood it would be unwise for him to do, either because they might result in physical harm or because they would violate neighborhood or cultural norms?

Can the child switch easily from dialect of the neighborhood to standard English and back? Does he know the neighborhood circumstances under which each is appropriate?

Does the child talk to his parents, especially about matters not related to his own conduct?

This list is just a beginning. It needs to be greatly amplified before we can winnow the list to "best bets" for actual Head Start measurement. The team designing the measurement battery needs first to come up with a complete list of candidate behavioral objectives and then to invite a group of Head Start teachers and directors to critique them, rejecting unlikely ones and adding some of their own. Certainly a list could be generated which is far more imaginative and more face-valid than the ones assumed in past evaluations.

Now let us turn to the various types of measures which might be employed.

INDIVIDUALLY ADMINISTERED TESTS

These have been the work-horse measures in all previous Head Start evaluations. They have fairly high reliability and tend to be moderate in cost if they are not too long and tester training is not too elaborate. Average cost varies between the high cost per testing of the Binet, for instance, and the much lower cost of the PSI or N.Y.U. booklets. Individually administered tests are good for measuring some aspects of school readiness. They also are useful in measuring theory-based developmental gains and general knowledge, but I have tried to argue that these areas should not receive major attention. Finally, they may be useful in measuring selective aspects of cognitive process, especially where response style is able to tell us something

about process; and in the measurement of social competency, where Berko-Brown types of questions and child interviews are useful.

In general, we have erred in the past by looking at too narrow a slice of the Head Start child's experiential transformation over the year. To the extent the choice of individually administered tests is responsible for this myopia, they should not be emphasized. Of course, if there is little time for developing new measures, a new evaluation may still have to rely heavily on these instruments.

Individually administered tests of cognitive performance are currently available in a wide variety, although in many cases their ready application to Head Start can be questioned. Readers are referred to the ETS summary of available tests (1968), and the Huron Institute report on the PVHS battery (Walker, Bane, and Bryk, 1973). In general there is a great need for better empirically based, external-validity based, and criterion-referenced measures. Some new tests in this category deserve attention. In particular the ETS CIRCUS (1974) developed by Bogatz and other Sesame Street test developers, might be adopted. The ETS CIRCUS attempts to extend principles of Sesame Street test development into more general preschool and early elementary school testing. It is a promising criterion-referenced battery and might lend itself wholly or in part to Head Start measurement. CIRCUS has the additional advantage (or hazard?) of including various tests administered to several children at once by a single tester. If reliability can be maintained, the cost advantages of such a scheme are obvious.

Piagetian clinical measures also should be explored, and scoring for response style should be mandatory on all individually administered instruments.

New tests also might be based on what we have learned from past Head Start testing. One approach would be to look for items or subscales from instruments used in earlier Head Start evaluations that explained a high proportion of generalized gains or between-program variance, then creating new scales from these for the new evaluation. This process of scavenging would require independent assessment of the validity and reliability of the new composite measures, but the new

instruments might be more valuable than the older ones. Alternatively, we might find out from past evaluations which subscales and items were most reliably administered, using only these in a reduced scale of one or two factors. The assumption in this case would be that reliability and cost were primary considerations, within some validity constraint. Test dimensions might be reduced with a significant increase in reliability per dollar.

CLASSROOM OBSERVATION INSTRUMENTS

Classroom observation schemes often have been of low validity in the Head Start classroom because they were designed for investigation of another kind of classroom setting or because they have tried to monitor too many aspects of classroom process at once. In general, the only kind of observation instruments that should interest us in a new evaluation are those enabling exploration of particular hypotheses regarding face-valid behavior changes over the Head Start year. These may by hypotheses closely linked to performance on individually administered instruments or they may be hypotheses not amenable to exploration in any other way, such as those concerning dual-focus monitoring or sociocentric play. Best bets should be made in advance about most likely face-valid changes and should determine what is observed. It is too late at the time of data analysis to dredge for interesting results. Classroom observation measures tend to be more costly than other measures, both to administer and to analyze, and they are apt to be less reliable, especially if they require collection of large amounts of information in a short observation interval. This is another reason to be clear in advance about hypotheses to be explored.

It probably will not be necessary to devise entirely new observation instruments. The ETS PROSE (Medley et al., 1971) has many interesting aspects, as do the Bankstreet measures, some of which assess motivation and curiosity (Stern and Gordon, 1967; Cohen and Stern, 1968). But it will require real skill to select components of current instruments and adapt them to the specific dimensions of classroom process that interest us.

HOME AND NEIGHBORHOOD OBSERVATION INSTRUMENTS

One area of measurement never attempted in a Head Start evaluation is naturalistic observation of the child in his neighborhood or home. Ethological or ecological assessment of Head Start's global effects would be valuable, especially if measures could be devised to explore questions about cognitive process and increased social competency and awareness. It is of obvious importance to policymakers that we measure the child's actual conduct in the world outside the Head Start center. Measures of behavior outside class have a built-in external validity other instruments cannot claim. The OCD is now interested in global evaluation and may in the future want to emphasize Head Start's effects on the family in the fuller context of neighborhood and home (Bonfenbrenner, in press).

In general, home observation strikes me as a high-risk high-gain venture. We know little about how to do it for Head Start children or about which hypotheses to explore. (Does Head Start give a child more poise in dealing with his mother?) But if effects could be demonstrated it would be powerful evidence of Head Start's value. Among the few good measures in this domain at present are those of White and Watts (1973), looking at parent child interaction in the home, those of Watts (Watts et al., 1972), for infants and toddlers in daycare centers, and those of the Schoggens (1971). None of the measures are fully appropriate for use in Head Start, but they might be adapted.

Neighborhood observation outside the home and the Head Start center is also tempting, but it raises even greater difficulties. Observers might follow children from place to place as they played, in the fashion of some of Piaget's earliest work or the work of Barker (1968) and other post-Lewinians. There are problems with such an approach, however, unless we could agree on indices of increased social awareness or competency and could do the assessment in structured settings. Two possibilities for structure come to mind. We might ask that children go with their mothers or other family members to certain neighborhood stores and services and then observe their interactions there. Alternatively, we might assign the child various tasks to

perform in the neighborhood, reminiscent of a treasure hunt. If we want to know whether children can find the fire department or know how to talk to a policeman, we may be well advised simply to design a task that has them do this. Children might be given a list of things to do and then be observed while they did them, or assessed afterward according to whether they were able to do them. Control children might be given the same tasks.

PARENT, TEACHER, AND SIBLING INTERVIEWS AND RATINGS

These approaches generally fail on grounds not of reliability but of face validity. It is not terribly convincing to be told by parents that their child is now more competent than before, or to be told by a Head Start teacher that Head Start children are performing better than controls. Moreover, blind procedures that might enhance validity are clumsy and expensive. But here again there is a realm of imaginative, face-valid measures that might be considered. We might, for instance, collect data from kindergarten or first grade teachers on the placement of the last year's Head Start children. We might also interview parents about how children have changed in their preferred activities. Both of these approaches could be valuable, at least in the preliminary stages of developing observational instruments. They would help us gather information from teachers, parents, and other neighborhood people on best bets for specific areas of behavior to be assessed.

INSTRUMENTS FOR COLLECTING INCIDENTAL FACTS ABOUT REDUCTIONS IN SOCIAL COSTS

One category of instrument overlooked in the past that should be given attention under the rubric of cognitive effects measures and elsewhere in the evaluation is the catchall category of facts about [illegible] by Head Start. David Weikart (personal communication, July, 1973), for instance, has had success in promoting his program simply on the grounds that it results in fewer children being assigned to MR classes in school, with resulting cost reductions to the taxpayer. Weikart can make the case that early education, at

least in his program, is cost-effective. Such effects probably would not show up as dramatically among children in field sites as among children in the Weikart lab school program, but there is no doubt that this "cognitive effect" criterion is important.

Such a measure shifts the burden of proof for cost-effectiveness from Head Start to later programs, which will have to cope with untreated problems in the event a child does not attend preschool. Head Start probably has a number of such benefits, resulting from screening procedures of various kinds and from the child's being more aware and better socialized than before. We might explore incidence of undiagnosed problems of sight and hearing in the year after Head Start, incidence of children's involvement with the juvenile courts, and incidence of problem behavior in the kindergarten or first grade.

BALANCE AMONG TYPES OF INSTRUMENTS

Among the various kinds of measures, it remains for us to decide an appropriate mix given what we know about currently available tests in each category, how much money we have to spend to develop and administer tests, and how much time we have to design new instruments. Money, timing, type of instrument, predicted levels of face validity and reliability of administration, all of these must be weighed simultaneously before we can tell OCD "what it should want." These considerations cannot be sorted out fully here, but some generalizations can be made.

First, we need to keep the cognitive-effects evaluation *simple*. The fewer instruments the better, if the purposes of the evaluation are well-served by the ones chosen. I believe in George Miller's magic seven plus or minus two, preferring to err to the minus side where bureaucrats and legislators are involved. Most consumers of Head Start evaluations simply are not able to digest more than five or so measures in any given domain and make sense out of results. If we could present our findings on five or six dimensions of cognitive effects, perhaps showing differential program effects as Lesser, Fifer, and Clark (1965) did in their profile analysis of patterns of mental

ability, this would be interesting and comprehensible to a wide audience. Anything more complicated, involving lots of second and third order interactions and differential effects on multiple instruments, serves only to confuse everyone.

Second, it is important to decide how much additional instrument development is necessary in the area of cognitive assessment. As I see it, we can assume one of two stances on the matter, one quick and expedient, the other slower and with potentially higher yield. These are extensions of the reliability-based and validity-based strategies mapped earlier. The first choice is to give much more centrality to cognitive measures and to opt for conservative, highly reliable, and politically compelling data. Planners forgo any attempt at new cognitive instrument development, accept some version of Position 4--that cognitive effects should be a moderator variable--and adopt a limited assortment of the best currently existent measures. Some will be the same as the ones in the PVHS battery: perhaps an IQ measure, in acknowledgment of its political currency, along with certain criterion-referenced measures such as the WRAT, the PSI, or some of the NYU booklets. Perhaps one or two other measures in the works can also be selected (e.g., ETS CIRCUS tests).

If we adopt this approach, heavy emphasis must be placed on individually administered tests, to get highest possible reliability per dollar. No current instruments in other domains can match the individual tests in this regard. Classroom observation would have to be limited drastically, to explore only a few specific hypotheses about transfer effects of individually tested competencies to observed activities in the classroom. Cognitive process assessment would be limited to what could be learned from coding schemes for cognitive style on the individually administered tests. There would be no assessment of social competency or social awareness in the cognitive domain except what could be gleaned in the individual testing situation. Parent and teacher interviews would be downplayed. Certain face-valid facts of interest to cost analysts, of the sort mentioned in the previous section, would be collected.

In general, cognitive effects assessment of this sort would serve the purpose of demonstrating that something reliable was happening in Head Start--something we knew about sponsored programs before but that needed to be shown more carefully. It would not involve any re-conceptualization of cognitive effects measurement. There would be only three differences from past evaluations: the battery of cognitive measures would be more limited; it would reflect an emphasis on face-valid, criteria-referenced, short-term program effects; and it would be much more carefully administered than before.

The second option is more to my liking. If planners could get the concession of more time from OCD it would make sense to spend a year developing a new battery of tests, limited in number but designed more carefully with Head Start in mind. A battery developed with this much lead time could include individually administered measures but also other kinds of measures, striving for equal levels of reliability for all. Agreeing upon a set of behavioral objectives in the domains of school readiness, cognitive process, and social awareness and competency will take time. It will then take more time to devise instruments that measure these objectives to everyone's satisfaction. Teachers or other Head Start field personnel have to give their opinions about which objectives are most important. Then instruments must be developed or adapted for the subset of behavioral objectives chosen, and these measures pre-tested. This is not a process that can be accomplished in less than one year. A new battery evolved in this fashion might include two individually administered tests, two observation schemes (one in the classroom and one in the home), one wild card instrument assessing child competence in various tasks either around the neighborhood or in the kindergarten and first grade classroom with older children, or both, and perhaps an inventory of social cost-benefit indices.

Regardless of which option is chosen, it is abundantly clear that we are not ready to launch another three-to-five year longitudinal study immediately. Under option one we would be doing little more than replicating certain aspects of PVHS, with better data but with

no possibility of changing to better measures. Under option two, we need a period of test development and then a field trial before being ready for another major evaluation.

MEASUREMENT STRATEGY AND EVALUATION DESIGN

Design of the evaluation is closely related to choice of instruments. If the OCD selects a design comparing a more costly but presumably more effective sponsored program with traditional programs and non-Head Start controls, for instance, then the final layout will have three levels--a sponsored group, a traditional group, and a control group. Assuming for illustration a sample of 600 children, each treatment group would have 200 children. These might be all the children from a limited number of centers, or if OCD was willing to pay more, could be a group chosen randomly from among the children in all centers of the appropriate treatment type. Notice that even with this simple three-level design and with no mention of other independent variables, the study would be down to two hundred children per level.

If the OCD is interested in a good study, without hopelessly confounded variables preventing valid inference, then it needs to realize that as more and more independent variables, covariates, and other controls are introduced, cell sizes can diminish rapidly to nothing, or next to nothing, leaving us with the problem that characterized the PVHS design. This should not be permitted to happen. I will try to be more specific, to show concretely how hard it is to avoid the temptation of trying to investigate too much at once.

Judging from all we have learned in past Head Start evaluations, the following variables are important to stratify on or control:

geographic region
urban/rural
ethnicity
SES
IQ
age (four or five, or broken out at monthly intervals)
previous preschool

No evaluation can gloss over these sources of differential effects. Not only must good data be collected but comparison groups must be well planned. Returning to the hypothetical estimate of 200 children per experimental level, let us make the further, not unreasonable, assumption that for reasons idiosyncratic to the study design, cell size cannot drop below 15 and still allow reliable estimates. Now evaluators probably could only consider *three* additional independent variables to cross the sample on, not the four or five we might like. Reviewing the candidate variables on our list this could lead to difficulties.

SES and IQ would not be the problem. The range of SES in the Head Start population is greatly circumscribed and does not seem to explain much of the variance in outcomes in past evaluations. SES can be entered in most analyses as a moderator variable or covariate. Evaluators also can covary on IQ, unless the design is intended to look at special benefits for low-IQ children. There are large enough numbers in each treatment group to make sizable differences in group IQ means unlikely. *Ethnicity* is more of a problem. No evaluation can disregard it, but to make it a prominent independent variable does not seem advisable. In any event, most Head Start children are black. Probably the best solution is to make sure of a roughly comparable racial mix for each cell in the design, simply choosing children or centers for inclusion in the study with the understanding that ethnicity will be controlled by initial stratification.

Geographic region and *urban/rural* also have to be considered. They cannot be finessed, since programs in one part of the nation often are quite different from those in another, and since a program in the country usually differs from one in the city. We need representation in each of these areas for policy-relevant inferences about Head Start's nationwide effects. There is, however, the possibility of post hoc pooling of data in the event that no statistically significant differences are found between groups.

Finally the real bugaboos--*age* and *previous preschool*. There is no question that a test for a four year old is not the same as a test for a five year old, and maturation-related changes in children's thinking make a tremendous difference in the utility of given measures.

Thus, for instance, the PSI is generally acknowledged as a test for four year olds, with definite ceiling effects for fives. To make matters more complicated, in the PVHS study it is clear that previous preschool made a difference in five year olds' performance on most of the individually administered tests: Gains are not as great if a five year old is in his second Head Start year. There are a number of possible explanations for this phenomenon, ranging from the least important (increased familiarity with the pre-tester in a second year of the program) to the most important (reduced marginal utility of Head Start influence in a second year of the program) suggesting that we should have only first year children.

In the past, some Head Start evaluators have established age by year as an independent variable, others have chosen age by month, others have introduced age by month as a covariate or dummy variable in regression equations. But no result is fully satisfactory if the same test is measuring different things for children of different ages. Perhaps children should be chosen for the evaluation from an age range no wider than a year, or centers should be selected with children in a narrow age range. There is much to say for concentrating on four year olds, skirting the problem of previous preschool as an additional variable, but in some regions, especially the South, many Head Start programs lead directly into first grade. There are few (pre-kindergarten) children because there are few kindergartens. This issue deserves a red flag. It is directly related to the choice of relevant measures and their validity, and it is precisely the kind of issue that in the past has been overlooked, resulting in a nightmare for data analysts. Let us not be afraid to delimit the study somewhat if it will mean we can place more trust in what we find.

It might also be interesting to reconsider the number of testings during the Head Start year, moving from the pre-post design of past evaluations to a time-series design with three testings, or a design with true randomization of subjects and a single criterion testing in the spring.

Evaluation planners should at least consider reducing the number of instruments in the battery, or reducing sample size, and using the

[illegible] used to administer some instruments three times during the [illegible]. This would yield information about trends and about the times [illegible] the year when most gain is taking place. If it were discovered, for instance, that children were experiencing a large portion of Head Start improvement during the interval from October to Christmas and improving only marginally thereafter, this would be a [illegible] of obvious importance. There are also problems with three testings, however. In the absence of alternative forms for most preschool instruments, practice effects are likely. Also, the administration of the tests presents a logistical problem of large proportions and a potential irritant to Head Start center staffs. One testing would have to begin almost as soon as the previous one ended.

[illegible] opposite notion, administering criterion-referenced tests [illegible], would be interesting if there could be true randomization [illegible] children or an adequate approximation of it. Such a design might [illegible] approximated with a full battery administered at post-test [illegible] only a minimal number of [illegible] instruments given at pre-test. [illegible] variant of this approach would be analogous to the National Assessment strategy, by which a large bank of items are given and [illegible] children [illegible] many more [illegible] be given. For such a design [illegible] valid, however, it would be necessary to assume [illegible] of the design were similar on all important background variables and the interactions of these variables. [illegible] sample would have to be chosen with this assumption in mind.

[illegible] National Assessment strategy [illegible] to administer a large bank of items at pre-test and post-test, with randomization of children to treatments and randomization of items to children. [illegible] would be given pre and post [illegible] then be analyzed with items rather than children [illegible] unit of analysis. The potential advantage of such an approach [illegible] would permit inferences about effects [illegible] wider range [illegible] tested [illegible] considerations [illegible]

In addition to the assumption that children are similar on salient third variables from cell to cell, it would be necessary to assume that magnitude of gains can be compared from item to item. Also, items would have to be analyzed individually and not as part of scales, meaning that no independent validity or reliability estimates would be possible based on item-scale characteristics.

A final area that has never received adequate consideration in Head Start evaluations is the issue of decision rules for program success. Such rules should be made explicit *before* the evaluation begins, so that it is clear that a gain of X amount on Y scale, or [illegible] on items B, C, and D, constitutes a sufficient demonstration of Head Start success. This consideration takes us back to the beginning of the report, where it was suggested that until there is some understanding of what it takes to convince appropriate audiences, an evaluation should not be initiated. Once such an understanding is attained, decision rules based on it should be made explicit.

Such rules have not been formally stated in the past because evaluators have wanted to anticipate various different configurations of findings and did not want to commit themselves or the OCD in advance to [illegible]. The desire to keep options open is understandable, [illegible] thinking of it as traditional optimism of [illegible] the early statement of a decision rule can only [illegible] while [illegible] unanticipated and [illegible] emerges in the data. But it is nonetheless important [illegible] such rules early rather than rationalizing them later. Otherwise [illegible] interpretations of [illegible]

V. CONCLUSIONS

The Westinghouse-Ohio study tried to demonstrate systematic, sizable cognitive effects for a randomly selected group of Head Start centers, comparing children in these centers with non-Head Start controls. It found only slight effects when all programs in the sample were aggregated and mean gains were assessed. It would be a mistake for the next Head Start evaluation to recreate the Westinghouse-Ohio study, even with a better design and better measures of cognitive effects. Whatever the measures selected, effects probably will not be large enough in such a design to command respect from policymakers, even if sample size is large enough to give them a fair chance of being statistically significant. Substantively, much is obscured by analyzing program gain data only at the highest level of aggregation. It is not surprising that overall effects are slight; programs differ widely from one another and we know some are good and others bad. It is better to ask which programs are doing well and why.

It is also important to shift the terms of the debate about fadeout gains. This is not a fair criterion for Head Start evaluation. Longitudinal effects might be possible to demonstrate, at least into first and second grade--they have been demonstrated in smaller studies--but it would be unwise to spend the money necessary for a careful longitudinal evaluation design exploring this aspect of gain, especially when the magnitude of effects probably is not great. This is a criterion of program funding not imposed on other federal programs or other levels of schooling and should simply be resisted.

The Planned Variation Head Start Study was an elaborate natural experiment that told us less than we had hoped. It was initiated before we were sure (a) we could implement various sponsored programs as though they were part of the same treatment, from the same template, (b) we could control all necessary variables in order to make valid inferences, (c) we had a battery of measures that would tell us about the differential effects of programs, and (d) these tests would be well-enough administered to be trustworthy. If we have learned

anything from PVHS it is that bigger is not necessarily better, and that natural experiments in education, while they represent a significant improvement over *post hoc* survey research, have problems of their own. The next evaluation, if it chooses to explore the interaction of program type and child group, should perhaps develop hypotheses from the PVHS data, but it should be much smaller and more carefully designed and executed. Randomization of children, clear definition of treatments, and adequate controls are needed.

One strategy for evaluation planners in coming months would be *reliability based*. It would begin with a fixed budget and brief time frame as constraints; instruments would be selected from those currently available or nearly developed; sample size would be determined by dividing per child cost of reliable test administration into the total sum available for test administration. This approach would arrive at better data on a limited number of familiar measures. It would also be compatible with a decision to make cognitive effects measurement less important in this evaluation than previous ones, using cognitive instruments as moderator variables in a study that emphasized effects outside the cognitive domain. Perhaps such a study would focus on the areas of health and nutrition, social development, and effects on the family. If OCD must initiate a new evaluation soon, a reliability based strategy is sensible.

If there is sufficient time to devise a new Head Start cognitive effects battery, then a *validity-based* strategy may be possible in primary the evaluation, with substantial efforts made to develop new and more valid instruments. Test development could proceed in the areas of school readiness, cognitive process, and social awareness and competence. Some currently available instruments can be adopted, but it is certain that a battery departing significantly from the present one will take no less than a year to develop. The extra time would be worth taking if cognitive effects are again to play a central role in the evaluation.

Regardless of which strategy planners adopt, one *sine qua non* of the coming evaluation is that the field operation be well administered.

In past Head Start evaluations, the most basic aspects of test administration and data collection have gone wrong, resulting in data of dubious value. Under no circumstances should this error be repeated. It makes more sense to administer a few measures well than a lot of measures poorly. This is true even if it means to some extent sacrificing external validity resulting from a large, representative national sample and a large number of instruments.

BIBLIOGRAPHY

Ali, F. and J. Costello, "Modification of the Peabody Picture Vocabulary Test," [illegible], 86-91, 1971.

Ball, S. and G. A. Bogatz, *The First Year of Sesame Street: An Evaluation*, Educational Testing Service, Princeton, N.J., 1970.

Barker, R. G., [illegible], Appleton-Century-Crofts, New York, 1963.

Barker, R. G., *Ecological Psychology*, Stanford University Press, 1968.

Barker, R. G. and P. Schoggen, [illegible], Jossey-Bass, San Francisco, 1973.

Bereiter, C., "Educational Implications of Kohlberg's Cognitive-Developmental View," [illegible], 2 (1), 25-31, 1970.

Bereiter, C., "An Academic Preschool for Disadvantaged Children. [illegible] from Evaluation Studies," in J. Stanley (ed.), [illegible] *the Disadvantaged*, Johns Hopkins Press, Baltimore, 1972.

[illegible] and R. Brown, "Psycholinguistic Research Methods," in P. H. [illegible] (ed.), [illegible], Wiley, New York, 1960, pp. 517-560.

[illegible] Effects of Preschool Programs for Disad[illegible] [illegible] doctoral dissertation, Harvard Graduate School of Education, [illegible].

[illegible], "[illegible] Implementation of Planned Variation in Head Start: [illegible] and Summary of First Year Report," Office of Child Development, Dept. of HEW, Washington, D.C., 1971.

[illegible], "Planned Variation in Head Start and Followthrough," in [illegible] (ed.), [illegible], Johns Hopkins Press, Baltimore, Md., 1973.

[illegible], "[illegible] Early Intervention Effective?" Office of Child Development, HEW, Washington, D.C., in press.

[illegible], "The Measurement of Change," Harvard Graduate School of Education, [illegible], 1972.

[illegible], 1972 (2 vols.).

Buros, O. K. (ed.), *Seventh Mental Measurements Yearbook*, Gryphon, Highland Park, N.J., 1972 (2 vols.).

Butler, J. A., "Item Components of Preschool IQ Gains," Harvard Graduate School of Education, 1973 (unpublished).

Caldwell, B., "The Preschool Inventory: Technical Report," Educational Testing Service, Princeton, N.J., 1967.

Cicirelli, V. G. *et al.*, "The Impact of Head Start: An Evaluation of the Effects of Head Start on Children's Cognitive and Affective Development," Office of Economic Opportunity, Washington, D. C., 1969.

Cohen, D. K., "Social Experiments With Schools: What has been learned?" Paper prepared for a conference on social experimentation, Brookings Institution, August 1973.

Cohen, D. K. and V. Stern, *Observing and Recording the Behavior of Young Children*, Teachers College Press, New York, 1968.

Cole, M. and J. S. Bruner, "Some Preliminaries to Some Theories of Cultural Difference," *NSSE Yearbook*, University of Chicago, 1972.

[illegible]ller, A. and J. Victor, "Early Childhood Inventories Project," Institute for Developmental Studies, N.Y.U. School of Education, New York, 1971.

Cooperative Tests and Services, "Preschool Inventory Revised Edition--1970 Handbook," Educational Testing Service, Princeton, N.J., 1970.

[illegible], "[illegible] and Social [illegible]," [illegible] March 1966, 105-124.

Educational Testing Service, "Disadvantaged Children and Their First School Experiences. Theoretical Considerations and Measurement Strategies," Educational Testing Service, Princeton, N.J., 1968.

Educational Testing Service, "Enumeration Test," in V. C. Shipman (ed.), Disadvantaged Children and Their First School Experiences: Technical [illegible] Series, Educational Testing Service, Princeton, N.J., 1972

Educational Testing Service, [illegible], Educational Testing Service, Princeton, N.J., 19[illegible].

Featherstone, H., "Cognitive Effects of Preschool Programs on Different Types of Children," Huron Institute, Cambridge, Mass., 1973 (prepared for the U.S. Office of Child Development).

[illegible], M. [illegible] and [illegible], McGraw-Hill, New York, 19[illegible].

Gibson, E., [illegible], [illegible] Pick, and H. [illegible], "A Developmental Study of the Discrimination of Letter-like Forms," [illegible], 1962, [illegible], 897-906.

Hainsworth, P. K., and M. C. Siqueland, "Early Identification of Children with Learning Disabilities: the Meeting Street School Screening Test, Crippled Children and Adults of Rhode Island, Inc., Providence, R.I., 1969.

Harris, C. W. (ed.), [illegible], University of Wisconsin Press, Madison, 1963.

Hertzig, M. [illegible], [illegible] Birch, A. Thomas, and O. A. Mendez, "Class and Ethnic Differences in the Responsiveness of Preschool Children to Cognitive Demands," [illegible] (1), 1968.

Hess, R. D., "Parental Behavior and Children's School Achievement: Implications for Head Start," in E. Grotberg, (ed.), [illegible], Educational Testing Service, Princeton, N.J., 1969.

Hildreth, G. H., [illegible] Griffiths, and M. E. McGauvran, "Metropolitan Readiness Tests, Form A," Harcourt, Brace and World, New York, 1965.

[illegible], Holt, Rinehart and Winston, New York, [illegible].

[illegible], 1972.

Kagan, J. and H. A. [illegible], 1962.

Karnes, M. B., "Evaluation and Implications of Research with Young Handicapped and Low-Income Children" in [illegible] Stanley (ed.), [illegible], Johns Hopkins Press, Baltimore, Md., 1973.

[illegible], "Reply to Bereiter's Statement on Kohlberg's Cognitive-Developmental View," [illegible].

[illegible], University of Pennsylvania Press, Philadelphia, 1972.

[illegible], [illegible] Fifer, and D. H. Clark, "Mental Abilities of Children from Different Social Class and Cultural Groups," [illegible] Research in Child Development, 1965, [illegible], Serial No. 102.

Levine, S., F. F. Elzey, and M. Lewis, "California Preschool Competency Scale Manual," Consulting Psychologists Press, Palo Alto, 1969.

[illegible]an, M., "The Estimation of a Moral Judgment Level Using Items [illegible] Alternatives Form a Graded Scale," doctoral dissertation, [illegible]iversity of Chicago, December 1970.

[illegible], R. [illegible] and P. V. Smith, "Choosing a Future: Strategies for De-[illegible]ing and Evaluating New Programs," *Harvard Educational Review*, [illegible], [illegible] (1).

[illegible], [illegible] and P. V. Smith, "Accumulating Evidence: Procedures for [illegible] Contradictions Among Different Research Studies," [illegible], 197[illegible], 41 (4), 429-471.

[illegible], "A Comparative Analysis of Preschool Curriculum Models," [illegible] Anderson and H. G. Shane (eds.), *As the Twig Is Bent: [illegible] Childhood Education*, Houghton Mifflin, Boston, 1971.

[illegible], [illegible] M., T. J. Quirk, C. G. Schluck, and N. P. Ames, "Personal [illegible] of School Experience: A Manual from PROSE Recorders," [illegible] Memorandum 71-8, Educational Testing Service, Princeton, [illegible], June 1971.

[illegible], "Predictive Validity of the Metropolitan Readiness [illegible] and the Murphy-Durrell Readiness Analysis for White and [illegible] Pupils," *[illegible] and Psych. Measurement*, 1967, [illegible], 1047-1054.

[illegible], [illegible] and S. A. Kirk, "The Development and Psychometric [illegible] of the Revised Illinois Test of Psycholinguistic [illegible]," University of Illinois Press, Urbana, 1968.

[illegible] and [illegible], [illegible], 1968, [illegible].

[illegible], [illegible] and P. Schoggen, "Environmental Forces in the Lives of [illegible] Year Old Children in Three Population Subgroups," George [illegible] College for Teachers, Nashville, Tenn., January, 1971.

[illegible], "Disadvantaged Children and Their First School Experiences: Structural Ability and Change in the Test Performance of [illegible] Preschool Children," Educational Testing Service, Princeton, [illegible], 1972.

[illegible], V. C., "Disadvantaged Children and Their First School Experiences, ETS--Head Start Longitudinal Study," in [illegible] Stanley (ed.), [illegible], Johns Hopkins Press, Baltimore, Md., 197[illegible].

[illegible]man, V. [illegible] et al., "Disadvantaged Children and Their First School Experiences: Structure and Development of Cognitive Competencies and Styles Prior to School Entry," Educational Testing Service, Princeton, N. J., 197[illegible].

Smith, M. S. and J. Bissell, "Report Analysis: The Impact of Head Start," *Harvard Educational Review*, 1970, 40 (1).

Smith, M. S. et al., "Some Short-term Effects of Project Head Start: A Preliminary Report on the Second Year of Planned Variation--1970-71," Office of Child Development, Dept. of HEW, Washington, D.C. 1973.

Stanford Research Institute, "Implementation of Planned Variation in Head Start: Preliminary Evaluations of Planned Variation in Head Start according to Follow Through Approaches (1969-70)," Office of Child Development, Dept. of HEW, Washington, D.C., 1971.

Stearns, M. S., "Report on Preschool Programs: The Effects of Pre-School Programs on Disadvantaged Children and Their Families," Office of Child Development, Dept. of HEW, Washington, D.C. 1971.

Stern, V. and A. [illegible], "Preschool Environment Inventory," Bankstreet College of Education, mimeo, New York, fall 1967.

Stodolsky, S., "Defining Treatment and Outcome in Early Childhood Education," in H. Walberg and A Kopan (eds.), *Rethinking Urban Education*, Jossey-Bass, San Francisco, 1972.

Terman, L. M. and M. A. Merrill, *[illegible] Intelligence [illegible]*, Houghton Mifflin, Boston, 1960.

Walker, [illegible] *[illegible] for Preschool and Kindergarten* [illegible], Jossey-Bass, San Francisco, 1973.

Walker, D. [illegible] M. [illegible], and A. S. Bryk, "The Quality of the Head Start [illegible]," [illegible], [illegible], Mass., 1973.

Watts, [illegible], [illegible] Halfar, and N. Apfel, "Home Scale: An Observation Instrument for the Analysis of the Human or Material Environment of Children [illegible] Years," Harvard Graduate School of Education, Laboratory of Human Development, mimeo., May 1972.

Weikart, D. P., "[illegible] Intervention: A Preliminary Report of the Perry Preschool Project," Campus Publications, Ann Arbor, Mich., 1967.

[illegible], "[illegible] Curriculum, Teaching, and Learning in [illegible]," in J. C. Stanley (ed.), *[illegible] Programs* [illegible], Johns Hopkins Press, Baltimore, 1972.

[illegible], "[illegible] Cognitive Effects of Head Start Programs: A [illegible] Planned Variation--1971-72," Huron Institute, Cambridge, Mass., 1973.

[illegible], "[illegible] Payoffs [illegible] on Money Invested in Early Childhood Education, [illegible] Editorials: When should Schooling Begin?" [illegible], 1972, 53 [illegible].

White, B. L., "Preschool: Has It Worked?" in J. Chaffee (ed.), [illegible], the Educational Commission of the States, Denver, July/August, 1973.

White, B. L. and J. C. Watts, *Experience and Environment*, Prentice-Hall, Englewood Cliffs, N.J., 1973.

White, B. L. *et al.*, "Brookline Early Education Project," Brookline Early Education Project, Brookline, Mass., mimeo, 1972.

White, B. L. *et al.*, "Federal Programs for Young Children: Review and Recommendations, Office of Child Development, Dept. of HEW, Washington, D.C., 1973 (4 vols.).

Wright, H. F., *Recording and Analyzing Child Behavior* [illegible], Harper and Row, New York, 1967.