No Simple Answer: Critique of the "Follow Through" Evaluation

Ernest R. House, Gene V. Glass, Leslie D. McLean, and Decker F. Walker*

This study was supported by a grant from the Ford Foundation to the Center for Instructional Research and Curriculum Evaluation, University of Illinois at Urbana.

A distinguished panel examines in depth the Abt Associates, Inc. findings on the effectiveness of Project Follow Through. This panel disputes the report's conclusion that so-called "basic skills" approaches to the schooling of poor children are superior to other methods.

A recent U.S. Office of Education-sponsored report[1] has been widely publicized as suggesting a simple answer to the teaching of reading and math in the elementary school. Unfortunately, no simple answer has been found. The report concluded that so-called "basic skills" approaches to the schooling of poor children are superior to other methods. However, a careful inspection of the study reveals that the label "basic skills" is misleading and that approaches so named are no more effective than are other approaches. In fact, an approach that works best in one town may work worst in another.

The report on the effectiveness of Project Follow Through (a school-based extension of Head Start) was prepared by Abt Associates of Cambridge, Massachusetts. It reached an erroneous conclusion that "models that emphasize basic skills succeed better" by misclassification of the early childhood education models, by inadequate measurement of results, and by flawed statistical analysis. (Many of the errors derive from steps taken prior to Abt Associates' involvement.) A reanalysis shows that even the small advantages claimed for some models cannot be accepted at face value. In fact, participation in Follow Through classes was not shown to be either superior or inferior to schooling normally provided by the schools.

The major finding of the study is valid, however. The effectiveness of a teaching approach varies greatly from one school district to the next. This finding should be honored widely and serve as a basis for educational policy. Local schools do seem to make a difference. The peculiarities of individual teachers, schools, neighborhoods, and homes influence pupils' achievement far more than whatever is captured by labels such as "basic skills" or "affective" education.

The Follow Through evaluation compared 13 models of early childhood education in which over 20,000 students were taught for a four-year period. Many of the difficulties, such as in measurement and sampling, were derived from political considerations of the original evaluation design and were beyond the control of Abt Associates. Abt Associates wrote the final report and analyzed the data, but did not design the study or collect the data. We examined the report in detail and reanalyzed some of the data. We were led to our conclusions by the following considerations (some of which were noted by the analysts themselves):

[1] *Education at Experimentation: A Planned Variation Model.* Volumes IV—A-D. Cambridge, Massachusetts: Abt Associates Inc.; Richard B. Anderson, Project Director; Linda B. Stebbins, Deputy Project Director; Elizabeth C. Proper, Director of Analysis. April 15, 1977.

* Valuable assistance was given the panel by Elizabeth J. Hutchins.

Classification Problems

• *The classification of Follow Through models as "basic skills," "cognitive/conceptual," and "affective/cognitive" is misleading and mistaken.*

Nearly all Follow Through models taught reading, writing, and arithmetic. Those models labeled "cognitive," in fact, emphasized no less than those labeled "basic" what the public understands as basic skills—the ability to read with comprehension and to do problems drawn from ordinary life that require arithmetic. The models labeled "basic" concentrated on the mechanics of reading and arithmetic. Furthermore, many models seem to fit equally well in two categories. The whole classification scheme was vague and ill-defined. No check on the validity of the classification is reported, even though it shaped the findings.

• *Likewise, the distinction among outcome measures as "basic skills," "cognitive," and "affective" is untenable.*

For example, two very similar subtests from the Metropolitan Achievement Text (MAT) were arbitrarily placed in different categories.

Measurement Problems

• *The Metropolitan Achievement Test (MAT) was a good choice from among standardized measures of achievement in reading, language, and mathematics; but the Raven's Coloured Progressive Matrices was a poor choice to test for more advanced academic outcomes.*

The design of the Raven's makes it insensitive to school instruction. It is more an intelligence test than a test of school achievement.

• *The evaluation measured very few of the goals stated by the developers of the Follow Through models.*

Not only were outcomes such as improvements in personality and character not measured, but even such straightforward skills as the ability to read aloud, to write a story, or to translate an ordinary problem into numbers went unassessed. Both explicit and implicit goals of primary education in almost all schools are much broader than the measures used, and it would be reckless to suppose that the results of the testing indicate the attainment of these broader goals.

• *The tests strongly favored the models that concentrated on teaching the mechanics of reading, language, and arithmetic.*

Models emphasizing spelling, punctuation, capitalization, and similar details were favored by the tests over models emphasizing reading *comprehension*, mathematics *problem solving*, and the *use* of academic skills, or models emphasizing nonacademic outcomes.

• *Attempts to measure children's self-concept and tendency to take responsibility for their academic successes and failures produced unconvincing results.*

The instruments used required sophisticated observation and judgment of one's feelings and behavior. The correlations among the measures were low. The results on one measure did not resemble results on others. The pattern of results for different sites and models showed few positive effects, and these could easily have arisen by chance. In view of such considerations and the history of difficulty in personality assessment with young children, reliance on only these instruments was unwise. At the very least, the tests should have been given to some students a second time to determine if the scores were stable.

Analysis Problems

• *Questionable statistical definitions were employed in assessing model effectiveness.*

Abt Associates analyzed test scores so that the number of students in the models influenced how good the models appeared to be. When this irrelevant influence is removed from the analyses, a different order of effectiveness is found. The Abt Associates' ordering of models by effect is thus suspect.

• *An arbitrary choice of an analysis method made models labeled "basic skills" look better than other analysis methods would have.*

Abt Associates chose arbitrarily among statistical methods for testing whether the differences among Follow Through models were reliable (that is, "statistically significant"). In particular, their analyses favored showing models labeled "basic skills" to be reliably superior to the other types of model; our equally defensible

analyses show no reliable differences. The truth may well lie between, but no one can know exactly where.

• *Reliance on a single basic data analysis technique was unwise.*

Even under ideal circumstances, statistical adjustments such as analysis of covariance (ANCOV) cannot eliminate initial differences among groups, that is, differences among children when they entered the programs. ANCOV is likely to yield results biased against the more disadvantaged groups, an inescapable flaw the effect of which is difficult to estimate.

Fairness Problems

• *The scope of the measurement was biased.*

The evaluation was based on an exceedingly poor sample of the full domain of goals of the different Follow Through models. There were dozens of model goals not assessed. The outcome measures used favored the model emphasizing reading, language, and mathematics mechanics. The evaluation was biased not because of which measures were *included,* but because of which were *excluded* from the study.

• *The evaluation did not deliver what it promised.*

The original evaluators at Stanford Research Institute and the Office of Education were unable to deliver on promises they had made regarding what would be measured and assessed. In this sense, the evaluation was unfair to the sponsors of models who had been assured that instruments responsive to some important, but allusive learning outcomes could and would be developed. On the other hand, sponsors continued to accept large sums of money from the government even after this shortcoming of the evaluation became evident.

• *Other studies contradict the Abt findings.*

As evidence that many other outcomes exist, the model sponsors' own evaluations measuring progress toward their own goals conflict with some of Abt Associates' findings. For example, a group of the sponsors constructed measures of productive language and demonstrated gains there. Others have shown gains on traditional achievement tests. Although these studies are by no means above criticism, they demonstrate the possibility of arriving at rather different conclusions.

Our study has led us to the conclusion that "models that emphasize basic skills" are not better—at least not as demonstrated by the Follow Through evaluation. No approach was demonstrated to be better than the others. In addition, the differences between performance of Follow Through and non-Follow Through students were small—well within the range attributable to artifacts of the study. With almost all programs of this type, the new programs do no better on standardized measures overall than do comparison groups. Whether to attribute this phenomenon of no significant differences to the programs themselves or to their evalaution is debatable. The Abt Associates' analysts did well not to emphasize this finding.

The truth is more complex. Particular models that worked well in one town worked poorly in another. Unique features of the local settings had more effect on achievement than did the models. This variability in the benefits of school programs points up the shortcomings of one form of policy making. When fully understood, it can serve as the basis of a new educational policy that honors local individuality in place of general labels.

* The complete text of the critique will be published in the *Harvard Educational Review.*

Ernest R. House, University of Illinois; Gene V. Glass, University of Colorado; Leslie D. McLean, Ontario Institute for Studies in Education; and Decker F. Walker, Stanford University.