SCHOOL of PUBLIC POLICY

27
Westinghouse Learning Center/ Ohio University Report on Head Start

Douglas J. Besharov
Peter Germanis
Caeli A. Higney
and
Douglas M. Call

September 2011

Maryland School of Public Policy
Welfare Reform Academy
www.welfareacademy.org

Part of a forthcoming volume
Assessments of Twenty-Six Early Childhood Evaluations
by Douglas J. Besharov, Peter Germans, Caeli A. Higney, and Douglas M. Call

Note: This report is open to public comments, subject to review by the forum moderator. To leave a comment, please send an email to welfareacademy@umd.edu or fill out the comment form at http://www.welfareacademy.org/pubs/early_education/chapter27.html.

27
Westinghouse Learning Center/ Ohio University Report on Head Start

The federal Head Start program began in 1965 as an eight-week summer program for three- and four-year-old children to help "break the cycle of poverty by providing preschool children from low-income families with a comprehensive program to meet their emotional, social, health, nutritional, and psychological needs."[1]

In 1968, four years after the initiation of Head Start, the Office of Economic Opportunity (OEO) issued a request for proposals (RFP) to study the impact of the program. The RFP came as a response to the widespread perception that many of the early evaluations of the new Head Start program were "limited in scope and weak in design."[2] The RFP laid out many of the specifics of the evaluation, including its national scope and retrospective design. OEO awarded the contract to the Westinghouse Learning Corporation and Ohio University ("the Westinghouse group"). The study was not an experimental, randomized design. Rather, a sample of children who had attended a Head Start center between 1965 and 1968 and a matched sample of children from the same grades and schools who had not attended Head Start were administered a series of tests covering cognitive and affective development. This evaluation is included in this volume as it represents an early attempt to measure the impact of Head Start.

The Westinghouse group concluded that the summer Head Start program was "ineffective in producing any persisting gains in cognitive or affective development . . ."[3] The full-year program was seen as "marginally effective" in producing some cognitive gains, but ineffective in achieving improvements in affective development. The cognitive gains, however, failed to achieve

---

[1]U.S. Department of Health and Human Services, Administration for Children and Families, Head Start Bureau, "Head Start History," (Washington, DC: HHS, 2002), http://www.acf.hhs.gov/programs/hsb/about/history.htm (accessed November 8, 2005).

[2]Westinghouse Learning Corporation and Ohio University, *The Impact of Head Start: An Evaluation of the Effects of Head Start on Children's Cognitive and Affective Development, Volume 1*: *Report to the Office of Economic Opportunity* (Athens, Ohio: Westinghouse Learning Corporation and Ohio University, 1969), 13.

[3]Westinghouse Learning Corporation and Ohio University, 1969, 243.

the suggested criterion of practical relevance. Although the Westinghouse group conducted the evaluation reasonably well given the constraints placed on it, there are major questions related to selection bias that undermine the credibility of the findings. Moreover, the uneven implementation of Head Start centers during the program's early years limits the generalizability of the findings.

Program Design

Program group. The Westinghouse group selected a nationally representative sample of Head Start centers, some of which operated summer programs between the summer of 1965 and the summer of 1968 and others which operated full-year programs between 1965–1966 and 1967–1968. The final sample of children included 1,980 Head Start children and 1,983 comparison group children.

Services. About 70 percent of the children participated in the summer program and 30 percent participated in the full-year program. They received the basic Head Start package, although there was considerable variation in services provided across the centers.

The Evaluation. The Westinghouse group compared the cognitive and affective outcomes of the children who attended Head Start with those for children who were eligible for Head Start but did not enroll in the program. Although the purpose of Head Start extends beyond the goal of improving cognitive and psychological development for disadvantaged children, this evaluation only measured the impact of Head Start on these two factors.

Center selection. A random sample of 300 Head Start centers was chosen from a total of nearly 13,000 centers. The sample was then further reduced, with the goal of selecting 100 centers for the analysis. Centers were included only if they had operated for at least two years. The large oversample of centers was necessary because the Westinghouse group expected that many centers would not participate or would have to be excluded for other reasons. Indeed, a total of 225 centers were screened to obtain a final sample of 104 centers, consisting of 75 operating summer programs and 29 operating full-year programs.

Centers were excluded from the analysis for several reasons: they had an insufficient number of comparison group families in the target area (the major reason, affecting fifty centers); they did not operate a Head Start program before 1967 (twenty-one centers); school officials refused to participate (nineteen centers); they did not have summer program directors (eight centers); they were too small (seven centers); or various other reasons (sixteen centers). When one center was lost, it was replaced by another. This process was designed to maintain a 70:30 ratio between summer and full-year programs, which represented the distribution of programs nationally.

In selecting centers, the Westinghouse group used simple random sampling rather than stratified random sampling. They reasoned that no single characteristic would affect the findings,

and that stratification on a large number of factors would have required a much larger sample of centers. In addition, they did not have much information on the characteristics of centers beyond their addresses and whether they operated full-year or summer programs.

Child assignment. Children selected for the study were required to have lived in the areas served by the selected Head Start programs (target areas) and attended the same school system from the time of Head Start participation (or nonparticipation for the comparison group) until the time of the study, so as to assure that both groups of children were exposed to comparable neighborhood and school environments.[4] Only children who had not attended another Head Start center or preschool program and who would have met Head Start's eligibility requirements were included in the comparison group. A random sample of ten former Head Start children was selected first. A comparison group of ten children was then selected based on a matching procedure that included variables related to school performance, including sex, race/ethnicity, and kindergarten attendance. (Only eight children within each group were included in the analysis.)

The Westinghouse group carried out testing of former Head Start and comparison group children in first, second, and third grades. Selected children were tested in their schools. Sample children who did not take the test were replaced with children from the oversample. In cases where more than two children were lost, a child outside the original sample population served as a replacement. (This affected 1.7 percent of the entire sample.) The final sample included 3,963 children (1,980 Head Start children and 1,983 comparison group children), with roughly 1,500 in the first and second grades and about 850 in the third grade.

The Westinghouse group attempted to measure the short-term psychological and intellectual impact of Head Start on participating children. They analyzed Head Start's effects in two ways. First, they used the target area (or center) as the unit of analysis, using the mean score for each group, without adjusting for any socioeconomic characteristics beyond the adjustments made during the initial matching process. They then used the individual as the unit of analysis, controlling for parental occupational and educational level, as well as per capita income. Since the latter findings tended to confirm the former, this review focuses on the findings using the center as the unit of analysis, unless otherwise noted.

Major Findings

The Westinghouse group concluded that the summer Head Start programs were "ineffective in producing any persisting gains in cognitive or affective development . . ."[5] Full-year

[4]To exclude Head Start children who had migrated from the target area meant that the findings were not representative of the original group of Head Start children. To include them, however, would have made finding a similar comparison group more difficult and would have complicated the data collection effort.

[5]Westinghouse Learning Corporation and Ohio University, 1969, 243.

programs were seen as "marginally effective" in producing some cognitive gains, but ineffective in achieving improvements in affective development. The cognitive gains, however, failed to achieve the suggested criterion of practical relevance.

Cognitive. The impact of Head Start on cognitive development was based on the Metropolitan Readiness Tests (MRT) for children in first grade and the Stanford Achievement Test (SAT) for children in second and third grade.[6] Each test included a number of subtests, so outcomes were compared for both total scores and subscores. In addition, the Illinois Test of Psycholinguistic Abilities (ITPA) was administered to children in all three grades as a measure of children's powers of receptive and expressive communication, general intellectual development, and school readiness.

Summer program. There were few statistically significant impacts for Head Start children on the various measures of cognitive development, including language development, as measured by ITPA scores. (There were just twenty-two statistically significant differences in 418 tests, about what would be expected on the basis of chance alone. In seventeen of the twenty-two differences, the findings favored the comparison group.)

With respect to school readiness and achievement, there were no overall effects on the MRT for first graders, but nine out of ten statistically significant subscore or subgroup differences (out of 101 possible) were found favoring the comparison group. Similar findings were found using the SAT for second graders, with a statistically significant overall effect favoring the comparison group and twelve positive subscore or subgroup effects. There were no findings favoring the Head Start group. Although there was no overall effect on the SAT for third graders, there were five favorable subscore or subgroup effects favoring the Head Start group (and none favoring the comparison group).

As a result, the Westinghouse group concluded that the summer program was not successful in improving either the area of language development or school readiness and achievement.

Full-year programs. The study found some positive effects for the full-year program. With respect to language development, for all three grades, there were twenty-three statistically significant differences (out of 187 tests), with twenty-two favoring the Head Start group. Most of these were concentrated in the second grade. Similarly, there were seventeen statistically significant findings favoring the Head Start group on the MRT and SAT. Most were in the first grade, fewer in the second grade, and none in the third grade. In addition, the gains were greatest for children who attended centers in central cities, in the Southeastern part of the county, or in

[6]The SAT was not appropriate for children beginning the first grade, so the MRT was substituted to measure their learning readiness.

which the enrollment was predominantly black.

School readiness/performance. See above under "Cognitive" findings.

Socioemotional development. To measure the impact of Head Start on socioemotional development, the Westinghouse group used several tools: Children's Self-Concept Index (to measure the "degree of positive self-concept of children in primary grades"); Classroom Behavior Inventory (CBI) (to "assess children's motivation to achieve school learning"); and Children's Attitudinal Range Indicator (CARI) (to measure "positive and negative attitudes toward papers, home, school, and society"). There were few statistically significant findings with respect to Head Start's impact on the affective development of children for either the summer or the full-year programs.

Health. Data apparently either not collected or not reported.

Behavior. Data apparently either not collected or not reported.

Crime/delinquency. Data apparently either not collected or not reported.

Early/nonmarital births. Data apparently either not collected or not reported.

Economic outcomes. Data apparently either not collected or not reported.

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

Benefit-cost findings. Apparently a benefit-cost analysis was not performed.

Overall Assessment

Although the Westinghouse group conducted the evaluation reasonably well given the constraints placed on it, there are major questions related to selection bias that undermine the credibility of the findings. Moreover, the uneven implementation of Head Start programs during the program's early years limits the generalizability of the findings.

Program theory. Apparently, there is no specific theory detailed beside the general expectation that early intervention programs promote school readiness and improve developmental outcomes for children.

Program implementation. The rapid implementation of Head Start on a national scale with relatively little direct federal oversight resulted in considerable differences in local implementation. First, the recruiting practices of Head Start centers varied considerably. Some enrolled the most disadvantaged families, leaving a comparison group of less disadvantaged

families, while other Head Start centers served families that volunteered for the program and were probably more enthusiastic and motivated than the families in the comparison group.

Second, there was considerable variation in the duration of program operation among the full-year programs, ranging from three months to eight months.[7] This factor was not captured in the evaluation.

Third, these early Head Start programs differed in the experience and training of teachers, and the pupil/teacher ratio, and many other programmatic dimensions. As Lois-Ellin Datta, then Assistant Director of the Teaching and Learning Program at the National Institute of Education, observes in a 1979 analysis:

> Early in the program's history, there was no cadre of teachers trained in the pre-school education of disadvantaged children. The majority of teachers in the summer 1966 Head Start program, for example, had no special training beyond six-day college-sponsored training programs. One-third of these teachers had no previous pre-school teaching experience. Over 20% had no previous experience working with disadvantaged children. Center directors in the 1966 Summer program reported a mean of 14.3 children per teacher, with the number ranging from three to twenty-nine."[8]

This diversity makes it difficult to make any generalizations about the program as it operated in those early years or even shortly thereafter. Indeed, Datta notes that Campbell and Erlebacher argue for random assignment of children and programs to: "reduce the possibility of very poor programs canceling out the effects of very good programs. A study using only a few programs is particularly vulnerable to this problem. Westinghouse studied the graduates of only 104 Head Starts [centers] out of the 8,000 in operation in 1968."[9]

Edward Zigler, director of the Yale Bush Center in Child Development and Social Policy and the Sterling Professor of Psychology at Yale University, explains in a 1992 analysis that the negative findings should not have been a surprise, given the uneven quality of Head Start programs across the country:

---

[7]Marshall S. Smith and Joan S. Bissell, "Report Analysis: The Impact of Head Start," *Harvard Educational Review* 40 (1) (Winter 1970): 57.

[8]Marshall S. Smith and Joan S. Bissell, "Report Analysis: The Impact of Head Start," *Harvard Educational Review* 40 (1) (Winter 1970): 57.

[9]Lois-Ellin Datta, "Another Spring and Other Hopes: Some Findings from National Evaluations of Project Head Start," in *Project Head Start: A Legacy of the War on Poverty*, edited by Edward Zigler and Jeanette Valentine (New York: Free Press, 1979), 420.

> In short, there was no mystery behind the highly uneven quality of the Head Start programs in 1970. Despite the flaws in the Westinghouse report methodology, I doubt that any national impact evaluation at the time would have showed that Head Start had long-term educational benefits. Even if, as I suspected, a third of the programs were wonderful, their effects would most likely have been canceled out by an equal fraction of programs that were poorly operated.[10]

Assessing the randomization. The groups were not randomly assigned.

Assessing statistical controls in experimental and nonexperimental evaluations. The Westinghouse group based its findings on an "ex post facto" or "after only" research design. In the absence of random assignment, a matching procedure was used that paired children according to race, sex, and whether or not they had attended kindergarten. If a child was lost from the sample between the time of the parental interview in the summer and the fall testing, the Westinghouse group substituted a child from the oversample. This procedure was not described in enough detail to determine how it may have affected the comparability of the two groups. For example, it is unclear whether, if one sample member was dropped, its matched pair was also dropped and both were replaced, or if a replacement from the oversample pool was added for the one child. Replacing the pair would have been more likely to maintain the comparability of the two groups.[11]

Because the children were matched on the basis of just three child-specific variables, there were a number of statistically significant differences between the Head Start children and comparison group children.[12] The differences between Head Start and comparison group families varied by the type of program (summer or full-year) and grade level. In most cases, the Head Start children were from families with a slightly lower socioeconomic status (SES) than the comparison children. These disparities appeared to be more serious with respect to the summer program than the full-year program. The Westinghouse report compared program and comparison group

---

[10]Edward Zigler and Susan Muenchow, *Head Start: The Inside Story of America's Most Successful Educational Experiment* (New York: Basic Books, 1992), 154.

[11]In sixty-six cases (less than 2 percent), the sample loss exceed two children and additional children from the original population had to be recruited for the study. This loss, too, could have introduced bias.

[12]Ideally, the comparison of the families' socioeconomic characteristics should be based on the characteristics at the time the Head Start children went into the program. However, this information was collected at the time the parents were interviewed, which was in the summer of 1968, shortly before the children were tested. Thus, collection occurred about one year after program participation ended for first graders, and about three years later for third graders. This time gap could have created a misleading picture of the comparability of the two groups if Head Start participation itself affected some of the characteristics. For example, if Head Start participation encouraged mothers to stay at home with their children, it would have affected variables such as maternal employment and family income.

parents on sixteen demographic items. At the 5 percent level of significance, only one statistically significant difference would be expected. For the summer program, however, there were nine statistically significant differences for first graders, four for second graders, and just one for third graders. For the full-year program, there were three statistically significant differences for first graders, one for second graders, and three for third graders. The fact that so many differences existed suggests that the matching of children on the basis of a limited number of characteristics was imperfect. (This also complicates comparisons of Head Start's effects across grade levels.) Indeed, Harvard researchers Marshall Smith and Joan Bissell conclude in a 1970 analysis: "The non-comparability between the two groups of children raises strong doubts about the results' validity."[13]

In the analysis using the centers as the unit of analysis, the Hollingshead Two-Factor Index was used as a covariate. This index is based on the occupation and education of the father (or head of household) and is a measure of family SES. This analysis indicated that the Head Start group was more disadvantaged than the comparison group for all but the third graders who had taken the full-year program. The index, however, did not capture many of the other statistically significant differences between the groups and thus could not rule out other sources of potential bias. In particular, a major concern is possible differences in unmeasured characteristics that affect performance. For example, if children with the most enthusiastic and supportive parents enrolled in the program, the effect of Head Start may have been overstated. On the other hand, if Head Start centers sought out and enrolled children from the most disadvantaged families, the impact of Head Start may have been understated.

The Westinghouse data have been reanalyzed by other researchers, who have reached somewhat different conclusions. For example, Smith and Bissell, using a slightly different analytical approach,[14] found similar overall findings, but reached different conclusions about Head Start's effectiveness because they focused on distributional effects, rather than changes in average performance. They focused their reanalysis on full-year programs for first graders, because of their "desire to confound as little as possible the impact of Head Start and subsequent public school experiences."[15]

Smith and Bissell found that the Head Start children in the full-year program outscored the

[13]Marshall S. Smith and Joan S. Bissell, "Report Analysis: The Impact of Head Start," *Harvard Educational Review* 40 (1) (Winter 1970): 54.

[14]They made some changes in the data set to correct what appeared to be invalid information and used somewhat different covariates, but otherwise their analysis was quite similar.

[15]Smith and Bissell, 1970 79. They argued that this allowed them to focus on programs that have been operating for several years and were not likely to have some of the implementation problems experienced in the initial years (which would have affected the second and third grade children).

comparison group children on the MRT by about four points, placing them in the 44th percentile, compared to the 36th percentile for comparison group children. They also report that Head Start had important effects on the distribution of individual scores, with 40 percent of the comparison sample falling into the "below average" category on the MRT test compared to just 31 percent of the Head Start sample. These impacts were even larger for urban black children in full-year Head Start centers, where the Head Start children scored in the 48th percentile compared to the 32nd percentile for comparison children. Similarly, just 26 percent fell into the "below average" category, compared to 48 percent of the comparison group children. They conclude that these differences were large enough to be considered "educationally significant." Although the focus that Smith and Bissell placed on distributional effects is important, the very narrow focus of their reanalysis (one test for first graders) cannot be regarded as repudiation of the Westinghouse findings.

Burt Barnow, then at the Department of Labor, and Glen Cain, a professor of economics at the University of Wisconsin-Madison, also reanalyzed the Westinghouse data, attempting to correct for possible selection biases in who attended Head Start itself.[16] They focused primarily on the individual data (rather than the group data). They also replaced the variable used by the Westinghouse group to measure SES (the Hollingshead Index) with a "more complete" measure of SES. They examined a number stratifications consisting of program type (summer and full-year), grade level,[17] and single vs. two-parent family.

Most of the findings reported by Barnow and Cain confirmed the pessimistic findings of the original Westinghouse study, but they found a few statistically significant positive effects. In particular, black children who lived with both parents and participated in the summer program experienced a gain of about 1.6 points on the ITPA in the first grade. (They note that this was equal to about a 6.4 to 8 point increase in IQ points.) For white, first grade children, statistically significant positive effects were found from participation in both the summer and full-year programs, but only for those children living in single-parent families. These positive findings differed from those reported in the Westinghouse study. Barnow and Cain suggest that differences in statistical modeling could account for the difference. They also report, however, that Head Start effects were smaller and not statistically significant in the second and third grades. They did not conclude that this represented "fade out," because the selection procedures used to enroll children could have changed over the years or the program itself might have improved.

Donald Campbell and Albert Erlebacher, psychologists at Northwestern University, in a 1970 reanalysis, argue that the Westinghouse group compared nonequivalent groups, which led to

[16]Burt Barnow and Glen Cain, "A Reanalysis of the Effect of Head Start on Cognitive Development: Methodology and Empirical Findings," *Journal of Human Resources* 12 (2) (1977): 177–197.

[17]There were too few third graders for the full-year analysis, however.

an understatement of Head Start's effects.[18] They explained that the comparison group was from a higher socioeconomic group than the Head Start group, which, in turn, understated Head Start's effectiveness. They also argue that the statistical analysis employed by the Westinghouse group created a further bias. In response, Victor Cicirelli, the research director for the Westinghouse evaluation, while agreeing with some of the technical points, contends that any such biases were small, given the underlying similarity of the groups[19].

The various reanalyses of the Westinghouse data illustrate the type of debate generated by most quasi-experimental evaluations and the resultant uncertainty regarding the findings. Indeed, several of those critical of the study argued that random assignment would be needed to resolve the issue of selection bias. For example, Barnow and Cain conclude:

> The clearest message that our study reveals is the methodological one that evaluations must "model" the selection process that has assigned children to treatment and control groups. The selection process is not adequately known with existing Head Start data. Clearly, a random assignment is one design that readily permits an unbiased statistical estimation of the effect of Head Start.[20]

In the absence of random assignment, they recommended improving the measurements of the selection criteria to permit researchers to better model the selection process.

Sample size. Although the overall sample size was large, the findings were analyzed separately by type of program (summer and full-year) and by grade. As a result, some subgroup analyses were not possible, especially for the third grade cohort. Moreover, the small number of full-year centers and the many subgroup analyses performed meant that differences on key outcome variables would have had to be very large to be statistically significant.

Attrition. If children dropped out between the time of the parental interview in the summer of 1968 and testing in the fall, they were replaced by children in the oversample and, in some cases, from outside the matched sample group altogether. The extent of such attrition is not reported, except that in sixty-six cases, the attrition was large enough to exceed the oversample within a group within a center. No data are provided comparing the characteristics of those

---

[18]Donald T. Campbell and Albert Erlebacher, "How Regression Artifacts in Quasi-Experimental Evaluations can Mistakenly Make Compensatory Education Look Harmful," in *Compensatory Education: A National Debate*, edited by Jerome Hellmuth (New York: Burnner/Mazel, vol. 3, 1970): 185–210.

[19]Victor G. Cicirelli, "The Relevance of the Regression Artifact Problem to the Westinghouse-Ohio Evaluation of Head Start: A Reply to Campbell and Erlebacher," in *Compensatory Education: A National Debate*, edited by Jerome Hellmuth (New York: Burnner/Mazel, vol. 3, 1970): 211–215.

[20]Barnow and Cain, 1977, 195.

dropping out by their respective category. Thus, it is not possible to assess the potential effects of attrition. (As discussed below, since 225 centers had to be screened to obtain the final sample of 104 centers, attrition was a problem in this area as well.)

Data collection. The data collection relied on a various standardized tests and parent surveys, and the data sources are appropriate for the questions studied.

Measurement issues. The instruments used to measure Head Start's cognitive effects were widely used standardized tests: the Illinois Test of Psycholinguistic Abilities, the Metropolitan Readiness Test (first grade), and the Stanford Achievement Test. However, the three attitudinal measures were largely untested, and "the results on these measures tended to be dismissed."[21] (It could also be that they were "dismissed" because there were virtually no meaningful findings.)

Generalizability. Although the Westinghouse group claimed that the 104 Head Start centers were randomly selected, the fact that 121 out of 225 centers were dropped raises the concern that the findings might not be generalizable, despite the goal of a nationally representative sample. A survey administered to participating and nonparticipating centers suggested that there were few differences among them on factors such as program type, geographic area, racial/ethnic classification of the center, size of the city, and other characteristics of the centers,[22] but a low (45 percent) response rate among nonparticipating centers leaves this an open question. For example, if poorly performing centers were more likely to refuse to participate, the findings of the analysis could be misleading.

Smith and Bissell contend that there were two problems with the selection criteria for centers.[23] First, they note that a stratified random sampling plan would have produced a more representative sample. Although the Westinghouse group argued that it did not have detailed information on the characteristics of all 12,927 Head Start centers, Smith and Bissell point out that "region and locality strongly relate to academic achievement," and that a stratified approach could have been adopted that would have permitted stratification by "region and by urban vs. local location."[24] Moreover, they point out that since location is also related to racial/ethnic composition, "adequate representation of racial/ethnic backgrounds might have been

[21]Zigler and Muenchow, 1992, 64. Marshall and Smith note that ". . .the three tests used to measure affective development were constructed specifically for this evaluation. Although the tests may prove to be useful, it is not clear how to interpret them." *See* Smith and Bissell, 53.

[22]Westinghouse Learning Corporation and Ohio University, 1969, 43.

[23]Smith and Bissell, 1970, 65.

[24]Smith and Bissell, 1970, 66.

approximated."[25] They note that the resulting Westinghouse sample failed to capture any information on many regional, urban/non-urban, and racial/ethnic breakdowns.

The second problem in sample selection stemmed from the high dropout rate of centers. That over half of the centers were excluded for various reasons raises concerns about the representativeness of the final sample. For example, 23 percent of the centers were excluded because they lacked enough comparable non-Head Start children in the local schools. If these centers were representative of all centers in the nation that were serving most Head Start-eligible children, this undermines the representativeness of the remaining sample. Smith and Bissell point out that these centers could have operated in middle-class or rural areas where most eligible children were enrolled or that the centers were especially effective in enrolling eligible children. Another 10 percent of centers were dropped because they had only been in operation for one year. Again, Smith and Bissell caution that the sample would overrepresent "those centers that were funded in the early days of Head Start, and any idiosyncracies in the allocation of the early funds were carried over into the study."[26] Other centers were dropped for administrative reasons, because the centers were small, or for other reasons. Each of these losses changed the representativeness of the final sample, creating greater uncertainty about the generalizability of the findings. It is also unclear how this "attrition" differed by type of program (summer vs. full-year). Smith and Bissell conclude: "The more important point is that for one reason or another *the resulting sample may well under-represent small centers, centers from sparsely settled rural areas, centers from large predominantly black inner-city areas, and centers from predominantly middle-class areas.*"[27]

Replication. There have been many evaluations of local Head Start programs, but there is no single completed evaluation of the scope of the Westinghouse study, until the Head Start Impact Study (see chapter 13).[28]

Some reanalyses of the data by other academic researchers produced somewhat different findings and conclusions, generally showing more positive effects overall or for some subgroups.

Evaluator's description of findings. The Westinghouse group concludes that the summer Head Start programs were "ineffective in producing any persisting gains in cognitive or

[25]Smith and Bissell, 1970, 66.

[26]Smith and Bissell, 1970, 67.

[27]Smith and Bissell, 1970, 68, emphasis in original.

[28]*See* Michael Puma, Stephen Bell, Gary Shapiro, Pam Broene, Ronna Cook, Janet Friedman, and Camilla Heid, *Building Futures: The Head Start Impact Study, Research Design Plan* (Washington, DC: Administration for Children and Families, March 31, 2001), http://www.acf.dhhs.gov/programs/core (accessed July 30, 2002).

affective development."[29] Full-year programs were seen as "marginally effective" in producing some cognitive gains, but ineffective in achieving improvements in affective development. Even the cognitive gains, however, failed to achieve the suggested criterion of practical relevance, although the Westinghouse group concedes that developing this criterion is highly subjective.

The Westinghouse group examined several alternative interpretations and explanations of the findings, including: (1) the possibility that the program was effective, but the limitations of the design obscured these effects; (2) there was bias in the selection of centers; (3) the populations attending the summer and full-year programs were different, accounting for their different impacts; (4) some programs were poorly implemented; and (5) the subsequent schooling experiences of Head Start children "nullified" the early gains. The Westinghouse group examined each of these possibilities, but concludes that their findings represented an accurate picture of Head Start's effectiveness. Other analysts, however, raised these and other objections and came away with different conclusions.

Evaluator's independence. To carry out an independent evaluation, OEO selected the Westinghouse Learning Corporation and Ohio University through a competitive process.

Statistical significance/confidence intervals. Statistical significance was measured and reported at the 5 percent level.

Effect sizes. Effect sizes were not reported for each measured outcome. Rather, the Westinghouse group selected a criterion of one-half of one standard deviation (0.5 SD) as a difference large enough to "be considered of sufficient practical relevance and worthwhileness."[30] The Westinghouse group then compares the difference between the Head Start and control groups on selected testing instruments to the standard deviation of the population (on which the instrument is standardized) to determine if Head Start's effect was at least 0.5 SD. None of the differences found in the report, either in total, or for various subscores or subgroups, met the Westinghouse group's established criterion of practical relevance.

The Westinghouse group justifies their selection of 0.5 SD as the benchmark for practical relevance by referring to this convention in educational and psychological literature. They note that "Cohen also speaks of 0.5 SD as a convention to use, calling it 'the operational definition of a medium effect size.'"[31] And, using grades as an example, they equate this to an increase from a C

[29]Westinghouse Learning Corporation and Ohio University, 1969, 243.

[30]Westinghouse Learning Corporation and Ohio University, 1969, 164.

[31]Jacob Cohen, "Some Statistical Issues in Psychological Research," in B. Wolman, ed., *Handbook of Clinical Psychology* (New York: McGraw-Hill, 1965), 101, quoted in Westinghouse Learning Corporation and Ohio University, 1969, 164.

to a C+.[32] (See Appendix 1 for a further discussion of effect sizes and their interpretation.)

Sustained effects. The evaluation examined impacts through age fourteen, nearly ten years after program completion.

Benefit-cost analysis. Apparently not performed.

Cost-effectiveness analysis. Apparently not performed.

[32]Westinghouse Learning Corporation and Ohio University, 1969, 165.

Commentary

Burt Barnow*

The paper appears to summarize my work, the original study, and the commentators fairly. Of course, there are more recent evaluations of Head Start that make use of better designs and should play a much larger role in assessing Head Start's utility.

Note: This report is open to public comments, subject to review by the forum moderator. To leave a comment, please send an email to welfareacademy@umd.edu or fill out the comment form at http://www.welfareacademy.org/pubs/early_education/chapter27.html.

*Burt Barnow was a research economist at the Department of Labor and is currently a professor at Johns Hopkins University.