

4 Million and Counting: Reflections on 40 Years of Head Start Research in Support of Children and Families

Presenter: John M. Love

1965--41 years ago. It's hard to remember what an amazingly eventful year 1965 was: On January 4, President Lyndon Johnson proclaimed the Great Society in his State of the Union address. In February, the U.S. began bombing North Vietnam in operation Rolling Thunder. On Feb. 2, Malcolm X was assassinated. March 7 became known as "Bloody Sunday," when civil rights marchers began their trek from Selma to Montgomery and were violently confronted by state troopers. A month later, in a one-room schoolhouse in Texas, President Johnson signed into law the Elementary and Secondary Education Act of 1965, which we all know in its latest incarnation as No Child Left Behind. On August 11, the Watts riots began. That year, in Britain, Winston Churchill died—and J. K. Rowling was born. And at the Grammys, *Mary Poppins* won best recording for children.

So with a spoonful of sugar and a ton of sweat and tears, a vast new program for children and families was underway. Within 3 years, Donald Rumsfeld, the new head of the Office of Economic Opportunity (OEO) would have Head Start as part of his responsibility, guiding the country through an early phase of a different war—the War on Poverty.

Research began immediately. One of the first studies in the new Head Start context is a dissertation by an enterprising young University of Chicago doctoral candidate, Diana Slaughter, who investigated the maternal antecedents of African American Head Start children's academic achievement. One reason we don't hear about that study and the other 506 that came out in the first six years of Head Start is because everybody—researchers, program administrators, and politicians—had to cope with the fallout of the so-called "Westinghouse report."

In reflecting on my own involvement in Head Start research since the early 1970s, I have tried to think about lessons for the research we are doing now and the research to come. I see moments of brilliance, and I see some ventures into blind alleys. I see important themes in the 40 years of Head Start research. I am going to focus on the Head Start research *process*. I will be selective, leaving out parts that others might have emphasized.

Head Start was soon subjected to a major accountability test. In 1968-1969, the Westinghouse Learning Corporation and Ohio University located Head Start graduates in 1st, 2nd, and 3rd grades, matched them with children who had not been in Head Start, and drew their comparisons. One of the major legacies of the Westinghouse report is the perception that Head Start effects fade out by third grade. But this study was limited to cross-sectional comparisons. Third graders, who had been through the rocky summer of the 1st year of Head Start, were compared with 1st graders who came along later. The poorer performance of the 3rd graders was described as a "fade-out effect," even though there were no data on the effect for these children before it faded. I wonder the extent to which the "myth of fade out," as Steve Barnett has called it, stems initially from a study that was not designed for answering that question. But many other studies took place in the early years. Their diversity foreshadowed the range of research that we see in the 21st century.

Although Westinghouse was the biggie (politically), the most rigorous quasi-experimental study of Head Start's early years was reported *just last year*. Jens Ludwig at Georgetown and Douglas Miller of UC Davis collaborated to apply an ingenious regression discontinuity analysis to data from Head Start's first decade. Because OEO provided technical assistance to the 300 poorest counties in the country to help them prepare proposals for Head Start funding, Ludwig and Miller looked for "discontinuities in Head Start funding at the OEO cutoff for funding eligibility" (Ludwig & Miller 2005, p. 11). Using OEO's cutoff, they found a large discontinuity for Head Start funding, but not for other federal social spending. Applicants in the counties above the cutoff (the "treatment" group) were much more likely to receive Head Start funds, and this funding difference "persisted through the late 1970s" (p. 3). The authors argue that the results probably apply to Head Start programs between 1965 and the late 1970s, showing that mortality rates from causes that could be affected by Head Start—*anemia, meningitis, and respiratory problems*—were significantly lower as a result of Head Start funding; Head Start children were more likely to complete high school, by about 3 to 4 percentage points (based on Census data analysis); former Head Start children completed about one-half year more of schooling than control children did (NELS2000 data); no effects were found on 8th grade reading or math scores (NELS 2000 data). It is no wonder that health-related effects were found. Look at what Head Start accomplished in its first two years, according to OEO: a) 98,000 children with eye defects were discovered and treated, b) 900,000 dental cases were discovered, with an average of 5 cavities per child, c) 740,000 children were discovered without vaccinations against polio and were immunized, and d) more than 1,000,000 were found not to be vaccinated against measles and also were vaccinated.

Back in the 1960s and 70s, research continued at a rapid pace, in spite of—or perhaps because of—the criticisms of the Westinghouse study. As the responsibility for Head Start shifted from OEO to the Office of Child Development in the Department of Health, Education, and Welfare, program improvement and research were taken very seriously. OEO was spending between 1 and 2 % of Head Start's \$340 million budget on research and evaluation (Vinovskis 2005, p. 103). For perspective, if ACF spent what seems like such a small percentage on research today, we would see a budget of \$100 million or more for Head Start-related research.

I have spent some time on ancient history because I worry about history repeating itself. We easily denounce the Westinghouse study—as though we could not imagine anyone doing an *ex post facto* study these days. I see more and more studies using the ECLS-K data set to answer questions about Head Start and other preschool programs' "effectiveness. To give just a few examples, in just the past year, we have seen such findings as children who had attended prekindergarten scored about a fifth of a standard deviation higher ($ES=.19$) on a reading and math skills assessment at school entry than comparable children who spent the previous year at home (Magnuson, Meyers, Ruhn, & Waldfogel 2004); entering preschool between ages 2 and 3 is better than either earlier or later for effects on later cognitive skills (Loeb, Bridges, Bassok, Fuller, & Rumberger 2005), and preschool programs lead to higher school achievement outcomes, but more behavior problems in kindergarten (Rumberger & Tran 2006).

All this, when everything thing that is known about the children's preschool experience came from parents' recollections at the beginning of kindergarten. Now, I am sure these researchers know more than I do about how to control for all the selection factors that may have affected which kinds of children attended which kinds of programs and at what ages and for how long,

but I still worry. The Early Childhood Longitudinal Study has produced a wonderfully rich data set, and the studies I have cited explore important issues, but ECLS-K was not designed to answer causal questions about the effects of children's experiences before kindergarten.

Two years before Head Start began, 123 children participated in a preschool program that eventually made them the most policy-famous children on the planet. Those 123 children in Michigan, cast a large shadow over the 24 million who have gone through Head Start since then. There is no doubt that the cost-benefit results of the Perry study have had a powerful positive influence on public policies supporting early childhood education. Ironically, however, the very success of that experiment has had a negative influence in that it holds up a standard for expected effects that represent 1960s conditions and not 21st century community realities.

For example, after one year of preschool, the treatment-control impact on the PPVT had an effect size of .83. The effect size was 1.22 for the Stanford Binet impact. As everyone knows by now, contemporary interventions like Head Start and Early Head Start have more modest-sized impacts, mostly under a quarter of a standard deviation. A number of writers seize on this contrast to say Head Start is not as effective as the older program. I find that comparison troubling. The size of the impact depends on how well the program children did—but also on how well the control group children did. The context in the 1960s was vastly different than it is today. Few alternatives to the intervention were available. Few states, for example, had universal *kindergarten* programs, let alone preschools. And the Perry children were a very disadvantaged sample. The control group entered the study with a mean PPVT score of 62 and an average Stanford-Binet of 78.5. And remember the health statistics in those days, when three-quarters of a million kids in Head Start needed to be vaccinated against polio? And children had an average of five cavities?

To compare effects across studies, we should at least try to approximate comparable samples. I will show what I mean with an example from the Early Head Start impacts on aggressive behavior problems. The effect sizes show the magnitude of the *reduction* in parent-rated aggressive behavior problems on the Achenbach scale. The impact for African American children was *three times* the size of the overall impact. If one wants to make comparisons with Perry or Abecedarian, we should look not at the .11 but at the .35, since those programs enrolled only African American children. What I am trying to convey here is in no way a criticism of the Perry Preschool study, but rather a sense of how we might best view the “Perry phenomenon.”

In February 1972, in my first year at High/Scope, Dennis Deloria and I spend a week at Abt Associates working with colleagues on a proposal to conduct an evaluation of the Home Start Demonstration Program. We struggled with the section on instruments to propose for measuring child outcomes. I wrote, “while we can propose strong, highly reliable, and valid measures of children's cognitive development, few such measures exist in the social-emotional domain.”

Many have tackled this problem. In fact, in the late 1970s and early 1980s Head Start made a huge push to develop measures across all domains of development. The push actually began in 1974 when OED commissioned the Rand Corporation to design a “national evaluation of the social competence effects of the Head Start program” (Raizen & Bobrow 1974, p. iii). One of the first things Rand did was to warn about “difficulties” in obtaining “interpretable and meaningful data” in any national evaluation without better

child outcome measures. Rand was guided by the *OCD-Head Start Policy Manual*, which—thanks to the efforts of Ed Zigler, Clennie Murphy, Ray Collins, Jenni Klein, Sol Rosoff, and others—had just been published in January 1973. This first systematic articulation of performance standards grew out of OCD’s Head Start Improvement and Innovation (I&I) effort. The performance standards clearly defined “social competence,” which became the centerpiece of Rand’s recommendations. Social Competence is, “the child’s everyday effectiveness in dealing with his environment and later responsibilities in school and life. Social competence takes into account the interrelatedness of cognitive and intellectual development, physical and mental health, nutritional needs, and other factors that enable a child to function optimally” (Raizen & Bobrow, 1974, p. 3).

The Head Start Bureau had wanted the national evaluation to begin in fall 1974. Rand got a one-year delay for more planning, including test development (Raizen & Bobrow 1974, p. 4). But it was another 25 years before the national impact study got underway with the contract to Westat.

After accepting the 1-year delay, OCD funded a multi-year effort by Mediac Associates and several subcontractors to develop measures that would address the concerns Rand and others raised. Aptly called “the Head Start Measures Project,” it identified four domains for developing measures: cognitive, social-emotional, health and physical development, and applied strategies. This was more than 10 years before the nation’s governors created the first education goal with its five domains that were notable for including “approaches to learning” as a domain parallel in importance to the cognitive, social, language, and health domains. As everyone knows, when Head Start published its *Leaders Guide to Positive Child Outcomes* in September 2003, it took a similarly broad, comprehensive view, with a framework comprising eight broad domains (ACYF 2003). Clearly, Head Start has been leading the way.

In 2003, ACF’s Office of Planning, Research and Evaluation launched a new initiative called “Design Options for the Assessment of Head Start Quality Enhancements.” The basic concept is that when a new program idea arises, it is put to a rigorous experimental test *before* it’s widely disseminated. The Head Start Bureau (now the Office of Head Start) has been criticized for rolling out Early Head Start for tens of thousands of families before the programmatic approaches were tested. Although that criticism is only partially valid, a strong case can be made for getting new interventions “right” before requiring 49,000 teachers in 1,600 programs to change their practices. With a planned variation approach, one would try out, for example, a new plan for improving teacher training on a small scale, collect data using a rigorous experimental design, and then, if the results support it, spread the idea more widely.

The first planned variation study in Head Start, however, began 34 years earlier. In 1969, eight curriculum developers were funded to implement their curricula in multiple communities each. Results were, unfortunately, not very conclusive. This happened shortly after the U.S. Office of Education launched the Follow Through planned variation experiment. After the Head Start results were in, but before the Follow Through study was completed, Alice Rivlin and Michael Timpane convened a special meeting of the Brookings Panel on Social Experimentation to assess what had been learned about doing such experiments. The title of the conference, and subsequent book, was *Planned Variation in Education: Should We Give Up or Try Harder?* (Rivlin & Timpane, 1975). Recognizing the flaws in the two experiments, participants asked themselves: if we care about effective education for poor children, should we try to carry out better planned

variation studies, or is the basic notion of planned variation doomed to failure? *Should we give up or try harder?* (Rivlin & Timpane 1975, p. 2, emphasis added). In the Head Start arena, it looks like ACF has decided to try harder.

In the 1970s, the Office of Child Development's I&I initiative launched a series of demonstration programs that were sort of planned variation experiments. They studied variations but without the experiments. Still, new program ideas were tried out, typically in a limited number of program sites, through grants that augmented existing Head Start programs. Such demonstration programs were a major programmatic and research initiative of Head Start that had its heyday in the mid-1970s, but have continued in every decade since.

The Office of Child Development morphed into the Administration on Children, Youth and Families (ACYF), which continued trying out ideas in demonstration programs and evaluating them through the 1980s and 90s. But these studies still did not address the basic questions program and policy folks had about the effectiveness of Head Start. That really began after the so-called "Blueprint Committee," in 1990, urged Head Start to address questions about which Head Start practices maximize benefits for children and families with different characteristics under diverse circumstances (U.S. Department of Health and Human Services 1990).

My perception is that Head Start has a love-hate relationship with experimental designs. I have worked on two experimental studies—Home Start in the '70s and Early Head Start since 1995. But we did not see a national impact study of Head Start with a representative sample until the Head Start Impact Study. Of course, we have the FACES study, which came about in response to the government's need for performance measures. For the better part of a decade, it has yielded rich information about programs and their children in a nationally representative sample. This is very useful for knowing what happens in a representative sample of Head Start programs, but it does not tell us about Head Start's effectiveness.

We need diversity in research designs, and experiments are not always possible, but one of my concerns is that observational studies can produce vastly different findings from clinical trials—just think about the common medical wisdom about the benefits of hormone replacement therapies for post-menopausal women that was turned upside down by clinical trials demonstrating the risks of taking hormones.

Randomized trials are hard, and we do not always do them well. Just this spring, the results of a huge dietary study produced "disappointing" results in failing to show the benefits of low-fat diets for women. After 8 years, the treatment and control groups had similar rates of some cancers and cardiovascular disease. Then clinicians observed that the intervention group did not reduce the fat content of their diets to the target levels—the treatment was not implemented. The investigators acknowledged the flaws in the execution of the study. We need to make the same acknowledgments when, for example, a home visiting program fails to show a difference in children's school readiness—but we look at its implementation, and we see that the home visits occurred less frequently than called for, and that many families left the program early.

In the Head Start Impact Study, the intervention group didn't take all the pills they were supposed to, and the control group got some of the same medicine. These problems could just as well have occurred in a nonexperimental comparison design. But the rigorous protocol for a

randomized trial increases the investigators' vigilance for noncompliance. The main point I want to make is that those who do not think randomized designs are necessary should understand that design decisions may affect study conclusions and their policy implications.

Partnership is the theme of this 2006 conference, but research-program partnerships have existed from the early days of Head Start research. At the time of the Home Start evaluation, we did not highlight the notion of partnership, but in 1976, I wrote about two people, Dr. Esther Kresh, the National Evaluation Project Officer and Dr. (Ruth) Ann O'Keefe, National Home Start [Program] Director, who were instrumental in establishing a close-knit integration of evaluation and program activities from the initial conceptualization of the project that minimized most of the major problems typically faced by other evaluators on similar projects.

Head Start has spawned many research partnerships and consortia, including Head Start-University partnerships, the Child Outcomes Research and Support Consortium, the Early Promotion and Intervention Research Consortium, the Head Start Quality Research Center (HSQRC) Consortium, and the Interagency School Readiness Consortium. ACF research staff has collaborated with the National Institute of Child Health and Human Development (NICHD) on the Study of Early Child Care and Youth Development and the Early Head Start father studies and with the Department of Education on the ECLS-K and ECLS-B studies.

In Early Head Start, we saw the most elaborate partnership I have ever experienced. ACF created a consortium that included program and research staff at the federal level, program staff and university researchers at the local level, the national evaluation contractor and local researchers (with program directors at times), and interactions with the national T&TA system. The national and local researchers jointly created a publications policy that went beyond restrictions and encouraged collaboration. To this end, the consortium structure included nine working groups of like-minded researchers who have been collaborating on analysis and writing in areas as diverse as childcare, fathers, and measures.

One outcome of the research collaboration has been an incredible number of papers, presentations, book chapters, journal articles, and books. And a unique feature of the national impact reports was the integration of local researcher perspectives. The interim and final reports included appendixes devoted to brief local research reports—21 of them in a special volume of our final report, titled "Local Contributions to Understanding the Programs and Their Impacts." In addition, if you read Volume I, the main report on our implementation and impact findings, you'll find 22 full-page text boxes interspersed, summarizing local research studies that complement the findings being reported from the national evaluation.

I noted how the success of the Home Start evaluation was at least partially a result of just a few people who were committed to collaborating. My understanding is that the same is true for the Early Head Start study. Without an associate commissioner for Head Start like the late Helen Taylor and an evaluation project monitor like Helen Raikes, the spirit of partnership would not have infused the collaboration activities we tried to carry out.

So we come to the end of these reflections. We have seen a small part of the legacy from 41 years of Head Start research that continues to undergird today's research. We have seen: Issues in comparing effects across studies, particularly from different eras; refining research questions

and conducting subgroup analyses to learn how the program can be more effective, and to get away from simply asking *whether* it is effective; the need—still—for measures that really get at the outcomes Head Start is committed to achieving; the important role of experimental designs and use of planned variation studies; the value added of research-research and research-program partnerships, and the importance of a few highly committed leaders.

Head Start research has also confronted, dealt with, and sometimes helped to move us forward on a number of issues I have not touched on today, including concerns about assessing language-minority children, research aimed at understanding program implementation, and effectiveness for children with disabilities, research to understand how programs can be more effective with the highest risk families, and many other areas.

We now share a common heritage in this body of research that contains many lessons. But many challenges remain. Let me mention ones I think are particularly important. We have not solved the challenges of achieving fidelity of implementation when small-scale programs are taken to scale. In fact, there is evidence that we haven't even solved the challenges of replicating a successful program on a *small* scale.

Thirty years ago, Abt Associates completed a huge multi-curriculum, multisite, national, planned variation study of Follow Through *without* being able to conclude that one curriculum was clearly better than any other. The major problem was more variability across the sites where a particular curriculum was implemented than there was across the different curricula. On Monday, some of you attended the PCER poster symposium where one of the underlying themes related to the challenges of successfully implementing curriculum changes in public school preschools today. Clearly, we have a ways to go.

In summer 1965, the Office of Economic Opportunity launched the first Head Start programs as part of the War on Poverty. Head Start has been the doorknob for 24 million children so far. Research in the early years included practically every design, measurement, and analytic challenge that program evaluators can imagine. The ways in which researchers dealt with those challenges provided lessons that have guided (and sometimes been ignored) in subsequent research and evaluation studies. By paying attention to those lessons, we can ensure more effective use of research and evaluation for program improvement and policy formation. The next 24 million children and families are depending on our research—our thoughtful and careful research—to increase their chances in life—by reducing chance.