SCHOOL of PUBLIC POLICY

10
Even Start Family Literacy Program (1st National Evaluation)

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/chapter10.html.

10
Even Start Family Literacy Program (1st National Evaluation)

The Even Start Family Literacy Program was originally authorized by Congress in 1965 and reauthorized in 1994. The Even Start program is intended to: "help break the cycle of poverty and illiteracy by improving the educational opportunities of the Nation's low-income families by integrating early childhood education, adult literacy or adult basic education, and parenting education into a unified family literacy program."[1]

Robert St. Pierre, Janet Swartz, Beth Gamse, Stephen Murray, Dennis Deck, and Phil Nickel, researchers at Abt Associates Inc. (the "Abt team"), conducted the first Even Start evaluation which covered the program between 1991 and 1993. The evaluation used random assignment, but was plagued by numerous problems in the randomization process. Other problems included a small sample size and high levels of attrition in the study sites. These problems were so serious that no conclusions could be drawn regarding the impact of the program. The second Even Start evaluation covered the 1993 to 1997 period. No control or comparison groups were used in this evaluation, however, so it could not address questions about the program's impacts and, hence, it is not included in this volume.[2] The third evaluation was conducted by Abt Associates for the 1997 to 2001 period and this random assignment evaluation appears to have been implemented without the problems that plagued the first evaluation (see chapter 9).

Program Design

Program group. At the time of the evaluation, Even Start was targeted to families with an adult who was eligible for adult education programs under the Adult Education Act or was within the state's compulsory school attendance age and had a child younger than age eight. For evaluation purposes, in five sites the sample was restricted to families that had a child ages three to four.

---

[1]Cited in Beth C. Gamse, Dylan Conger, Dean Elso, and Maria McCarthy, *Follow-Up Study of Families in the Even Start In-Depth Study: Final Report* (Cambridge, MA: Abt Associates Inc., 1997), 1.

[2]Fumiyo Tao, Beth Gamse, and Hope Tarr, *Second National Evaluation of the Even Start Family Literacy Program: Final Report* (Washington, DC: Fu Associates, 1998).

The families randomly assigned to Even Start were 40 percent single-parent families; 27 percent with a high school diploma or GED; 14 percent white, 24 percent African American, and 50 percent Hispanic; and 68 percent of parents were between the ages of twenty and twenty-nine.[3]

Services. At the time of the evaluation, Even Start was a family focused program, providing services to both children and parents. Grantees had considerable flexibility in designing services to meet the needs of the low-income families, but all were required to offer three core services: (1) adult education (included GED preparation classes) to develop basic educational and literacy skills, and on a more limited basis, vocational skills; (2) early childhood education services to meet the needs of children from birth to age seven; and (3) parenting education to help parents support the educational growth of their children. The context in which these core services were provided depended on the population that the center served. Region, availability of transportation, and need for individualization of support were considered in determining the most appropriate mode for providing services, either through home visits or center-based activities. The core services were typically provided in an integrated manner—in the same setting, with the same instructors, and through parallel or joint activities. In addition, Even Start projects offered referrals to a range of support services, such as child care, transportation assistance, health care, nutrition, and an array of other family support services. When these services were not locally available, some Even Start projects provided them directly.

The Evaluation. The first national evaluation of the Even Start program, funded by the Department of Education and carried out by Abt Associates, included an "In-Depth Study" (IDS) of ten projects.[4] These projects were selected from the first cohort of seventy-three Even Start grantees. Selection was based on location, willingness to participate, and the level of program implementation. Only five of the ten projects employed random assignment.[5] In each site, about forty families were randomly assigned, half to a program group and half to a control group. (Families were recruited in the summer/fall of 1991 and randomly assigned between the fall of 1991 and January 1992.) Because of the small number of families in each site, the Abt team pooled the data for analytic purposes, resulting in a total of 101 Even Start families and 98 control group families. They focused on families with a child three or four years old in order to examine the program's impact on school readiness and other preschool outcomes.

---

[3]Robert St.Pierre, Janet Swartz, Beth Gamse, Stephen Murray, Dennis Deck, and Phil Nickel, *National Evaluation of the Even Start Family Literacy Program: Final Report* (Washington, DC: U.S. Department of Education, Office of the Under Secretary, January 23, 1995), 153–4.

[4]The random assignment portion of the first Even Start evaluation is a small part of the total evaluation effort that cost about $2.9 million. The third Even Start evaluation (the second with a random assignment component) cost about $3.6 million.

[5]Some projects dropped out of the study because they were not able to implement random assignment or because they were at full capacity and could not enroll a new cohort of families.

Data were collected on a "target" adult and a "target" child at the time of entry into Even Start (or the control group), about nine months later, and again about eighteen months after entry. All families were included in the data collection, even if they never enrolled in any of the three core services or participated for just a short period of time. This procedure led the Abt team to conclude that the findings, although unbiased, would be a conservative estimate of Even Start's impacts.[6] The findings described below were limited to the subset of children or adults from whom data was available for each of the three data collection points.

Major Findings

For Even Start children, there were no statistically significant effects on various cognitive tests or measures of school readiness eighteen months after program enrollment.[7] For adults, there was no impact on literacy, but participants appear to have been more likely to have received GEDs.

Cognitive. The Abt team used the Peabody Picture Vocabulary Test-Revised (PPVT-R) to measure child literacy-related skills. Eighteen months after enrollment, there was no statistically significant difference between Even Start and control group children on the PPVT-R vocabulary test. Nor did the Abt team detect any statistically significant differences eighteen months after enrollment "using a brief set of items designed to tap children's emergent literacy skills."[8]

School readiness/performance. The Abt team assessed the children's school readiness using the Pre School Inventory, an inventory of concepts considered "important for preschool children to know before entering school."[9] After eighteen months, there were no statistically significant differences on the Pre School Inventory between Even Start and control group children.

Socioemotional development. Relevant tests apparently not administered or results not reported.

Health. Data apparently either not collected or not reported.

Behavior. Data apparently either not collected or not reported.

---

[6]St.Pierre et al., 1995.

[7]St.Pierre et al., 1995.

[8]St.Pierre et al., 1995, 8-17

[9]St.Pierre et al., 1995, 8-1

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. The Abt evaluation found no statistically significant effect on the functional literacy level of adults in the program, as measured by scores on the Comprehensive Adult Student Assessment System (CASAS). But Even Start apparently increased the GED attainment of participants. The percentage of Even Start adults with a high school diploma or GED was initially lower than the control group (21 percent vs. 29 percent), although the difference was not statistically significant. However, the two groups were essentially even at the first follow-up, about nine months after enrollment (32 percent vs. 31 percent). At the second post-test, eighteen months after enrollment, a higher percentage of adults in the Even Start group had a high school diploma or GED (40 percent vs. 33 percent). Thus, the program group gained 19 percentage points, compared to just 4 percentage points for the control group, a statistically significant difference.[10]

The Abt team found no statistically significant differences in the reading and writing habits of the two adult groups eighteen months after enrollment, as measured by self-report questions about the frequency and type of reading and writing activities. The Even Start evaluation used the Pearlin Mastery Scale to measure parents' sense of control and the Center for Epidemiologic Studies Depression Scale (CES-D) to assess the percentage with high depressive symptoms. There were no statistically significant differences on these two measures between the groups after eighteen months. The Home Observation for Measurement of the Environment (HOME) inventory was used to assess the home environment. There was a statistically significant impact favoring Even Start families in a measure reflecting the number of different types of reading materials in the home.

Finally, there were no statistically significant effects on employment or receipt of government assistance among adults in the program.

Benefit-cost findings. The cost of Even Start for families in the IDS during the 1991–1992 program years was $5,366 (in 2005 dollars), of which 65 percent was provided by federal Even Start funds. This amount generally included the costs associated with one adult and one child per family. Given the absence of positive findings, it is doubtful that the program produced savings to offset these costs.

[10]The level of GED attainment was significantly higher in the sites conducting random assignment than nationwide, so these findings cannot be generalized to other Even Start sites operating at the time. In over half of all sites, less than 5 percent of adults attained a GED.

Overall Assessment

The first National Even Start evaluation used random assignment, but was plagued by numerous problems in the randomization process. In addition, the small sample size and high levels of attrition in the study sites created serious data problems. These problems were so serious that no credible conclusions could be drawn regarding the impact of the program.

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. The Abt team also hypothesizes that "Even Start also should have direct, but longer-term, effects on the economic self-sufficiency of adults due to increased participation in adult education and mediated by subsequent enhanced literacy skills."[11] Within this context, the evaluation is appropriate.

Program implementation. Little information has been published regarding implementation of the Even Start program in the five randomized study sites. Many of the children who began participating, however, had dropped out by the time of the second post-test. There were also high levels of nonparticipation among adults, as the Abt team describes:

> Maintaining parent participation is a continual challenge, according to Even Start project staff. For example, some staff expressed concern that the requirement for families to participate in adult education, parent education, and early childhood education places a great burden on families. Others cited the fact that many adults who are unemployed are not used to the structure of the school and may not have had positive school experiences; both of these factors affect attendance in program activities.[12]

High levels of nonparticipation, in turn, would be expected to limit the magnitude of program effects. Additionally, the measured or differential impacts of Even Start on children may have been reduced by the high levels of participation in early childhood programs among control group children.

Assessing the randomization. The original evaluation plan of the IDS called for random assignment in ten sites, with twenty families assigned to the program group and twenty to the control group, for a total of 400 families. However, only five of the ten sites implemented

---

[11]Robert St.Pierre, Beth Gamse, Judith Alamprese, Tracy Rimdzius, and Fumiyo Tao, *National Evaluation of the Even Start Family Literacy Program: Evidence from the Past and a Look Towards the Future* (Washington, DC: U.S. Department of Education, Office of the Under Secretary, 1998), http://www.ed.gov/pubs/EvenStart/index.html (accessed April 3, 2006).

[12]St.Pierre et al., 1995, 32.

random assignment, and it appears that significant problems occurred even in those.[13] Moreover, a number of the assignment procedures could have created bias. Consider the following:

- In Phoenix, Arizona, twenty-five families were randomly assigned to each group, but five program families dropped out or moved. The five program families were simply dropped from the analysis, undermining the random assignment.

- In Golden, Colorado, random assignment was carried out in two waves. Initially, ten families were assigned to each group. Then four were assigned to each group, but then three were simply added to the program group. This last addition was not randomly assigned, undermining the comparability of the two groups.

- In Birmingham, Alabama, twenty families were randomly assigned to each group, but then twelve program families dropped out or moved. Twelve new families were added to the program (as was one cross-over), for the total of twenty-one program families. The Abt team cautions that, "There is a control group, but it is not statistically equivalent to the program group."[14] Nevertheless, the cases were included in the overall sample and no adjustments or caveats were added to the evaluation to deal with this nonequivalence.

- In Albuquerque, New Mexico, fourteen families were initially assigned to Even Start and thirteen to the control group. A number of the families dropped out because they were not satisfied with their assignment and there was a second stage in which families were randomly assigned (but it is not clear how). In December 1991, there were thirteen program families and seventeen controls. Afterwards, four families were added to the program group to increase sample size. Although the exact procedures were not clearly described, it is apparent that random assignment was subverted in this site as well.

- In Reading, Pennsylvania, twenty families were assigned to the program group and nineteen to the control group. Within the first week, fourteen program families dropped out and were replaced with five families from a randomly assigned waiting list. By the end of the assignment process, the program group included eight nonrandomly selected families.

Each of the five sites experienced problems with random assignment, suggesting that the evaluation should be considered a quasi-experimental research design. Even so, the absence of meaningful statistical controls renders any findings highly suspect.

---

[13]Robert St.Pierre, Janet Swartz, Stephen Murray, Beth Langhorst, and Phil Nickel, *National Evaluation of the Even Start Family Literacy Program* (Washington, DC: U.S. Department of Education, 1993).

[14]St.Pierre et al., 1993, 30.

Assessing statistical controls in experimental and nonexperimental evaluations. The Abt team utilized random assignment, but because of problems in the randomization process described above, selection bias may have been a serious problem. They based estimates of program impacts on a regression model that included the pre-test value of the outcome measure in question as well as assignment to program or control group.[15] There were a number of differences in baseline characteristics, including: parent has high school diploma or GED (27 percent vs. 19 percent); spouse not working (27 percent vs. 14 percent); government assistance (53 percent vs. 44 percent); and English as the primary language of the parents (63 percent vs. 54 percent). None, however, were statistically significant, and the differences were "most likely an artifact of random assignment with a relatively small sample."[16] Since there were some problems with the randomization procedures, adding some of these variables as covariates might have improved the precision of the estimates and reduced potential bias.

In addition to possible selection bias within the research sample, University of Michigan professor Maris Vinovksis describes in an assessment of the Department of Education's evaluative practices how the selection of sites for the random assignment portion of the Even Start evaluation may have been geared toward showing positive results:

> The stringent, but understandable, criteria for selecting the 10 sites limited the representativeness of the IDS sites relative to the other Even Start projects. For example, only 32 of the 73 first-year grantees expressed interest in participating in the IDS analysis. Moreover, Abt decided that "projects for inclusion should be high-quality projects that can reasonably be expected to be successful at achieving their goals"—thereby potentially biasing the assessment toward a more positive outcome.[17]

Sample size. The Abt team pooled the data from the five projects, but even then, the sample remained small and large impacts would have been needed to produce statistically significant findings. Thus, the program may have affected some child outcomes, but impacts may have been too small to be detected, given the small sample size. A small sample also means that differences in baseline characteristics would have to be very large to be statistically significant, making it more difficult to assess the comparability of the program and control groups.

Attrition. Attrition varied depending on the outcome analyzed, averaging about 20 to 30

[15]There were a number of significant differences between Even Start and comparison group families at baseline, so some baseline covariates were included in the model, but the small sample limited the number of covariates possible.

[16]St.Pierre et al., 1995, 193.

[17]Maris A. Vinovskis, "Missing in Practice? Systematic Development and Rigorous Program Evaluation at the U.S. Department of Education,"(paper, Conference on Evaluation of Educational Policies, American Academy of Arts and Sciences, Cambridge, MA., May 13-14, 1999, revised January 2001), 20.

percent for most outcomes. Attrition was somewhat higher for control group families. No information was provided on whether there were statistically significant differences between nonrespondents and respondents or between program and control group members in the follow-up sample.

Data collection. The data collection relied on a wide range of tests and survey results. The data sources were appropriate for the questions being studied.

Measurement issues. The Abt team used well-known tests to measure child outcomes and assessed the validity and reliability of those measures. Some children, however, participated in Even Start for up to several weeks before taking various "pre-tests." If this time lag tended to boost scores on the pre-test, it could have understated program impacts, since most impacts are measured by comparing the change for each group between the pre-test and the post-test.

Generalizability. The five sites with random assignment were not representative of Even Start projects generally. The research sample differed from Even Start families nationally in a number of respects. For example, among families in the evaluation, 36 percent of program families and 45 percent of control group families spoke Spanish as their primary language, compared with 22 percent of Even Start families nationally. In addition, the particular approaches of the evaluated projects varied considerably. Some were home-based, while others were center-based. Finally, the evaluation focused on children three or four years old. Actual eligibility in the Even Start program is from birth to age seven, limiting the generalizability of the findings.

Vinovskis offers the following assessment:

> Given the unrepresentativeness of the participating IDS Even Start projects, the considerable heterogeneity among them, the nonrandomized nature of some of the controls, and the small sample sizes of the ones with controls, we should not have expected much definitive information about the impact of the overall Even Start Program on participants. Even looking at the five remaining IDS projects together cannot provide much reliable data, and looking at differences among individual models is almost impossible.[18]

Replication. Although the five sites represent a form of replication, the samples were too small to permit site-specific impacts. A second random assignment Even Start evaluation was conducted by Abt Associates Inc. for the years 1997/1998 to 2000/2001, but with a larger sample and stricter randomization procedures (see chapter 9).

[18]Maris A. Vinovksis, *History and Educational Policymaking* (New Haven, CT: Yale University Press, 1999), 129–130.

Evaluator's description of findings. The Abt team summarized the Even Start findings as follows:

> On each of the measures, there were no statistically significant differences between children in Even Start and the control group at the second post-test using data from the "In-Depth Study." While these results are discouraging, they need to be considered in light of two findings that work against two program effects: (1) the majority of children in the In-Depth Study control group were enrolled in some type of early childhood program by the second post-test; and (2) nearly half of the Even Start children in the In-Depth Study were no longer enrolled in an Even Start early childhood program by the second post-test. The impact of this involvement by the control group in early childhood education is seen most dramatically on the PSI [Preschool Inventory], where significant program effects reported at the first post-test erode by the second post-test. The results suggest that, when control group children enroll in an early childhood program, they learn many of the same skills the Even Start children learned in preschool. What is not known is the long-term effects of this lag time in learning. Since most of the control group children are in kindergarten by the second post-test, it appears that they are learning many of the basic readiness skills (e.g., shapes, colors, directionality) during their first year of public school. If many of the Even Start children learned these skills in a preschool program, they may be progressing to other school skills (e.g., prereading) during their kindergarten year. Thus, the finding that the differences between Even Start and control children on basic readiness skills are diminished when control children reach kindergarten does not preclude other long-term impacts of Even Start on school performance.[19]

This seems to be a fair appraisal based on the evaluation's findings, but the methodological weaknesses of the study call for more tentativeness in the negative conclusions.

Evaluator's independence. Even Start was evaluated by an independent evaluator, Abt Associates, Inc. Abt Associates was, however, under contract with the U.S. Department of Education.

Statistical significance/confidence intervals. Statistical significance was measured and reported at the 1 percent and 5 percent levels.

Effect sizes. Effect sizes were calculated by diving the difference in mean scores between the Even Start group and the control group by the standard deviation (of the measure of interest) and were reported in standard deviation (SD) units.

Most effect sizes fell in the range of 0.02 standard deviations (SD) to 0.15 SD. These

[19]St.Pierre et al., 1995, 182.

effect sizes, however, were not based on statistically significant impacts, so they were not discussed in detail. (See Appendix 1 for a further discussion of effect sizes and their interpretation.)

Sustained effects. The evaluation did not examine post-intervention impacts.

Benefit-cost analysis. Apparently not performed.

Cost-effectiveness analysis. Apparently not performed.

Commentary

Robert G. St.Pierre*

The evaluation reviewed by Besharov and his colleagues was the first national study of the Even Start Family Literacy program. Their review provides an accurate description of the strengths and weaknesses of the "In-Depth Study" (IDS), a portion of the national evaluation that was designed to assess the impact of Even Start projects. But the IDS was only a small part of the entire evaluation. This commentary provides an overview of the series of three national Even Start evaluations, and suggests that the design of the current (third) national evaluation will remedy some of the shortcomings of the first evaluation and therefore will offer a better assessment of program effects than the study that Besharov and his colleagues reviewed.

The national Even Start evaluation has been an ongoing effort from the time that the program was first funded in 1989 through the present. Every four or five years the Department of Education has held a competition to conduct the evaluation and to improve on its design. Each time the national evaluation has been re-designed, it has included a national survey component to describe projects and participants, as well as a smaller scale attempt to understand program effectiveness.

First National Evaluation (1989-1990 through 1992-1993)

The first national evaluation (1995) by St.Pierre and his colleagues was broad in scope, addressing questions such as: What are the characteristics of Even Start participants?[1] How are Even Start projects implemented and what services do they provide? What Even Start services are received by participating families? and What are the effects of Even Start on participating families? These questions were addressed through the National Evaluation Information System (NEIS) which was used to collect data on participant characteristics, project implementation, and participant outcomes from all projects using paper and pencil or optically scannable forms. Literacy assessments were administered at program entry and exit to one adult and one child in each Even Start family. In addition, the evaluation included an experimental component (the IDS) in which families in five volunteer sites were assigned to be in Even Start or a control group

*Robert G. St.Pierre was vice president and principal associate at Abt Associates Inc.

[1]Robert St.Pierre, Janet P. Swartz, Beth Gamse, Stephen Murray, Dennis Deck, and Phil Nickel, *National Evaluation of the Even Start Family Literacy Program: Final Report* (Cambridge, MA.: Abt Associates Inc. under contract to the U.S. Department of Education, Planning and Evaluation Service, January 1995).

and were measured three times during an eighteen-month period.

The first national evaluation described how early Even Start projects implemented the program. Documentation of program implementation helped the Department and grantees agree on the definition of key program terms by answering questions such as: "What counts as adult education in Even Start?" and "Who counts as a program participant?" The Department used data from the first evaluation to identify areas where Even Start projects needed technical assistance, in particular, improving the literacy focus and intensity of their parenting education components, engaging adults in adult education, and recruiting and retaining families.

While national data from the NEIS showed consistent child and adult gains on literacy outcomes, the lack of a national level control group meant there was no way to know whether the observed gains resulted from participation in Even Start. The five-site IDS provided experimental evidence about the effectiveness of Even Start, but suffered from a small sample size, problems with random assignment, and a lack of information on control group services.

Information from the first national evaluation was used to improve the program through legislative changes. Findings that showed a fairly low year-to-year retention rate were used to modify the legislation to require year-round services so that families would remain involved in Even Start throughout the summer and into the next year. The study served as a catalyst for other substantive changes made to the legislation including: focusing program targeting on those most in need, requiring that projects serve at least a three-year age range of children, allowing projects to serve teen parents, and allowing the involvement of ineligible family members in appropriate family literacy activities. The finding that there was a relationship between the amount of participation in Even Start and child-adult test gains provided evidence that resulted in an amendment in 1996 requiring Even Start services to be intensive.

Second National Evaluation (1993-1994 through 1996-1997)

In the second evaluation (by Fumiyo Tao and colleagues from Fu Associates in 1993), the national survey was improved, converted to a computer-based system, and renamed the Even Start Information System (ESIS).[2] While program and participation information continued to be collected from all Even Start projects, the administration of literacy assessments was restricted to children and adults from a 10 percent sample of projects (approximately 60 out of 600), called the Sample Study. No control or comparison group was included in the second evaluation.

Implementation data collected through the ESIS allowed tracking of changes in the population served over eight years, and the Sample Study provided national data on the size of

[2]Fumiyo Tao, Beth Gamse, and Hope Tarr, *Second National Evaluation of the Even Start Family Literacy Program: Final Report* (Washington, DC: Fu Associates, Ltd. under contract to the U.S. Department of Education, Planning, and Evaluation Service, 1998).

gains made by Even Start participants. Early data from the second evaluation corroborated the positive relationship between service intensity and family outcomes found in the first national evaluation. However, without a control group, the Sample Study could not address lingering questions about the program's effect.

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/chapter09.html.