Even Start Revisited:
A Counter to the *Third National Even Start Evaluation Program Impacts and Implications for Improvement (2003)*

Drucie Weirauch

The Goodling Institute for Research in Family Literacy

College of Education

Penn State University

Contents

Abstract

The William F. Goodling Even Start Family Literacy Program has undergone several national evaluations to measure its effectiveness. A key finding in a recent evaluation, conducted by Abt Associates on behalf of the U.S. Department of Education, has led legislators to question the value of the Even Start program. The indicting key finding, "Even Start children and adults made gains on literacy assessments, but not more than adults and children in the control group, two-thirds of whom received no adult or early childhood education services" is addressed in this critique. This paper questions the validity of the Evaluation's argument regarding the efficacy of Even Start. It voices concerns about the experimental design, site selection, and that the Experimental Design Study (EDS) occurred before accountability measures were in place. Further, it provides evidence of Even Start's more recent success, culled from multiple statewide evaluations that indicate its effectiveness for families since Even Start's reauthorization with the *Learning Involves Families Together (LIFT)* Act of 2000 and the *No Child Left Behind Act* of 2001. Curiously, although the authors of the *Third National Even Start Evaluation* acknowledge several shortcomings of the study, these have not been widely recognized by policy makers. Instead, policy makers focus on the one key finding—that Even Start does not work. This paper brings these shortcomings to the forefront in an attempt to show that exemplary Even Start programs are efficacious and do, indeed, impact the knowledge, skills, and lives of adults and children.

The Third National Even Start Evaluation: Program Impacts and Implications for Improvement, 2003 (hereafter referred to as *The Evaluation*) was undertaken from 1997-2001 to measure the effectiveness of Even Start and to provide information on this federally funded program's implementation. The methodology included analysis of data collected through the Even Start Performance Information Reporting System (ESPIRS) from 1997-2001 and an Experimental Design Study (EDS) conducted with 18 programs from 1999-2001. One key finding, in particular, is an indictment of this federally funded program, and it has contributed to legislators questioning the efficacy of the model. This, in turn, has led to a reduction in funding for Even Start with the threat of being de-funded completely. That key finding is:

> While Even Start children and parents made gains on literacy assessments and other measures, children and parents in the 18 Even Start programs that participated in the EDS did not gain more than children and parents in the control group—about one-third of whom also received early childhood education or adult education services. (2003, p. 1)

While *The Evaluation* reported other key findings, they are beyond the confines of this report. However, they included, in brief:

- Even Start families are very disadvantaged, much more so than Head Start families in terms of education and income.
- Even Start children and adults score very low in literacy compared to national norms.

- Families do not take full advantage of the services offered by Even Start projects, participating in a small amount relative to their needs and the program's goals.
- The early childhood classrooms are of overall good quality, but have insufficient emphases on language acquisition and reasoning.
- The extent to which parents and children participated in literacy services is related to child outcomes.

Of these, the first two are evidence that Even Start is serving those most in need, the poorest families and the adults and children with very low literacy as measured by standardized instruments. Indeed, facts from the Office of Elementary and Secondary Education (2003) indicate that Even Start families are significantly poorer than Head Start families. In 1997, 41% of Even Start families had annual incomes under $6000, contrasted with only 13% of Head Start families. Even Start parents are far more educationally disadvantaged than Head Start parents, with only 15% of Even Start parents having a high school diploma or GED, contrasted with 72% for Head Start.

The third bullet, participation and retention, clearly indicates the barriers that Even Start families must overcome in order to participate regularly. These barriers are myriad, from dispositional (fear of schooling, lack of self-confidence), institutional (location of services or schedule), and situational (lack of child care, transportation) (Cross, 1982). Programs work endlessly to resolve the barriers they can address to increase participation of their families.

The fourth bullet about the quality of the early childhood classroom is actually rather positive. That language and literacy and reasoning were not emphasized may be due to the fact that, during the time of the EDS, these were not emphasized in the legislation.

The last finding has been corroborated by at least two studies, which will be discussed later in this paper.

This paper is written in response to the first key finding. It provides a critique of the evaluation, an analysis of how the design may have affected the data, and a counter to its findings based on more recent data extracted from statewide evaluations from ten states. The report is based on a thorough review of *The Evaluation*, information gleaned from evaluation meetings conducted by the Goodling Institute for Research in Family Literacy at Penn State University; an analysis of ten statewide evaluations, provided by attendees at three hosted evaluation meetings; and research studies by the Goodling Institute and the state of Colorado.

Background

In October 2003 and 2004 meetings were hosted by Penn State's Goodling Institute for Research in Family Literacy at the National Even Start Association (NESA) annual conference in San Diego. In February 2003, a similar meeting was held at the National Center for Family Literacy (NCFL) annual conference in Orlando. Invited to these meetings were state evaluators and independent local evaluators recommended by state directors of Even Start. The

intent of the meetings was to dialogue about the nature of evaluation (both statewide and Even Start independent local evaluations) and to share frameworks and guidelines, ties to performance standards, and measures and assessments for the four components. Further, there was some discussion about how to measure the quality of collaboration and integration of components, two aspects unique to Even Start. Discussion was framed by how evaluation is used to inform program improvement, professional development, and policy. The goals of the meetings were to create a report on the best practices for state and local evaluation, share chosen instruments, discuss recommendations for policy, and share outcomes and findings to determine whether the national evaluations have presented an accurate portrayal of Even Start programs as implemented on the state and local levels. In the meetings, it became clear that individual states are finding ample evidence of the efficacy of Even Start as an intervention to improve the literacy of children and their parents as well as other important dimensions in their lives. The evaluators voiced grave concern about *The Evaluation* and its potential impact on continuation of the Even Start program.

This report addresses the first key finding from *The 2003 National Even Start Evaluation: Program Impacts and Implications for Improvement* and raises concerns about the validity of the argument of *The Evaluation* regarding the efficacy of Even Start. It provides evidence of Even Start's more recent success, using data extracted from statewide evaluations provided by eleven states (Nebraska, Massachusetts, Oregon, California, North Carolina, Connecticut, Colorado, New York, Kentucky, Texas, and Pennsylvania) and information

provided by attendees of the meetings held at the two conferences. Evaluators were present from the states listed above as well as from Texas, Florida, Georgia, New Mexico, Louisiana, Maine, Michigan, and Iowa. Statewide evaluations were not available from these states, however.

Critique of the Design

The Even Start Program overall has been placed under scrutiny and criticized because of results from *The Evaluation.* To the credit of those conducting the study, several shortcomings of the study were mentioned in the full report. However, it is the Executive Summary that is most often referred to by policy makers, especially the first key finding. That condensed report cannot provide the caveats and concerns that the full report modestly includes. The negativity of the key finding prompted the need to address *The Evaluation* regarding its data base, timing in terms of more recent legislation calling for accountability, site selection, and the experimental design overall. The shortcomings, considered more deeply in this paper, attempt to demonstrate that the first key finding of *The Evaluation*, that adults and children in Even Start made gains but not more than adults and children in the control group, misrepresents Even Start as an ineffective program, especially now that the program has had time to mature and develop, three years after the study data were collected.

Even Start Performance Information Reporting System (ESPIRS)

The Evaluation was partly based on the Even Start Performance Information Reporting System (ESPIRS) data 1997-1998 through 2000-2001. Indeed, chapters 2-5 of *The Evaluation* are based on the ESPIRS data, which reported on program and family characteristics, participation rates, and family progress. According to evaluators from the 18 states that attended evaluation meetings hosted by the Goodling Institute for Research in Family Literacy at the 2003 and 2004 NESA and the 2004 NCFL conferences, these data were suspect at best. More than one participant regarded the ESPIRS data as "flawed." In Pennsylvania, for example, programs did not consistently complete the ESPIRS forms; thus many were missing information and had to be discarded. During the time of the study, the method of entering ESPIRS data was changed, and programs had difficulty with the new technology of using a database; data, thus, were lost or inaccurately reported. Because each program reported the data directly to the federal government, there was no system in place to check for accuracy. Programs received late feedback and did not see the value of providing accurate and complete data. Programs were not required to report outcome data for parents and children. Because programs did not receive feedback for over a year, if at all, it could not be linked to program improvement. The accuracy of the ESPIRS data used for the national evaluation was, therefore, not reliable and findings based on those data are questionable.

Experimental Design Study (EDS)

The complement to the ESPIRS study was an Experimental Design Study (EDS) to study the effectiveness of Even Start. Its results and the basis for the first Key Finding appear in only Chapter 6. The EDS used a random assignment design, "the strongest approach for estimating the impacts of a program" (U.S. Department of Education, 2003, p. 154) and included 18 Even Start projects selected from the national 1,200 Even Start projects. The study consisted of two cohorts, which were studied over an 18-month period, as indicted below:

Cohort 1	1999-2000 (Fall '99 pretest and Spring '00 posttest)
(11 projects)	2000-2001 (No follow up in Spring '01)
Cohort 2	2000-2001 (Fall '00 pretest and Spring '01 posttest)
(7 projects)	2001-2002 (No follow-up in Spring '02)

Criteria for selection in the EDS included these minimal requirements:

- Minimally met Even Start's legislative requirements
- Had been in operation for at least two years
- Planned to operate through the length of the study
- Could serve at least 20 new families at the start of data collection
- Offered instructional service of moderate or high intensity
- Were willing to participate in a random assignment study

"However, no examination of the quality of instructional services was done as a part of the selection process" (p. 26), the evaluation concedes.

Of 115 eligible projects, only 18 volunteered to participate as the sample. The Follow-Up Findings from the Experimental Design Study (Ricciuti, St. Pierre, Lee, Parsad, Rimdzius, 2004) voices concern that 97 eligible projects refused to participate and admits that this fact, "does make us worry about the

generalizability of the findings" (p. 11). The report posits that the main deterrent to participation was the random assignment of families to participate in Even Start or the control group. The ethics of random assignment is taken up later in this report.

Each of the 18 EDS recruited families as usual and provided the families' names to Abt Associates staff who randomly assigned families to either participate in Even Start or to be in a control group. Two thirds (309 families) of the families received Even Start services and one third, the control group (154 families), were told they could not participate in Even Start for one year. Tom Sticht (2005) pointed out that a perusal of *The Evaluation* reveals that there are meaningfully significant, as differentiated from statistically significant, similarities and differences among the experimental and control groups which make statements about outcomes difficult to interpret or accept.

The main text of *The Evaluation* (not in the Executive Summary) states that, though there were some minimal requirements for the EDS projects, projects volunteered for the study instead of being randomly selected; thus results cannot generalize to the Even Start population on a strict statistical basis. Self- selection affects results. Interestingly, in addition to the small sample of only 18 programs, not all of the families were included in the data as "some families could not be found at the time of pre-testing and post-testing, some children accepted into the study were too young (under 2.5 years of age) to be pre-tested, and some parents/children were assessed but had missing data on selected items" (p. 155). This further limited the sample size.

Site Selection Inequities

Site selection problems of *The Evaluation* concern the demographics of the EDS as compared to national Even Start demographics. Although the plan for the study was to include both urban and rural populations and obtain a balance between high and low percentages of ESL families, the report acknowledges that this did not happen, “Due to the voluntary nature of the study, this plan could not be implemented perfectly” (p. 154). Indeed, while the EDS projects represent major kinds of projects funded by Even Start, the EDS families are more likely than the population to be Hispanic and urban; thus the sites in the EDS were not representative of national Even Start program demographics. While mentioned in *The Evaluation* Executive Summary, “Care should be given in applying the findings to Even Start projects as a whole” (2003, p. 9-10), the implications were not discussed in the report and will be considered here.

The discrepancy between the Even Start universe and the EDS is indicated in the table below.

	Experimental Design Study	Even Start Programs
Hispanic	75%	46%
Urban	83%	55%

With an over-representation of Hispanic and urban programs, the generalizability of the results to all of Even Start is questionable, as disclosed on page 154 in the report, “These data suggest that findings from the EDS are most relevant to urban projects that serve large numbers of Hispanic/ESL families.” Clearly, the majority of Even Start projects do not represent these demographics.

Though the researchers caution about the generalizability, policy makers have generalized the results from this study to the entire Even Start program.

Another site selection concern is about basic functioning. Of the 18 sites in the EDS, four (23%) were from Texas. Even Start in Texas, in 2003, was wrested away from the state education agency because of the poor overall quality of the state's program and its projects. The entire Division of Adult and Community Education, under which Even Start was managed, was eliminated. Since that time, Even Start is now under new leadership with Texas LEARNS, directed by the former director of an exemplary local adult education program in the state, and the Even Start coordinator hails from an excellent local program. While the leadership now is strong, in 1999-2002, when the study took place, Even Start in Texas was weak. That 23% of the 18 EDS projects were from Texas may well affect the data and the findings of the report.

Moreover, six of the 18 sites (33%) had been operating for just two years, starting in 1997 or 1998. This also is not representative of the Even Start universe, where 13% of Even Start programs had been in operation for only two years (U.S. Department of Education, 2001). Thus, the EDS had nearly three times the number of new programs than existed in the Even Start universe.

Even Start legislation at the time of the EDS required that at minimum, a successful Even Start project should:

> be implemented through cooperative projects that build on high quality existing community resources to create a new range of services (p. 4), and
>
> provide intensive family literacy services that involve parents and children, from birth through age seven, in a cooperative effort to help parents

> become full partners in the education of their children and to assist children in reaching their full potential as learners (p. 7), and
>
> (U.S. Department of Education, 2001).

While Even Start provides a start-up period of six months to new programs, typically programs require several years to find high quality partners who will collaborate with the program and find ways to effectively implement the complex model of integrating four components (adult education, parenting, and early childhood education and parent-child interactive literacy) to meet the legislative requirements. Further, it takes time to identify, hire, train and retain appropriate staff and to recruit families most-in-need. That 33% of the EDS programs were new and not fully developed to function effectively surely must skew the data, as at the time of the EDS, only 13% of Even Start programs had been in operation for only two years. This suggests that the results inaccurately paint a negative picture of the effectiveness of Even Start as a national program.

Clearly, the EDS did not include a representative sample of Even Start programs. Programs should have been randomly selected, but instead, due to problems encountered in the study, they self-selected to participate. As a result, the findings from the study cannot and should not be applied to the Even Start Program in general. A sample should be a small subset of the population and represent the population for a fair statistical analysis. It could be argued that the EDS created a sampling error in the site selection, due to the problems the research study incurred.

LIFT Act's Accountability (2002)

While the EDS comprehensively collected data during the study, albeit from a flawed data set, it is unfair to judge current Even Start programs' effectiveness based on data that are more than five years old. *The Evaluation* presents an unrealistically negative image of Even Start, which needs an historiocentric focus for several reasons.

The EDS study, conducted from 1999-2001, preceded the development and implementation of Performance Indicators (Standards), mandated in 2001 with the *LIFT* Act. Even Start programs, at the time of the study, were not held to the same accountability as they presently are with Even Start's reauthorization in 2001 and the *LIFT* Act's requirement of individually-established state Performance Indicators. These indicators, for the most part, were not in place and implemented until 2002 (after the EDS was completed), as participants at the two evaluation meetings testified. *The Evaluation* acknowledged that,

> During the period of this study, Even Start's guiding legislation stressed process factors such as collaboration with local service agencies and the recruitment and screening of eligible families, although it did require high-quality, intensive instructional components. The legislation was reauthorized in 2000 and 2001, and while all the previous requirements have been retained, the legislation now stressed more strongly the importance of the quality of instructional content. (p. 1)

While this brief acknowledgement appeared in the Executive Summary, greater attention to this caveat would document that Even Start programs during the time of the EDS (1999-2001) were different from current ones, which have improved to comply with more recent legislation and emphasis on accountability. Indeed, since the reauthorization, programs have been legislated to provide instruction

based on “scientifically-based reading research,” which is more tightly defined than the earlier legislative language of “high-quality instruction.” Further, staff qualifications have become much more restrictive, requiring the majority of existing teachers and all new teachers in adult and early childhood education classes to have an associate, bachelor or graduate degree in the appropriate field. As a result, program outcomes have improved since the implementation of performance indicators. *The Evaluation* reports that, “While the EDS sites represent functioning Even Start projects, they were not selected to be models of excellence” (2003, p. 9). Indeed, though Even Start programs were mandated to provide high-quality instructional services of sufficient intensity, the legislation at the time of the study did not specify what was meant by quality or intensity.

Time in the Program

The Evaluation report concludes that families are not staying long enough to meet goals and make significant gains. However, the study itself provides data for only nine months—fall to spring, with no follow-up of cohort one or two in their second year. Further, although year-round operation was a criterion for selection, the study found that the programs provided only seven months of instruction—too little for significant change.

The study occurred before Even Start legislation mandated year-round programming; a review of Statewide Evaluations from ten states (representing over 200 programs) indicates that programs offer year-round programming as now mandated by legislation. Thus, while the evaluation criticizes Even Start for

duration, programs now provide for duration and intensity as discussed more fully below.

Fidelity of the Model

While there exists the Keenan Model for family literacy, established in 1989 by the National Center for Family Literacy (NCFL), with its four components (adult education, parenting education, early childhood education, parent and child interactive literacy), implementation varies from site to site. *The Evaluation* made quite clear the differences of the EDS sites. Thus, even the concept of an evaluation of effectiveness of the Even Start model is questionable.

The variability of the EDS projects shows that a specific model was not being adhered to and there were wide variations in the settings. Indeed, integration of services, the core of the Even Start model and the focus of the study, was not a criterion for site selection. As the purpose of *The Evaluation* was to evaluate the effectiveness of the Even Start model as opposed to services families obtain for themselves, selecting projects that effectively implement the model with fidelity would have provided greater evidence of the quality Even Start. As Kirk suggests (2002),

> In order to conduct research on the effectiveness of a program, be it family preservation or any other program, a precise understanding of all of the program operations is necessary because the program operations comprise the 'independent variable' in the research study or program evaluation using an experimental or quasi-experimental design. In order to associate program outcomes with a program, one must have confidence that workers are following the prescribed service model closely, delivering the service with the intended intervention type, length of treatment, and 'dosage levels' to the proper service recipients (p. 5).

That programs were not selected based on “the quality of instructional services” (p. 26) and the fact that there were excessive number of new programs in the study brings up the concern about treatment fidelity (Kirk, 2004). Andy Hayes, (2001) also addressed the fact that national studies, when examining the efficacy of the model, do not study programs that are high-intensity, integrated, four-component programs, such as advocated by the NCFL and Congress, authenticated by the *Guide to Quality Even Start Family Literacy Programs* (RMC Research Corp. 2001) and validated by the National Research Council (NRC).

The importance of adhering to a particular model in a controlled study cannot be overemphasized, according to Kirk (2004):

> If participating programs do not adhere to the model, or if there is variation among settings that claim to use the same model, then the independent variable becomes amorphous and its relationship to the dependent variable becomes, at best weak, and at worst meaningless or misleading” (Kirk, 2004, p. 2).

A report to the U.S. Department of Education (American Evaluation Association, 2003) agrees that in order to test whether an educational program is effective, it must be tested by researching a “specific set of education practices or interventions that are thought to have an impact on a given set of educational outcomes” (p. 1), further testimony to the need for fidelity. The authors of the study acknowledge that the EDS programs were not selected because they represented exemplary models of Even Start. That a number of the EDS projects were not implementing a specific set of practices effectively is a flaw of a study that purportedly examined the effectiveness of a model. Even Start programs deserve the support of evaluators and researchers to test the efficacy of their

programs but only when administrators and practitioners are willing to adhere to the program model as intended.

The Evaluation impact study report acknowledges that one of the possible causes of lack of significant change in adults and children, as measured by pre- and post-testing of both the Even Start participants and the control group participants, is that Even Start programs are legislated to partner with local agencies (most of which provide educational services) to ensure that services are not duplicated. The variability of the sites of the EDS is evident in the numbers provided in the report. In the EDS study, only 12 of the 18 projects (62%) provided the Early Childhood Education component; two (11%) were shared by Even Start and a partner. The national Even Start average for programs providing early childhood education is 90%. Only five (25%) of the adult education classes were provided by Even Start and three (17%) were shared by partners. There were ten (55%) adult education courses provided by partners alone (p. 106). In the Even Start universe, 50-60% of Even Start programs (only) provide adult education. Again, the EDS sites were not representative of Even Start programs, using far more than the national average of collaborators to provide early childhood and adult education. While partners may well provide quality educational services, accountability is not necessarily shared by partners, and data may not be of the best quality, thus affecting the outcomes of the study.

The Evaluation admits that “Given Even Start’s intuitive appeal as an approach for enhancing parent and child literacy, we interpret the lack of

effectiveness as an indication that the Even Start approach needs to be strengthened" (p. 10). This paper argues that had site selection been more representative of true Even Start programs -- that is, fewer new programs, fewer from a state under scrutiny, and representative of the national rural and non-Hispanic populations– the study may have found that the Even Start approach is strong and effective when implemented as intended. Those families in the study who received Even Start services were not necessarily receiving the quality of services that exist in average Even Start program, due simply to the fact that the study had an overrepresentation of new programs and/or programs that did not implement the Even Start model with fidelity.

Further, if the EDS had occurred after the implementation of legislated performance standards, tighter staff requirements, and scientifically-based reading research in 2002, outcomes for children and adults who participated in Even Start likely would be better than those in the control group. While an experimental design in which families eligible for Even Start are randomly assigned to participate in the program or in a control may be a strong approach to estimate the effectiveness of Even Start, it is unlikely that local programs or states could undertake this approach due to the fact that it is a costly, time-consuming enterprise that requires considerable expertise. It would be prudent to undertake another national evaluation of Even Start using a different data base and a quasi- or experimental design now that Even Start has entered a new stage of accountability. Better, Even Start would benefit from small, targeted research studies that attempt to determine best practices, which will help to

describe an optimal service model. Here, random assignment would be far more appropriate than it was for a national evaluation.

Ethical Concerns of the Experimental Design

Beyond concerns regarding site selection and the timing of the EDS, exists an ethical concern with the study's experimental design and random assignment of families. Many, if not most, individuals who work with at-risk families in direct practice find the concept of random assignment to be ethically problematic. A critique by Kirk (2004) of the evaluation of the Family Preservation and Reunification Programs indicates that at some sites, staff had major concerns about random assignment and often would subvert the randomness of assignment, using a "triage" approach, sending the most-at-need families to the intervention and less needy families to the control group. Another concern was that the general random assignment strategy, in a practice setting, has a negative impact on the environment and can introduce measurement error in an experimental study. Thus, even if the EDS researchers truly were those who provided the random assignment of families, the very nature of a random design study changes the natural environment of the program.

The American Evaluation Association (AEA, 2003) cautions that randomized control group trials (RCTs), such as the EDS, are not the *only* studies capable of generating understanding of causality; they can, indeed, be misleading. The AEA suggests that a limited number of "isolated" factors are neither truly limited nor isolated in natural settings and are less capable of discovering causality than

designs sensitive to local culture. Sound policy decisions benefit from data illustrating not only causality but also conditionality. The AEA report agrees with Kirk that “denying control group subjects access to important instructional opportunities in critical medical intervention is not ethically acceptable even when the RCT results might be enlightening.

Much debate and discussion has recently been spurred by the U.S. Department of Education’s priority for evaluating educational programs using RCT methods as the only means to determine causality. The use of experimental and control groups with randomized assignment is seen as the “gold standard” of education research (Maxwell, 2004). Evaluators for the past decade have argued about the rigor of newer inquiry methods. Actual practice and published examples demonstrate that alternate methods and a mix of methods are both rigorous and scientific. It is unethical to dismiss such methodology as ineffective and promote only the use of RCTs (AEA , 2003).

Indeed, the social sciences present a challenge in that the use of controlled experiments, and the large number of relevant variables, provide an obstacle to efficient verification of the efficacy of a model. Even Start’s complexity and variations make it a very difficult program to evaluate for efficacy.

In conclusion, the findings provided by *The Evaluation* as to the efficacy of Even Start were based on flawed site selection that was not random as intended, an overrepresentation of urban and Hispanic population that makes generalizability difficult, and timing before Even Start’s age of accountability. There are ethical concerns about the design of the study due to the complexity of

the program, which makes fidelity of model difficult, leading to questionable data obtained in the unnatural environment of a randomized design. Indeed, *The Evaluation* raises more questions than it answers. In the words of Tom Sticht (March 15, 2005), "One thing is for certain, to use a 'fools gold standard' study, to kill the chances for education for tens of thousands of children and adults is unconscionable. It tarnishes the image of the United States as a major force in the United Nations Literacy Decade—a decade in which it is proclaimed that 'Literacy is Freedom'."

A Counter of Key Points

As discussed in the previous section, a concern with *The Evaluation* was that it occurred before Even Start's implementation of the *LIFT* Act of 2001 ushered in a new phase of accountability with state-developed performance indicators, staffing requirements, and emphasis on scientifically-based reading research. Since then, Even Start programs have improved. It is important to look at results regarding gains for children and adults from a number of states in more recent evaluations, culled from statewide evaluations from eleven states from a total of more than 250 programs—far more than the 18 EDS projects in the *The Evaluation* impact study. As suggested earlier, it is nearly impossible for states or local programs to use an experimental model with a control and intervention group. Programs have neither the expertise, time, nor funding to do so. Thus, this synthesis is comprised mostly of data that report outcomes for children and adults. It reiterates findings from the *Synthesis of Local and State Even Start Evaluations* (St. Pierre, Ricciuti, Creps, 1999) that children and adults are making

statistically significant gains in language and literacy. That report was based primarily on local evaluations (118 local evaluations and four state evaluations). This paper considers data, culled from eleven recent state evaluations, and shares results on outcomes for children and adults. Results indicate that far more gains are being made than reported in *The Evaluation*. The statewide evaluations (2001-2003) report substantial gains in children's learning for pre-school and infant children.

Children's Gains – Recent Statewide Data

Pre-School, Infant, Toddler

While states did not implement an experimental (or quasi-experimental) design for their evaluations, it is important to share that Even Start children are making statistically significant gains, suggesting that the gains are due to the program intervention and not due to chance. Most states developed performance indicators, as required by the *LIFT* Act, in regards to reading on grade level, attendance, and promotion. Thus, many states address indicators in terms of school age children. Still, some states collect data for younger children and show that Even Start infant, toddlers, and pre-school children are making statistically significant developmental gains.

In Nebraska, children taking the Teacher Rating of Literacy and Language (TROLL) had statistically significantly greater gains ($p=.007$ and .043) in language and literacy, specifically in oral language and in reading skills. This is in

direct contrast to the *National Even Start Evaluation* (2003) findings that Even Start children are far behind the national norm.

In Pennsylvania, pre-school children taking the Work Sampling System (WSS), Early Learning Assessment Profile (Revised) ELAP-R, and Child Observation Record for Pre Schoolers (COR) made significantly statistical gains (p.001) in all domains.

In Colorado's 14 programs, in 2003-2004, 91% of the Even Start infants, 86% of toddlers, and 88% of pre-school children were at age appropriate levels of development, suggesting again that Even Start children, who are the most at risk for school failure, are not far behind the national norm. This is an increase from previous years' results.

In California, 67% of children entering kindergarten were rated as "fully mastering" or "almost mastering" reading readiness behaviors.

In Kentucky, 95% of Even Start children were on target for reading readiness, far exceeding the Performance Indicator benchmark of 75%.

In Texas, 71% of the children served exceeded the expected language development for their age group.

School Age Children -- Recent Statewide Data

More data were available for school-age children, even though this is the age group for which family literacy often has less control, as educational services are provided by the school district and not directly by family literacy. It is surmised that parents who participated in family literacy (based on data from the

states) are more involved with their children's education and children who have participated in Even Start are more ready to learn than those who do not participate. Thus, results from schools showcase the value of Even Start in the evidence from statewide evaluations.

School age children fare well for reading on grade level—exceeding what *The Evaluation's* (2003) findings -- that Even Start children are far below the national average. Reading on grade level is one of three required Performance Indicators.

In Nebraska, teachers reported that by the fourth quarter, Even Start children were at the satisfactory or better level in key academic areas. In North Carolina, children far exceeded the indicator that 50% would improve reading skills with 82.3% in Grade 3 and 77% in grades K-2 improving their scores. In 2001-2002, likewise in Pennsylvania, teachers reported that 55% of the children read on or above grade level. This refutes the national evaluation findings that Even Start children are far below the national average.

About three-quarters of Even Start elementary students in Massachusetts were at or above grade level in reading, in their attitude toward school, and in social skills; over half of the children were rated at or above grade level in problem-solving.

In California, 64% of Even Start children in kindergarten to second grade met grade level content standards in reading and math, and 70% of the English language learners made progress in English skills.

Attendance, another of the three mandated Performance Indicators for children, is also a highlight for Even Start children. In Nebraska, teachers reported that Even Start children were slightly above average in attendance. Tardiness decreased during the program year from 4.6 to1.1 days. Likewise in Connecticut 100% of school age children met the attendance standard. In North Carolina and New York, school age children exceeded the performance indicators for attendance.

Promotion to the next grade level is the third mandated Even Start Performance Standard, despite its controversy, stemming from "passing on" despite a child's ability, misinterpretation of promotion for special needs children who are not "promoted," as well as other philosophical and educational issues. Still, it is a required standard, and states set their own benchmarks. Children in Connecticut, Pennsylvania, Colorado, and North Carolina far exceeded their state's indicator benchmarks for promotion with 100% promoted in Connecticut, 95% in Pennsylvania, 94% in North Carolina, and 92% in Colorado.

Early Childhood Education -- Quasi-Experimental Designs

Pennsylvania

A study by the Goodling Institute for Research in Family Literacy at Penn State University (Askov, Grinder, Kassab, 2005) used a quasi-experimental design to test two research questions: 1) Does pre-school children's participation in the family literacy program lead to gains in developmental skills, particularly literacy-

related skills; and 2) Does parental participation in a particular component of family literacy affect child development scores?

The early childhood assessments included the following: for children birth to 3 years, the *Early Learning Accomplishment Profile (ELAP);* for children ages 3-5, the *High/Scope Child Observation Record (COR)* and the *Early Accomplishment Profile-Revised (LAP-R).* Each of these instruments measures essentially the same developmental skills.

Only children who were enrolled in family literacy programs were included in the analysis. No children were denied access to the services; thus the design did not compromise ethics with random assignment. Further, the study occurred in a natural setting. Variables included age of the child at assessment, whether the child had participated in an educational program prior to enrollment in family literacy, and if the child had special needs. Other controls included the number of hours the parent participated in adult education, parenting education, and interactive literacy. Two groups comprised the analysis. The intervention group included children who had a pretest and a post test after being in the family literacy program for at least 90 days. The post test score was compared to the pre test scores of a comparable age group of children just new to family literacy, controlling for the variables above.

Results indicate that children who had participated in family literacy for at least 90 days were significantly higher ($p<0.05$) than those of comparable age who had just started the program for all domains on the COR and most of those for LAP-R and ELAP.

The second question, regarding the impact of a parent's involvement in different components of family literacy, revealed that the intensity of participation for adults in adult education had a significant effect on most of the developmental skills for infants and toddlers as measured by the ELAP. Thus, early language intervention that is provided by Even Start is critical.

Colorado

Colorado's study (Anderson, 2003) was a follow-up from one family literacy program of 15 Even Start families who had been out of the program for an average of 3.5 years. One part of the study was a teacher report of educational achievements of school age children who had been enrolled in Even Start and a comparison child group randomly selected from the teacher's class list by an Even Start staff member. Comparison children were not matched on demographic or risk factors.

According to teacher reports, of the children who had participated in Even Start, 53% were reading above grade level contrasted with 29% of the comparison group. Interestingly, of Even Start children, 47% were reading at grade level and none below grade level. In contrast, of the control group, only 43% were at grade level and a full 28% were reading *below* grade level.

The study also considered other important educational domains such as speaking and listening, writing, overall academic performance, behavior, relations with other students, family support, and motivation to learn. In all of these, Even Start children outperformed the control group. Even Start children

had slightly poorer attendance than the control group. Curiously, teachers reported rated the control group children higher for self-confidence and probable success in school, which seems to contradict the rest of the findings. It would seem that children who have participated in Even Start have an excellent chance of succeeding in school and have demonstrated this fact. As they continue to excel, their confidence will rise, especially if their teachers also begin to believe in their capabilities.

Colorado also considered data from the Colorado Student Assessment Program (CSAP), an assessment not used until the third grade. CSAP reading scores were available for only about 40% of the children. While this number is insufficient to draw conclusions, it is worth nothing that Even Start children's reading scores in several areas were better than those of the control group. Of the six Even Start third graders, one was advanced, three proficient and two partially proficient. In the control group of six third graders, none was advanced, three were proficient, one was partially proficient, and two were unsatisfactory.

Adult Gains -- Recent Statewide Data

The Evaluation looked at outcomes for adults in 18 programs. It found that adults in the EDS made gains, but no greater than those in the control group. However, most states report that their adults are meeting or exceeding the performance indicator benchmarks they have set and that frequently the gains for adult learning are statistically significant.

Most states use standardized tests to measure literacy gains for their adults, mainly using the Test of Adult Basic Education (TABE), Basic English Skills Test (BEST), Comprehensive Adult Student Assessment System (CASAS), and actual and official practice tests of the General Educational Development (GED) credentials.

Adults enrolled in Pennsylvania's 70 family literacy programs in 2002-2003 met or exceeded performance indicators for all but two of the nine adult assessments, missing the TABE math by only four points and the CASAS employability math by one point. It is important to note that adults enrolled in Pennsylvania's family literacy programs are held to the same accountability standard as adults who are enrolled in much less complicated adult education programs (not having the three other components). Nearly half of the adults who set earning a GED as a goal did so, and 86% obtained a high school diploma. Of those adults who had a goal to go on to post-secondary education or training, 52% did so. Over half of the families reduced or eliminated dependence on TANF or other public assistance.

Likewise, New York's 2002 statewide evaluation of 70 programs (3155 adults) reports that adults exceeded all performance indicators as indicated in the table below.

Assessment	**NY %**	**PI %**
TABE—1 grade level gain	67.7%	50%
NYSPLACE (ESL)	76.9%	50%
GED	51%	50%
Post Secondary Education/Training	79.9%	50%
Employment	83.3%	50%

Colorado's adults (14 programs) also made significant achievements with 41% earning a GED and 63% entering higher education or training. A special population of teen parents resulted in 80% staying in school and 78% graduating from high school.

Oregon (eight programs) reports that at 61% of enrolled parents improved their literacy skills, gaining at least one level or completing some or all of the GED tests.

Adults in California exceeded the adult education Performance Indicators for English GED (61%), Spanish GED (64%), and high school diploma (61%).

Adults in Nebraska and North Carolina far exceeded all Performance Indicators for adults at all levels--English speakers as well as English language learners.

In Kentucky, 69% achieved a GED and 96% a high school diploma. A full 100% with the goal of entering post-secondary education or training did so.

Over 80% of Massachusetts Even Start parents made significant academic gains in communication, reading, and understanding children's learning and writing and two-thirds made strong gains in English language acquisition, math and problem-solving skills.

Other Outcomes -- Recent Statewide Data

While Even Start is an educational program for adults and children, its primary focus is on the "Parent as the child's first and most important teacher." Statewide data indicate that parents take this concept seriously. Further, Even Start is intended to

help families become better community members. The data from statewide evaluations provide evidence of the following:

- Parents read more to their children, have more books at home, and take their children to the library more often than before participating.
- Parents are now more informed about children's development and age appropriate expectations.
- Parents are now more active in their children's classroom, volunteer more and talk more with teachers.
- Parents now take better care of their and their children's medical and dental health.
- Parents have registered to vote or voted for the first time.
- Many parents obtained a driver's license.
- Parents are now more active in their community.

Recent statewide and local evaluations provide evidence of the effectiveness of Even Start as an intervention, and therefore a national evaluation employing a different design is both timely and essential to further attest to its effectiveness. "Future evaluation work will be most helpful to Even Start if it is designed to find, demonstrate or test effective family literacy practices—to identify and determine which practices and procedures work best and hence can be used as a template, or model, for improving Even Start projects across the nation" (*Third National Even Start Evaluation*, 2003, p. 17).

References

American Evaluation Association (November, 2003). Response to the U.S. Department of Education "Scientifically-based Evaluation Methods." Retrieved November 11, 2004 from http.//www.eval.org/doestatement.htm.

Anderson, B. (Feb. 2003). Colorado Even Start Follow-Up Study. Colorado Department of Education: Center for At-Risk Education.

Askov, Grinder, Kassab. (2005). Impact of family literacy on children. *Family Literacy Forum, 4* (1), p. 28-39.

Cross, K. P. (1982). *Adults as learners*. San Francisco: Jossey-Bass.

Hayes, A. (2001). High-quality family literacy programs: Child outcomes and impacts. Retrieved on November 10, 2004 from www.famlit.org/policyand advocacy/policy and research briefs.

Hayes, A. (2001). High-quality family literacy programs: Adult outcomes and impacts. Retrieved on November 10, 2004 from www.famlit.org/policyand advocacy/policy and research briefs.

Kirk, R.S. A Critique of the Evaluation of Family Preservation and Reunification Programs: Interim Report. Retrieved on January 24, 2005 from www.nfpn.org/tools/articles/critique.php.

Kirk, R.S. , Reed-Ashcraft, K., Pecora, P.J. (2002). Implementing intensive family preservation services: A case of infidelity. In Family Preservation Journal, 6(1).

Maxwell, J.A. (2004). Causal explanation, qualitative research, and scientific inquiry in education. *Educational Researcher, 33*(2), pp 3-11.

National Center for Family Literacy (2004). National Center for Family Literacy responds to President Bush's budget request for Even Start. Press release. Retrieved on August 31, 2004 from http:/www.familit.org/AboutNCFL/CurrentPressReleases/Even-Start-2005.cfm.

RMC Research Corp., (2001, June). *Guide to Quality: Even Start Family Literacy Program, Volume I (revised)*. Manuscript submitted for publication.

Sticht, T. (March 2005). The Even Start Third National Evaluation Experimental Study: Is the Gold Standard Tarnished? AAACE. NLA list serv. March 15, 2005.

U.S. Department of Education, Planning and Evaluation Service, Elementary and Secondary Education Division, *Third National Even Start Evaluation: Program Impacts and Implications for Improvement* (2003). Washington, DC.

U.S. Department of Education, Office of Elementary and Secondary Education. *Even Start Facts & Figures.* Retrieved on November 19, 2004 from http://www.ed.gov/about/offices/list/oese/sasa/esfacts.html.