

BACKGROUND

The Third National Even Start Evaluation: Data that Distort Reality

INTRODUCTION

The 2003 Third National Even Start Evaluation, conducted by Abt Associates on behalf of the U.S. Department of Education, determined that Even Start children and parents did not gain more on outcome measures than a control group. This has been widely cited by the Administration as the primary reason to eliminate the program.

The following points respond to the results of the evaluation, critique the design, and address the timing, suggesting that the evaluation did not capture the reality and impact of Even Start particularly for limited English proficient populations.

EXPERIMENTAL DESIGN STUDY

The evaluation was conducted using an Experimental Design Study (EDS), the “gold standard,” which requires random assignment. There were inherent problems with this design.¹

SAMPLE SIZE: Out of 1,200 Even Start programs across the nation and 115 programs that were eligible to participate in the study (due to criteria set in the EDS), only 18 programs were included in the sample, which represents only 1.5% of the Even Start universe and 16% of the eligible programs. Sample size protocol suggests that 291 programs from the 1,200 existing total and 89 programs from the eligible 115 would provide an adequate sample size with a 5% sampling error.² Thus, the sample size was inadequate for the EDS.

DEMOGRAPHICS: The sample was not representative of the Even Start universe. The majority of the participants in the EDS were Hispanic (75%) and from urban communities (83%), while the Even Start universe comprises only 46% Hispanic and 55% urban programs. Thus, it is not possible to determine from the evaluation the extent of Even Start’s impact with other populations, including non-Spanish-speaking English language learners (ELL), native English speakers, migrant and Native American populations, and families in rural communities.

GEOGRAPHIC DISTRIBUTION: The 18 EDS sites were not geographically distributed. In fact, four programs (23%) were from Texas. In addition, at the time of the study, Even Start in Texas was weak. In 2003, Even Start and the entire Division of Adult and Community Education were closed down. (It is important to note that Texas currently has an admirable Even Start program serving Hispanic/Latinos.) That 23% of the programs were from a state with overall low quality at the time of the study exerts a negative influence on the data and findings. The lack of a geographic distribution weakens the finding's generalizability.

RANDOM ASSIGNMENT: Though families were randomly assigned to receive Even Start services, programs were not randomly assigned to participate in the EDS. Instead, programs who met the EDS criteria volunteered to participate. Self-selection can affect results. Indeed, 97 eligible programs declined to participate. Even the authors admit that this fact “does make us worry about the generalizability of the findings.”³ Thus, the results cannot generalize to the Even Start universe on a strict statistical basis.

KEY POINT: *The EDS, therefore, had problems with sample size and representative sample, and did not use random assignment as intended.*

LIFT ACT ACCOUNTABILITY

The EDS predated Even Start accountability requirements from the

Learning Involves Families Together (LIFT) Act of 2001.

PERFORMANCE INDICATORS: In 2001, states were required to develop Performance Indicators for Even Start programs to document outcomes for adults and children.

These reforms enhanced program quality, but were not implemented until 2002, well after data for the EDS were collected. These reforms worked, as illustrated in the Texas example above, where low performing programs were shut down.

Overall Program Quality: The LIFT Act also strengthened requirements for Even Start staff in terms of education level, use of scientifically based reading research in the early childhood and adult education classrooms, and guidelines for intensity and duration. These requirements also were not in place at the time of the EDS.

KEY POINT: *During the time of the study, Even Start programs were not held to the same accountability as they are today. Even Start, therefore, should not be judged on data (questionable data) that predates the accountability required today. In light of these accountability requirements, another national evaluation of Even Start is long overdue.*

HISPANIC/LATINO POPULATION

There are special issues and implications raised by the National Even Start Evaluation for the burgeoning Hispanic/Latino population.⁴

ASSESSMENT: Mostly English-language instruments were used to assess literacy growth for adults and children in the EDS, although most of the participants (75%) were Spanish speaking. A fair and comprehensive evaluation of the growth of Hispanic children's language development needs to include assessment of their Spanish-language development as well as their English acquisition. The evaluation does not provide information about the level of Spanish fluency for the parents or children or the impact of Even Start on the family's use of home language, even though Parent-Child Interactive

Literacy Activities, whether in English or other native language, is one of the key components of Even Start.

OUTCOME MEASURES: The primary child outcome measures included the Peabody Picture Vocabulary Test (PPVT) and the Woodcock-Johnson-Revised test, which have been widely criticized by bilingual scholars as being invalid for dual language learners. These measures do not capture the linguistic strengths and communicative abilities of pre-school children who are acquiring English. Parent measures also have questionable validity for Hispanic populations. Administration of the assessment was biased to using the English versions rather than the Spanish versions. Therefore, Hispanic adults and children were unable to display their true linguistic and conceptual competence and growth.

When the measures are flawed and invalid for the population assessed, then it is impossible to know the impact of a program.

CONCLUSION

The findings from The Third National Even Start Evaluation (2003) were based on an inadequate sample size that was flawed in its representation of the Even Start universe. Implementation of the “gold standard” EDS was not random as intended. The evaluation occurred before the LIFT Act and its subsequent standards that held Even Start to higher accountability. Assessments of the overrepresented Hispanic/Latino population were inappropriate and did not capture the linguistic ability and literacy outcomes of these adults and children. Even the evaluators admit that the Third National Even Start Evaluation has questionable generalizability: “Care should be given in applying the findings to Even Start as a whole.”⁵

ENDNOTES

1. Weirauch, D., “*Even Start revisited: A counter to the Third National Even Start Evaluation: Program impacts and implications for improvement*,” Goodling Institute for Research in Family Literacy, University Park, PA: 2006. Available at: www.ed.psu.edu/goodlinginstitute/pdf/3rd_ESNE2003critique.pdf
2. Krejcie, R.V. & Morgan, D.W. “Determining Sample Size for Research Activities,” *Educational and Psychological Measurement*, 30. 1970. pp. 607-610.
3. *Third National Even Start Evaluation: Program Impacts and Implications for Improvement*. U.S. Department of Education, Planning and Evaluation Service. Elementary and Secondary Divisions. Washington, D.C.: 2003. p. 11.
4. Espinosa, L.M. (n.d.) *Third National Even Start Evaluation and Latinos: Do the Data Distort Reality?*
5. *Third National Even Start Evaluation: Program Impacts and Implications for Improvement*, pp. 9-104.