An Evaluability Assessment of the Toyota Families in Schools Program

Prepared for the National Center for Family Literacy

Janet Quint

with
Anne Sweeney

Manpower Demonstration
Research Corporation

May 2001

This report was funded in part by the National Center for Family Literacy.

Dissemination of MDRC publications is also supported by the following foundations that help finance MDRC's public policy outreach and expanding efforts to communicate the results and implications of our work to policymakers, practitioners, and others: the Ford, Ewing Marion Kauffman, Ambrose Monell, Alcoa, George Gund, Grable, Anheuser-Busch, New York Times Company, Heinz Family, and Union Carbide Foundations; and the Open Society Institute.

The findings and conclusions in this report do not necessarily represent the official positions or policies of the funders.

Contents

Tables and Figures

Acknowledgments

Without the cooperation and support of staff members at the National Center for Family Literacy, this paper would not have been possible. We wish in particular to acknowledge Jeff Tucker for his general assistance to the effort and Heather Hill for helping us to understand and use the data in NCFL's large data base.

Heather Schweder, a consultant to MDRC, conducted site visits to the Toyota Families in Schools programs. Her thoughtful reports helped us grasp the operating issues facing these sites. We also thank local site staff for their candid responses to her queries.

At MDRC, Howard Bloom was instrumental in assisting us in developing the sample sizes that would be required by an evaluation and was a useful sounding board as we clarified our thinking about potential evaluation methodologies. Robert Granger provided detailed comments on an earlier draft of this paper, as well as ongoing support.

The Authors

I. Introduction

The Toyota Families in Schools (TFS) program is a new family literacy initiative developed by the National Center for Family Literacy (NCFL), with support from the Toyota Motor Corporation. The program adopts and adapts a model that NCFL has used earlier in programs for preschoolers and their parents; in TFS, the participating children are between the ages of 5 and 12 and attend Title I schools serving largely low-income populations. TFS was first implemented during the 1998–1999 academic year in three elementary schools in each of five cities ("sites") across the country; during the 1999–2000 academic year, 15 schools in another five cities were added, and by the end of the three-year demonstration period, the program will be in place at 45 schools in 15 cities. During the first year, participation in TFS was voluntary in all except one site; there, those parents who received TANF were required to attend TFS (or another approved program) in order to satisfy the requirements of the local welfare-to-work program.

TFS seeks to influence a broad range of outcomes for children and parents. For children, these outcomes include scores on standardized tests, school attendance, positive behavior, and attitudes toward learning. For parents, they include improved academic skills, increased critical thinking and problem-solving strategies, improved employability skills, enhanced knowledge of child development, and improved behavior management skills.

NCFL initially approached MDRC to design and conduct a definitive evaluation of the effects (or "impacts") created by the TFS program. After some discussion, the two parties agreed that because the program was just getting under way, and because at the outset it operated on such a small scale, it would be inadvisable to measure effects at this juncture. Instead, MDRC proposed to prepare this evaluability assessment, which would discuss the conditions under which a rigorous impact study could be conducted and the implications that doing such a study would hold for program operations.

To prepare for the evaluability assessment, MDRC staff reviewed program documents and statistics on the characteristics of program enrollees and their participation in program activities. The principal author also attended a NCFL-sponsored conference for sites that were completing the first year of operations and sites just coming on board. In addition, a consultant to MDRC visited the three first-year sites where, in the view of NCFL staff members, TFS had been implemented most successfully and conducted interviews with local site coordinators, adult education coordinators, and principals of the participating schools.

The essential conclusion we have reached is that a random assignment experiment is the most feasible and methodologically rigorous way to evaluate TFS. Although we consider other outcomes as well, for purposes of illustration, much of the following discussion centers on the use of an experiment to measure how TFS affects children's scores on standardized reading tests. We suspect that literacy gains are the main criterion by which funders and others will judge the effectiveness of family literacy interventions, and we doubt that any such intervention will be considered a success unless it raises reading scores. Furthermore, the same general principles that apply to measuring program effects on reading scores would apply to any other outcome measures that might be selected.

An experiment could be mounted with a reasonable level of effort under the following conditions, discussed in detail later in the paper:

- The program would have to be operationally strong enough to produce effects of at least middling size (since small effects could not be measured with any degree of statistical reliability).
- Programs would have to operate in a more uniform manner than at present, so that data could be pooled across sites.
- Between three and seven schools would have to agree to participate in the evaluation.
- If three schools were involved in the study, then at least 15 children in each school would be selected to enroll in TFS and another 15 would be assigned to a control group and excluded from the program; if four schools were involved in the study, then each school would need to include ten TFS children and ten control group children; and if seven schools were involved, each school would need to include five TFS children and five control group children.
- To operate on this scale, there would need to be an adequate pool of interested families in the communities served by the participating schools, and the schools would have to recruit twice as many eligible and interested families as they could serve.
- The schools would have to select families to participate in TFS randomly.
- Pretest scores (e.g., achievement scores from a previous year, the more recent the better) would need to be available for children in the study.

Deviations from most of these conditions would require a substantially larger number of families involved at each school or a substantially larger number of schools.

In addition, an experiment conducted under these conditions would involve a group of children who are extremely diverse in terms of their age, the language they speak at home, their pre-program level of literacy, and many other variables. Given this degree of heterogeneity, the estimates of program effects resulting from any evaluation are likely to be imprecise. If program administrators were to opt to focus the study on more homogeneous groups of children (and their families) — such as children in certain grades — the precision of the estimates would be increased. Narrowing the scope of the evaluation would entail a trade-off, however, since the estimates of program impacts could not be generalized to the entire group of TFS participants.

We also want to point out several other considerations that NCFL will have to weigh carefully. First, we fully recognize that an evaluation of this nature would require major changes in the way that TFS has operated to date. Second, we want to emphasize that the number of schools and participants specified above represents the number needed to measure the effects of TFS *overall.* Often, we find that programs have especially strong effects on particular subgroups of the population (e.g., children whose mothers have not completed high school, children from families in which English is not the primary language). If these subgroups do not constitute a

large share of the TFS population, the research sample might need to be expanded considerably in order to determine how the intervention affects them. Third, the cost of mounting an experiment on a small scale in multiple locations could be considerable and is yet another factor that NCFL will want to consider seriously.

Perhaps most important, we believe, and will argue later in this paper, that the first condition specified above has not yet been met: our examination has suggested that the program might not yet be strong enough to yield strong and lasting impacts. Our assessment has also pointed to pronounced differences in the way the program has been put in place in the different sites. We therefore suggest that it would be wise to defer an impact analysis until program operations have become more stable and more uniform.

The remainder of this paper is divided into eight sections. The next section examines the rationale for the TFS model and preliminary evidence about its potential effectiveness. Section III reviews salient features of the early implementation of the programs in three cities — the strongest performers during the program's first year — as these relate to a possible impact evaluation. Sections IV through VII discuss the issues associated with conducting a rigorous study of program impacts, both in general and as these issues pertain to the TFS program in particular. Section VIII first surveys the magnitude of effects that have been achieved in other demonstration programs; it then considers the TFS program model and its capacity to produce the desired changes. Finally, Section IX recapitulates the major themes and points toward their implications for action.

II. The Program Model and Its Rationale

In contemplating a possible evaluation of an intervention's effects, the first question to address is whether the program model is clearly specified and whether it makes sense – whether theory and preliminary data suggest that the program components, alone or in combination, might reasonably be expected to produce the hoped-for results.

TFS easily clears this first hurdle. The premise of TFS, like that of other NCFL programs, is that interventions that aim to enhance both parents' and children's literacy, as well as to increase the quality and quantity of literacy-related activities that parents and children perform together, will benefit both generations and will also produce greater academic gains for children than efforts that target children without also involving their parents. This premise is grounded in part in a large body of research literature that indicates that maternal educational attainment and parenting practices are strongly related to children's school achievement (see, for example, National Center for Educational Statistics, 1993; Snow et al., 1991; Sticht and McDonald, 1990).

Accordingly, the TFS program model includes four components:

- childhood education (essentially, what happens during the children's regular classroom hours);
- adult education (Adult Basic Education [ABE] classes; classes to prepare students for taking the General Educational Development [GED] test, sometimes

referred to as the "high school equivalency test"; and classes in English as a Second Language [ESL]);

- parenting education (sessions in which adults develop life skills, receive peer support, and learn about child development and various issues with which parents typically grapple); and
- Parent and Child Together (PACT) time (regularly scheduled time during which parents and children engage in literacy-related activities). The model calls for PACT activities to be child-directed.

An earlier study of the Toyota Families for Learning Program, the prototype for TFS, found that adults in the family literacy program registered greater literacy gains than adults in adult-focused literacy programs, while children in the family literacy programs realized larger gains than children in child-focused programs (Philliber, Spillman, and King, 1996). These results did not definitively confirm that the Toyota Families for Learning program actually *was* more effective than the adult- or child-focused programs. The characteristics of the participants in adult- and child-focused programs were different from those in the Toyota Families for Learning program, and this fact might explain the programs' different outcomes. The study indicated, however, the promise of the family literacy concept and suggested that more rigorous evaluation might be warranted.

III. The First Year of TFS: Some Key Considerations

Relevant data pertaining to the first-year implementation experiences of the three sites considered indicate several factors that make detecting program effects difficult: small scale, variability in participant characteristics, variability in program operations, variability in the amount of program services that enrollees actually received and a lower service intensity than planned, the availability of TFS-like services in the larger community, and the propensity of TFS enrollees to take advantage of these services.[1]

- **The TFS programs operated on a small scale within each school, and if current practices persist, they likely to continue to do so. To evaluate program impacts, it will be necessary to pool data across schools.**

Table 1 shows selected characteristics of the programs and their participants, by program location. (Appendix Table A presents some of this information for individual schools.) The TFS programs began enrolling participants several months after school began, between November 1998 and March 1999 (see Appendix Table A). As Table 1 shows, during its first year, 93 families with 117 children were served by TFS in these three sites — an average of 10 families and

[1]The numbers in the following sections are less precise than we would like because of the large quantity of missing responses to many questions about the characteristics of program enrollees and the activities in which they were engaged. We have worked with NCFL to obtain the most reliable data possible, but we are aware that some of our specific numbers — although not, we believe, the central arguments — may be off the mark.

Table 1

Selected Characteristics of Toyota Families in Schools' Programs and Participants, at Program Enrollment, by Site

Characteristic	Site 1	Site 2	Site 3	Total
Program				
Number of households enrolled	35	36	22	93
Number of children enrolled	48	42	27	117
Grade of focal child[a]				
Kindergarten	1		2	3
1	1		1	2
2		1	3	4
3			1	1
4		1	5	6
5	1	1	4	6
6				
Missing	32	33	6	71
Participant				
Relationship of adult to child				
Parent/guardian	27	21	20	68
Other relative	2	2	1	5
Other		5		5
Missing	6	8	1	15
Ethnicity of adult participant				
White (not Hispanic)	5	2		7
Black (not Hispanic)	16	20		36
Hispanic	8	6	22	36
Asian or Pacific Islander	4			4
Other	1	1		2
Missing	1	7	0	8
Primary language spoken in home[a]				
English	14	8	1	23
Spanish	6	2	17	25
Other	7			7
Missing	8	26	4	38
Highest grade completed in school				
No schooling	3			3
LTE grade 8	12	8	13	33
Grade 9-12, no diploma or GED	18	19	6	43
High school diploma or GED		2	1	3
Some college		1	2	3
Missing	2	6	0	8

(continued)

Table 1 (continued)

Characteristic	Site 1	Site 2	Site 3	Total
Annual household income[a]				
Less than $9,000	4	5	1	10
$9,000 to < $15,000	11		7	18
$15,000 to < $20,000	2		2	4
$20,000 or greater	2	1	4	7
Don't know / not sure	8	4	4	16
Missing	8	26	4	38
Primary source of income[a]				
Earnings	22	3	15	40
Alimony / child support	1			1
Government assistance	4	7	2	13
Other			1	1
Missing	8	26	4	38
Employment status of adult[a]				
Working full time	10	1	1	12
Working part time	6		1	7
Not working	10	9	16	35
Missing	9	26	4	39
Prior service receipt				
Welfare	10	13	7	30
Employment training	2	11		13
Vocational education	2	4	4	10
Vocational rehabilitation		2		2
ABE	1			1
ASE (grades 9-12)		1		1
GED preparation	5	22		27
ESL	2	1	13	16
Sample size	35	36	22	93

SOURCE: National Center for Family Literacy Toyota Families in Schools database.

NOTE: [a]Includes only data collected within 30 days after families enrolled in TFS.

13 children per school.[2] Although the host schools ranged in size (see Appendix Table A), the children in TFS made up a small proportion of the student body in each school.

Recruitment efforts during the first year were hampered by the late start of TFS at all sites.[3] Some program administrators also said that they had devoted most of their attention to implementing the TFS program components in the first year, and they planned more comprehensive recruitment strategies for the next year. Still, it is worth noting that sites that began recruitment in November averaged about the same number of families per site as those that began recruitment in February, suggesting that an earlier start would not, in and of itself, increase the number of participating households.[4] Furthermore, when asked by the MDRC field researcher what the ideal size program would be, many program administrators said that they would like to serve 15 families. Thus, although TFS will ultimately expand to include 45 schools, unless there is some outside pressure to expand individual programs it is likely to continue to operate on a very small scale at each school.

- **There was considerable heterogeneity in the group of participant families along several dimensions, including ethnicity, economic circumstances, and grade level of the children enrolled, along with marked differences by site. These differences among participants are important for two reasons. First, greater heterogeneity makes it harder to estimate impacts with confidence. Second, programs may affect different subgroups of participants in very different ways; consequently, understanding subgroup impacts is often a key goal of program evaluation.**

In Site 1, participants came from a number of racial and ethnic backgrounds; in Site 2, the majority of participants were African-American; and in Site 3, all were Hispanic (including many parents who were born in Mexico). (See Table 1.) Not surprisingly, in Site 3, Spanish was the primary language spoken in all but one of the participant's homes, whereas in Site 2, most participants were native speakers of English. In Site 1, participants in two of the three schools were English-speaking, while at the third school, families spoke a variety of languages including English, Spanish, Arabic, Kurdish, and Vietnamese.

As might be expected of a program that promises to increase literacy for parents as well as for children, the majority of adults enrolled in TFS had not completed high school; of those whose educational level was known, 42 percent had not gone beyond the eighth grade. The majority of families enrolled in the TFS program were economically disadvantaged, although only

[2]At some sites, adults who did not have a child of their own in the school were allowed to participate in TFS activities. In some instances, adults participated with younger members of their extended families. At a Site 1 school, for example, an aunt and her nephew participated, as did two adult-child pairs of cousins. The extended family enrollment scenario also occurred in Site 2. In Site 3, parents of children in the Head Start program located at that school participated in TFS, despite not having any school-age children.

[3]The effect of the late start on enrollment was particularly significant at a school in Site 1 where approximately 40 families who received TANF planned to enroll to fulfill welfare-to-work requirements. When TFS did not start in the fall, however, they were assigned to Families 1st , a welfare reform program.

[4]Recruitment strategies during the first year included mailings and phone calls to families in the school, advertisements in the school newspaper, announcements at Parent-Teacher Organization (PTO) meetings, word of mouth, and referrals from parents, teachers and principals.

13 households — seven of them in Site 2 — reported relying primarily upon government assistance. In two sites, the large majority of the adults participating in TFS were not employed; in the third site (Site 1), in contrast, 16 of the 26 participants whose employment status was known worked outside the home.[5]

Although NCFL was unable to collect grade-level data for most children, it appears that children were dispersed across all grade levels, with only a handful of children in any given grade at each school.

- **There were important differences in the scheduling and content of the core TFS components among the three program locations; while some variation is to be expected (and is, indeed, an appropriate response to differences among program participants), these differences may make it more difficult to pool data in an evaluation.**

At all sites, childhood education occurred in a child's regular classroom during the school day. The three TFS components in which adults were engaged — adult education, parenting education, and Parent and Child Together (PACT) time — occurred during the school day in Sites 2 and 3. In Site 1, however, where, as mentioned earlier, most participants worked during the day, adult education classes took place between 5 and 8 p.m. (with food provided for parents and children), and PACT was designed as a set of activities for parents and children to do together at home.

Site 2 participants were predominantly native speakers of English, and the site's adult education component offered ABE and GED classes. TFS enrollees could remain in this activity component for up to six months and partially fulfill their welfare-to-work requirement. In Site 3, the adult education component consisted primarily of ESL classes to meet the needs of its Spanish-speaking participants, while in Site 1, parents took part in ESL, ABE, or GED classes, depending on their needs.

Some parenting education activities were available only to TFS participants. (For example, at one school, the school counselor spoke to a group of TFS parents.) Other activities that sites counted as parenting education were available to all parents, not just those in TFS; for example, all parents at a Site 1 school were welcome to attend a workshop on alcohol and drug prevention, as well as one on "family fun ideas for summer." Some sites also engaged parents in volunteer time or work experience (e.g., by having them assist in the classrooms or in the school cafeteria).

PACT time, the fourth and perhaps most distinctive component of TFS, varied widely across the sites. This variation, in particular, reflects both the programs' adaptations to participant differences (e.g., Site 1's scheduling of PACT time as a set of take-home activities for working parents to do with their children) and the distinct educational approaches and philosophies of

[5]It is worth noting that Site 2 program administrators planned to relocate one of the district's three TFS programs to a different school during the program's second year. They felt that low enrollment at the first school — only eight participants signed on — was a result of the school's location in a working-poor neighborhood, rather than one with a high percentage of TANF recipients.

local program staff. In Site 3, for example, program administrators tried to establish an academic focus for PACT time and viewed teachers as providing the best educational role models for the parents. One principal noted that parents needed to observe teachers in action to see the best instructional practices, adding, "Parents don't know how to interact with their own child." With this perspective in mind, program administrations scheduled TFS activities during the school day, thereby restricting participation to non-working parents or those who did not work 9–5 jobs. At another school in Site 3, PACT time was limited to fourth-graders and their parents because the school principal felt that her fourth-grade teachers would be especially effective in working with family members. The few third-graders in the program did not participate in PACT time, although their parents could be classroom volunteers.[6]

- **Sites differed considerably in the extent to which enrollees actually participated in services that were offered; at some sites, enrollees received a larger dose of the program than their counterparts at others, and in general, the treatment was less intensive than had been planned. Reduced participation in voluntary programs is likely to mean reduced program impacts.**

Because of a substantial amount of missing information from two of the sites, it is difficult to determine precisely how many enrollees actually participated in TFS services, to what extent, or for how long. Table 2 presents our estimates, based on available data from NCFL, of the possible range of values of each of these indicators of participation in adult education and parenting education. (PACT time is excluded because NCFL did not collect data on this component.) The lower estimate presented in the table assumes that the individuals for whom data are missing did not participate at all. The higher estimate assumes that these enrollees had the same participation patterns as those for whom data were available. It seems likely that the true extent of participation lies somewhere within these broad ranges.

In general, the data suggest that during the first year the program was not very intensive. Adult participants spent about six to eight hours a week in adult education classes and about an hour to an hour-and-a-half a week in parenting education. These averages conceal a good deal of variation by school, however. For example, at two schools TFS participants spent less than two hours per month in parenting education, while at two other schools they spent at least 10 hours a

[6]PACT time created other issues for administrators. Some saw the guideline that the component be "child-directed" as inconsistent with their districts' emphasis on implementing a standards-driven curriculum; in their view, PACT time could detract from time teachers needed for instructional purposes. In the same vein, administrators questioned whether parents could help children without knowing what the children were learning. Finally, some administrators came to believe that the parents needed instruction in how to behave in the classroom. For example, the principal of one school remarked, "We need to have more training for them so that they know how to be a volunteer in class. This would include: confidentiality of information … [and] training in how to determine what is confidential…. Also, that you can't come into a classroom high or drunk, or smoking cigarettes." She also indicated two steps her school had taken to address such issues: creating waiting periods before allowing parents in classrooms and holding debriefing sessions with parents afterwards to provide resources and support for them.

Table 2

Participation in the Toyota Families in Schools' Programs by Site

Outcome	Site 1	Site 2	Site 3	Total
Ever participated in (%):				
Any adult education	74-100	56-100	100	73-100
Parenting education	57-80	44-84	95	61-86
Estimated average hours per month in:				
Adult education	18-24	33-59	21	24-33
Parenting education	3-4	3-5	10	4-6
Estimated average months per year in:				
Adult education	2.1-2.9	2.6-4.8	3.8	2.7-3.7
Parenting education	1.8-2.8	2.4-4.4	3.9	2.5-3.7
Estimated average total hours in:				
Adult education	38-70	86-283	80	65-122
Parenting education	5-11	7-22	39	10-22

SOURCES: National Center for Family Literacy Toyota Families in Schools database and site visit field notes.

NOTES: N/A = not applicable.

A low and high estimate of participation are shown. The lower estimate assumes that the individuals for whom data are missing did not participate at all. The higher estimate assumes that these enrollees had the same participation patterns as those for whom data are available.

month in this activity. Data collected through field research suggest that a number of enrollees left the program early and thus did not receive as much of the treatment as had been intended.[7]

- **At many sites various other educational enrichment programs and services were available to assist both adults and children; further, before enrolling in TFS, many participants had received some of the same services offered by TFS and might well have continued to seek out these services if TFS had not existed. These factors may make it less likely that TFS will produce large impacts.**

In general, other school-based initiatives were open to all children, not only to those in TFS families. For example, at one site, Americorps volunteers tutored children in the public schools, and students from a local university served as mentors. Adults and children at all of the TFS sites had additional educational opportunities through community programs. Two school districts also sent parents newsletters that encouraged their engagement in their child's education. In short, TFS was one among several options that parents who wanted to improve their literacy could select.

Further, many TFS participants had previously received education or job training services (see Table 1). Thus, 22 of the 36 adults enrolled in Site 2 had attended GED classes, and 13 of the 22 Site 3 participants had taken part in ESL classes. These findings suggest that if TFS had not been in operation, many TFS enrollees would not have been sitting at home doing nothing. Rather, they might well have sought services and programs elsewhere or displayed initiative in other ways (e.g., by getting a job). In estimating the impact of TFS, the likelihood of nonparticipants receiving other services must be borne in mind as a factor that could reduce the magnitude of the program's effects.

The remainder of this paper explores the challenges to evaluating the TFS program, given the information presented in this section, and considers in greater detail the capacity of the program to create measurable change.

[7]Program administrators commonly cited participants' needing to care for a young child and becoming employed as reasons for dropping out. One administrator noted that her biggest problem in terms of recruiting participants and ensuring their consistent attendance was being able to suggest day care providers for young children. To address this issue, the program maintained a list of neighborhood providers. Her counterpart at another site, when asked about families that left TFS, commented, "They left the program because of jobs that they were able to get. They couldn't afford not to take them." Staff members at one school also noted the crucial role of motivation in promoting continued engagement; as one person put it, "Participants need to understand it's not a quick fix."

Program administrators experimented with a variety of approaches to maintain participation. At one school, an Americorps volunteer called parents if they were absent. In a one-on-one initial meeting with parents at another site, staff members required parents to make a verbal commitment to attend regularly. At one site, financial incentives in the form of $20 gift certificates were awarded to adults who had a 95 percent attendance rate for the month.

IV. Some General Considerations Related to Program Impacts

A. Outcomes vs. Impacts

The objective of any impact analysis is to produce reliable quantitative estimates of the impacts, or *effects*, of a program. Many things in program participants' lives affect what happens to them and how they respond, apart from whatever influence the program may be having. In measuring program impacts, the real question is: What is the influence of the program above and beyond the other circumstances that may shape participants' outcomes?

To derive impact estimates, there must be some estimate of what would have happened in the absence of the program — i.e., a *counterfactual* — that provides a benchmark against which changes resulting from the program itself can be measured. The program's impacts, then, are the difference between the program's results and the counterfactual.[8]

One possible but not very convincing way to measure a program's effects is simply to compare baseline measures with subsequent measures of a particular variable of interest; we refer to these subsequent measures as *outcomes* of the program. Such a comparison would show how much participant behavior, attitudes, and so on had changed along these variables; that is, the earlier measures constitute the counterfactual against which subsequent changes are measured.

In this regard, we note that NCFL seeks to collect a considerable amount of data on TFS participants — both adults and children — when they enroll or shortly thereafter and to collect many of these measures again at the end of the academic year (or upon program exit, for families who leave the program earlier).[9] For adults, the intention is to collect baseline (or near-baseline) and follow-up measures of reading competencies, employment, service utilization, interactions with children, as well as information on their attendance in the program and on the services they received while enrolled. For children, the aim is to obtain baseline and follow-up measures of grades, scores on standardized tests, and other outcomes. Thus, except for the issue of missing data, NCFL can readily derive measures of program outcomes from the information in its database.

The problem with using outcomes as a measure of impacts, however, is that we cannot be certain that any differences are attributable to the program rather than to other factors. For instance, children's reading levels are expected to improve over time simply because they go to

[8]As an example, we suggested above that if TFS did not exist, some families would likely take part in education-related services similar to those TFS offers; such participation would constitute part of the counterfactual.

[9]Upon program entry, adult participants complete a form providing information on the size and composition of their households, their educational attainment, work history, and prior receipt of social and education services, and their reasons for participating in family literacy. The data collection schedule also calls for adults to be assessed within two weeks in one of three areas (depending on their goals): academic achievement (generally using the Tests of Adult Basic Education [TABE]), functional literacy/life skills (using the Comprehensive Adult Student Assessment System [CASAS], or basic skills for those with limited English proficiency (using the Language Achievement Scales [LAS]). Also within 30 days after entry, adults are to complete forms ascertaining information about their own literacy-related activities, household income, employment, and educational achievement; their children's school attendance, experiences, and literacy-related competencies; their educational expectations for their children; and the frequency and nature of their interactions with their children, especially around literacy-related activities. Participating schools also send NCFL the children's school records, as well as their scores on standardized assessments.

school. As another example, while TFS participants may engage in more positive interactions with their children over time, we cannot automatically assume that program participation made the critical difference. The TFS experience during the 1998–1999 academic year points to factors in the local program environments that could, quite aside from TFS, produce positive changes on this dimension: e.g., a new school principal whose mission was to increase parent involvement and a newsletter that was sent to all parents at one site describing educational activities parents could do with their children.

This is not to say that NCFL should curtail its baseline and follow-up data collection efforts. The TFS database has the capacity to yield a great deal of information that is critical to understanding how the program operates. The demographic and socioeconomic data collected upon enrollment can provide administrators with a statistical portrait of who is attracted to the program. And if outcomes are not in the expected direction, program operators may want to consider such issues as the quality and quantity of program services that are delivered, the extent of participant absenteeism and attrition and how these can be reduced, and the fit between the services provided and the characteristics, interests, and needs of the program's clientele.[10] The point is, rather, that these outcomes are not adequate measures of impacts. A better estimate of the counterfactual — one that "controls for" these alternative explanations of program results — is called for.

B. Impacts and Their Magnitude

The preceding discussion suggests that the magnitude of program impacts depends on a number of factors:

- the characteristics of individuals who are in the program's target group;
- the quantity and quality of services that program participants receive;
- the kinds of activities in which nonparticipants engage; and
- characteristics of the local environment, such as the availability and accessibility of adult education classes or conditions in the local labor market, that shape people's behavior.

We have seen that, on average, enrollees in TFS did not receive extensive amounts of program services and that other adult literacy initiatives were available in their communities.

[10]We suggest that these outcome data would be more useful to NCFL if baseline data on participants were collected as early as possible. It seems likely that adults register major learning gains during the first few weeks of program participation, as they are re-exposed to terms and concepts they may have once known but since forgotten. Thus, if baseline measures are taken after people have been in the program for a while, this early learning may not be captured. Instead, their baseline scores will be higher, and the measured outcome of the program lower, than is really is case.

Current practice at some TFS sites is to postpone administration of baseline tests for as long as three weeks after enrollment in the research sample, in order to make participants feel as comfortable as possible in their early days in the program. (In fact, it appears that pretests were never administered at all at one site.) While understandable and sensible from an operational standpoint, postponing baseline measurement is likely to result in underestimates of in-program gains.

Both these factors suggest that, unless the treatment is a strong one, the program's impacts are likely to be modest in size.

A further concept related to the magnitude of program impacts is that of statistical significance. "Statistical significance" refers to the probability that a given effect could have arisen by chance. (Thus, an effect that is significant at the .05 level of significance most commonly used by social scientists is likely to have arisen by chance only 5 times out of 100.) Whether program impacts of a given magnitude can be determined to be "statistically significant" is related to the number of people studied in the research (the "sample size"). In brief, the larger the expected difference between outcomes for program and control group members, the smaller the sample size needed to detect statistically significant effects. Conversely, if impacts are small, a relatively large sample size will be needed to assure that these impacts are not due to chance. The sample size issue is one to which we return in Section VII. First, however, we need to consider the most appropriate research designs for measuring impacts.

V. Measuring Impacts: Selecting a Research Design and Outcome Measures

As we have seen, the objective of any impact analysis is to produce reliable quantitative estimates of the effects of a program by comparing program outcomes with estimates of what these outcomes would have been in the program's absence. Different research designs offer different ways of estimating the counterfactual condition, with random assignment experiments generally being the most highly regarded. Given the scale and voluntary nature of the TFS program, a random assignment evaluation would be best suited for analyzing the program's impacts. Before discussing random assignment in greater depth (and to show why this evaluation design is recommended in the case of TFS), we first consider two alternative designs – the interrupted time series and the comparison-group design.

A. Interrupted Time Series Designs

In an *interrupted time series* design, repeated measures are taken of groups of people and statistical techniques are used to identify discontinuities in the aggregate data patterns. For example, let us suppose that prior reading scores were available on an annual basis from first grade on for children who began to participate in TFS while in fourth grade. Then gain scores from first grade to second, second to third, and so on could be compared with gain scores between fourth and fifth grade, after TFS was introduced. If gain scores were much larger after the children began to participate in TFS, there would be reason to believe that TFS made the crucial difference. We would be even more justified in attributing changes to TFS if the same pattern of greater post-TFS gains were found in multiple settings, so that school-specific factors such as the introduction of a new curriculum could be eliminated as alternative explanations of the impacts.

One problem is that this design is simply not very practical. It requires a great deal of historical data (four or five years' worth) to establish a trend line and deviations from it. Such data are unlikely to be available, especially for children in lower grades. A second problem is that large sample sizes are needed to establish that differences are statistically significant.

It is also worth pointing out that a time-series analysis cannot be used to measure effects on adults. While it may be possible to track at least some children's aggregate test scores over time, because children must routinely take standardized tests, adults do not face comparable requirements. It is hard to imagine how it would be possible to obtain data for parents at several points in time.

B. Comparison-Group Designs

As the name suggests, *comparison-group designs* compare recipients of program services with nonrecipients who are selected to be as similar to the program group as possible along the most salient dimensions. For example, we might compare TFS families with families in schools at which TFS has not been put in place, or with participants in other programs, or with a comparison group selected from a national data base. To make the groups more equivalent, we might restrict the comparison group to low-income families in which the mothers have relatively low levels of education. We would then track the behavior and outcomes of members of both groups over time.

The major problem associated with comparison-group designs is that of *selection bias*– the fact that there is no assurance that the program and comparison groups are indeed similar at baseline, or that all differences between them can be identified, measured, and then controlled for statistically. If the two groups differ in important but unmeasured ways — e.g., in motivation — the results of the study will be biased. Given the voluntary nature of TFS, the potential for selection bias is serious. Even if comparison-group members could be selected among other parents at the TFS schools, it would be reasonable to assume that those parents who join the program differ along important dimensions from those who opt not to participate. For example, TFS parents may be more than usually motivated to achieve the goals the program espouses, or they may perceive themselves as needing the services more, or they may differ in other important — but hard-to-measure — ways from those who don't enlist.

A further problem is that if the comparison group is drawn from a different environment from that of the program group (as would be the case, for example, if comparison group families came from non-TFS schools, or were selected from a national data base), one could not be certain that any differences arising between TFS participants and comparison group members were attributable to the TFS program, rather than to differences in the immediate environments of the two groups.

C. Random Assignment Experiments

Most evaluators believe that a *random assignment experiment* (sometimes referred to as a *controlled experiment*) is the methodology that yields the most rigorous and credible estimates of program impacts. In this design, families eligible for an intervention are assigned at random to either a program group or a control group in a procedure that can be likened to a lottery. Program group members are provided with the regular array of program services. Control group members are excluded from the program being evaluated but can participate in any other programs in their communities (or, in the case of TFS, in any other programs in the TFS schools). As with a comparison group design, follow-up data are collected for both groups.

The strength of an experimental design is that randomization allows the evaluator to assume that the program and control groups are equivalent prior to the intervention (or that any differences between them are random rather than systematic). Moreover, aside from the presence of the intervention for one group, the same environmental factors influence program and control group members alike. Consequently, if the evaluation is well conducted, we can conclude that the intervention, rather than other potential explanations, *caused* any observed differences that subsequently emerge in the outcomes of the two groups.

While the strengths of random assignment experiments are widely recognized, it is important to be aware of what such experiments cannot tell us. In particular, neither a random assignment design nor any other design allows investigators to "get inside the black box" — i.e., to determine which components in a multifaceted intervention make the critical difference — or if, indeed, it is not the components but something else that promotes change (e.g., the relationships that develop between program staff and participants). The program treatment is, rather, considered as a whole. Statistical analyses as well as implementation research can shed light on the key factors contributing to change, but quantitative estimates of component effects will be at best suggestive, not conclusive.

Further, because control group members can receive non-TFS services (and the data presented in Section III indicate that they are likely to do so), this research design will not compare families who participate in TFS with families who receive no services at all. (Neither would a comparison-group design.) Rather, it will capture the *marginal* effects of TFS, above and beyond whatever other services families may receive. This latter comparison seems more "true-to-life," since other programs, which seek to accomplish some of the objectives of TFS but do not provide all of its services, are widely available. But to the extent that control group members actually do receive TFS-like services, it may be more difficult to detect the effects of TFS per se.

We are not arguing that a random assignment study is the only way to evaluate the impacts of the TFS program, but we believe that random assignment is the only methodology that is likely to produce findings about program impacts that will be widely acknowledged as reliable and believable. In the next two sections of this paper we confine our discussion to a fuller explanation of this design. In Section VI, we consider general requirements associated with the implementation of random assignment experiments as they relate to current TFS practice. In Section VII, we consider the specific sample sizes needed to measure program effects on children's literacy.

VI. Random Assignment: General Requirements and TFS Practice

Within the context of a random assignment evaluation, one critical question is whether the objective of the evaluation is to assess the effects of an intervention on those *eligible* to participate or on those who actually *do* (or are likely to) participate. This choice has important implications for the point at which random assignment takes place, the sample size, and other design elements.

In the TFS context, the first alternative would mean studying a group of families who have been identified by teachers, principals, or others as suitable for TFS, whether or not they ultimately choose to take part. These TFS-appropriate families would be randomly assigned to

two groups; then program staff would have to recruit families in the "program group" actually to enroll in the program. This approach positions evaluators to answer a question of considerable interest to program operators, policymakers, and funders: What percentage of those eligible for a program can be induced to join it? It also addresses the question of the magnitude of impacts that could be achieved if the program were to move to a larger scale at a particular site. But since program impacts will be averaged over both the participants and nonparticipants within the program group, to achieve positive impacts program operators will need to do a strong "selling job" to persuade as many eligible families as possible to take part.

The second alternative entails studying a group of families that have expressed active interest in participating in TFS and are likely to do so. Here, randomization would occur once families have moved toward enrolling in the program. This approach is not geared toward answering the question of how many families can be persuaded to sign on, since the study subjects will be families that have already "bought into" the program. But it means that program effects will primarily reflect the effects of participation, not a mixture of some participation and widespread nonparticipation. The assumption underlying this paper is that this second alternative is the one of major interest to NCFL.

Implementing random assignment experiments successfully depends on several conditions being met:

- The sample size must be adequate.
- The program must attract more people than it can actually serve.
- Program staff need to understand and cooperate with random assignment.
- Parents must also accept the process.
- Uniform follow-up data must be collected on the same timetable for program and control group members.
- Data must be collected for all research sample members, including program group members who do not receive any TFS services and control group members who receive non-TFS services.

The sample size needed for an evaluation of TFS is discussed in Section VI. Here, we consider the other requirements. We recognize that in some instances, fulfilling these requirements could mean important changes in the ways that TFS programs currently operate.

More interested families than available slots. During its first year, TFS was generally able to serve all interested families. (One school in Site 1 had a waiting list.) Whether this will be true in future years — when TFS operations will start at the beginning of the school year instead of midway through it and when word of mouth about TFS may have spread through the community — remains to be seen.

If TFS attracts many more families than it can serve, random assignment can function like a lottery to allocate scarce program slots. If, on the other hand, the standard recruitment measures yield a shortfall of potential participants, then program staff would need to undertake additional

efforts to inform people about TFS and to identify families that are interested in the program and willing to participate in the random assignment selection process.

Staff and parental buy-in. Cooperation by program staff members and acceptance by parents are essential if a random assignment design is to yield the unbiased estimates it promises. Staff members must be willing to recruit more interested families than the program can accommodate. They cannot "game" the random assignment process. They cannot allow those assigned to the control group to receive the services designated uniquely for those in the program group (at least not until the research is concluded). And they also should not make extraordinary efforts to assist control families in finding services elsewhere in the community.[11] Parents must also be willing to subject themselves and their children to random assignment and to abide by its outcome.

Staff members and parents sometimes have ethical objections to random assignment because it means denying services to some individuals. Experience indicates that they can sometimes be won over to the procedure by the argument that random assignment is the fairest way of allocating scarce resources. (A "first come, first served" approach, by contrast, may reward the most assertive applicants for services, or those who are most "in the know" about what services are available.)[12] Staff members and parents may also be induced to accept random assignment by the arguments that the restrictions need not be permanent, that they are critical to the success of the evaluation, and that solid evaluation results, in turn, are required to persuade funders that the program should be continued — or to indicate that it needs to be changed to become more effective.[13]

Uniform data collection timetables for program and control group members. It is also critical to collect baseline and follow-up data according to the same schedule for sample members in both groups. Figure 1 illustrates this point.[14] In this hypothetical example, scores on a key indicator are rising for both program and control group members over time. If data are collected for program group members on an earlier timetable than for controls (perhaps because those in the program are easier to find), then the impact of the program — the difference in scores between the two groups — may well be understated. (In the example shown, the true impact of the program is 40 points: P_2-C. But if program group data were collected earlier, it would appear to be only 10 points: P_1-C.)

Data on all research sample members. In measuring the impacts of a program, follow-up data must be analyzed for *all* research sample members — including program group members who never participate (or drop out early) and control or comparison group members who receive

[11]To secure staff members' support for the study, program administrators may permit them to give families assigned to the control group a list of other resources to seek out on their own.

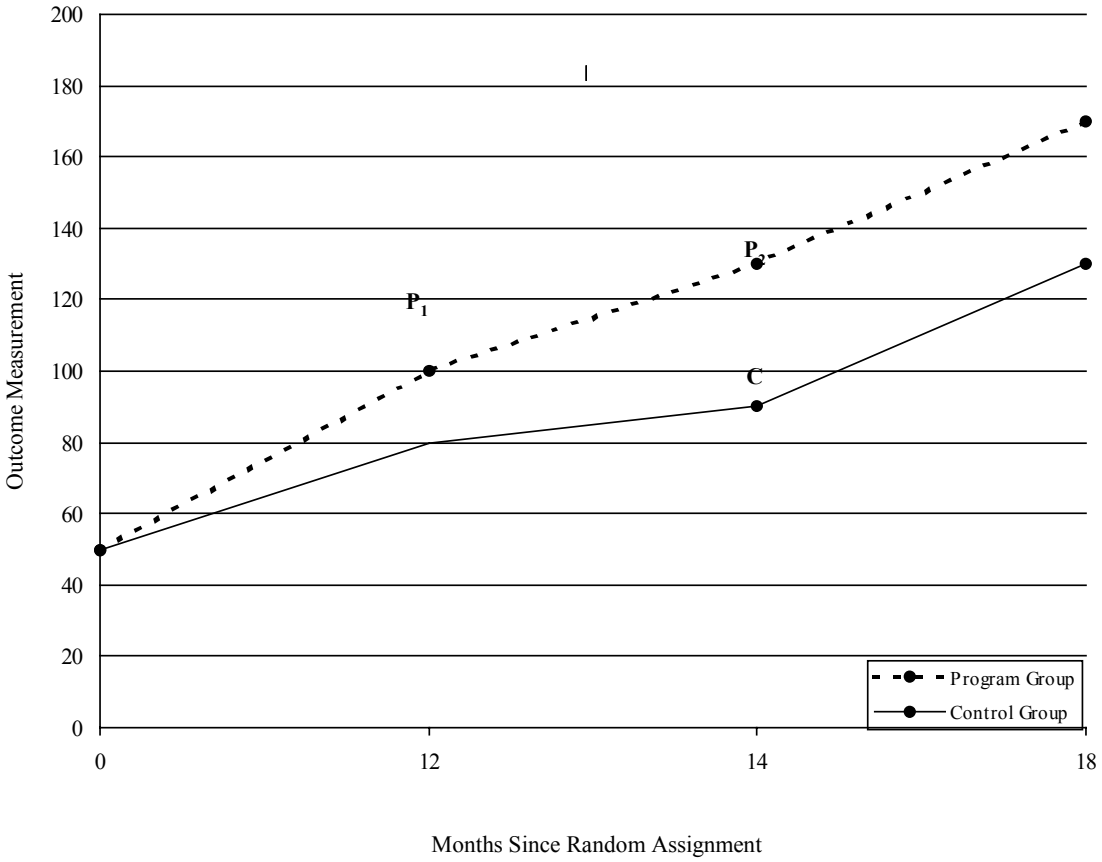
[12]In this regard, it is worth pointing out that social programs may not register the largest impacts on those participants who appear to be easiest to serve. This is because such participants might fare reasonably well on their own, without the program's assistance. Impacts may instead be larger for hard-to-serve participants who would do much less well without the program's aid.

[13]If program impacts are to be measured over only one year, for example, controls might be allowed to receive these services after the year has elapsed.

[14]We are indebted to Fred Doolittle for this discussion and example.

Figure 1

Effect of Different Timing of Data Collection



SOURCE: Adapted from Orr (1999), p.177.

NOTE: Correct impact estimate: $P_2 - C = 40$. Incorrect impact estimate: $P_1 - C = 10$.

services similar to those delivered by the intervention. This is essential in order to preserve the initial equivalence of the program and control groups; otherwise, one would be comparing outcomes for two groups of individuals who could be expected to differ on many counts. For example, differences can be anticipated between those in the program group who persist and those who drop out, as well as between program group persisters or dropouts and all control group members. A study that did not include all sample members would not yield reliable conclusions about program impacts.

VII. Sample Size: A Critical Issue

As noted above, the sample size must be large enough to determine with a high degree of statistical assurance that any differences between program and control group members did not arise by chance. In this section, we consider the sample size that would be needed to evaluate the impacts of TFS. The section addresses two questions: What size impacts are generally considered "large," "medium," and "small"? And, what sample sizes would be needed to detect such impacts? For purposes of discussion, in this section, we use scores on standardized tests as the critical variable of interest; in Section VIII, we examine the magnitude of effects achieved on a range of other outcome variables.

A. Establishing a Common Metric for Impacts: Effect Size

NCFL permits schools to satisfy the TFS data requirements by submitting whatever measures of children's reading skills the schools normally collect. In fact, the three study sites used different measures to assess children's literacy. This policy of allowing schools to use whatever standardized tests they would normally administer makes eminently good sense from an operational perspective; it helps minimize the additional data collection that is generally required by research and demonstration programs.

From the perspective of an impact evaluator, this policy creates the need to find a common metric of impacts across the reading tests that are employed. One common metric that can be used is *effect size.* Statistically, the effect size is the difference between the mean scores for program and control groups divided by the standard deviation, a measure of score dispersion around the mean for both groups.[15] As a practical example, suppose that students in the TFS program and control groups were to take a standardized test with a standard deviation of 20 points; if program group members scored 10 points higher, on average, than their control group counterparts, the effect size would be .5.

It is possible to calculate an average effect size as follows: Suppose that the 15 school districts in which TFS will be located by the end of the 2000–2001 academic year use 15 different standardized tests to measure reading ability for students whose native language is English. Suppose further that all three schools in each district use the same standardized test. Within each district, reading scores for students in the TFS program group could be compared with reading scores for students in the control group, and an effect size could be calculated. Then, an average

[15]In a normal distribution, 95 percent of the sample lies within about 2 standard deviations on either side of the mean.

effect size could be calculated across the 15 districts and 15 tests (weighted by the number of students taking each test).[16]

One question to consider carefully is whether the average effect size should be calculated *for each grade separately.*[17] (If so, a substantially larger sample size would have to be established.) On the one hand, one can argue that the objective of TFS is to produce effects for *schoolchildren*, not for first-graders or third-graders or fifth-graders. On the other hand, looking at effects across all grades introduces an element of both conceptual and statistical fuzziness, since, as we have noted several times, the greater the variation in the sample, the greater the measurement error (especially in small samples). Greater measurement error makes it less likely that a given effect will be statistically significant.

Furthermore, as noted previously, evaluations have frequently shown that some subgroups especially benefited from an intervention, while others were unaffected or even harmed. Both evaluators and program operators might want to know about these differential subgroup impacts. If TFS program planners and administrators are interested in learning about grade-specific impacts, they might want to focus recruitment efforts and limit the evaluation to families with children in those grades where TFS is hypothesized to make a crucial difference. (For example, they might want to look at the program's effects on children in first grade who are just learning to read.)

B. The Magnitude of Effect Sizes and the Sample Sizes Needed to Calculate Them

By convention, statisticians have come to agree, more or less, on effect sizes that should be considered large, medium, and small: .90, .45, and .15, respectively.[18] That is, a medium-sized effect is one in which the mean score for the program group is .45 of a standard deviation larger than the mean score for the control group. It is extremely rare for interventions to achieve large effects, so we ask: What would it take to detect medium-sized and small effects in TFS?

We begin with the assumption that the sample will be evenly divided between program and control groups.[19] Table 3 shows the number of schools in which an evaluation would have to be mounted in order to detect medium-sized and small effects, when two other conditions are varied. The first is the number of students in each school who are randomly assigned and participating in the study. We choose 10 as a number that seems easily attainable, 30 as a "stretch tar-

[16]NCFL will want to consider whether effect size should be used as a common metric in cases where the underlying tests used do not measure precisely the same underlying construct. This is the case for the reading tests administered to adults in TFS: the TABE measures "academic" knowledge, while the CASAS is designed to measure reading in a life skills context, and the LAS is geared toward non–English speakers only. On the other hand, all three tests are measures of literacy.

[17]Because of the large quantity of missing data in the TFS information system, it is impossible to estimate the number of students in each grade with confidence.

[18]Lipsey (1990), who proposes these specific values, bases these conventions on a meta-analysis of the effects of 186 educational, psychological, and behavioral treatment programs.

[19]It would be possible to adopt a random assignment ratio that included more program than control group members — 2:1, for example. However, any departure from a 50-50 ratio would entail a larger *total* sample size than would an even split.

Table 3

Number of Schools Needed to Detect Medium-Sized and Small Effects With a 50-50 Treatment - Control Split, With and Without Pretests, and With Different Sample Sizes

Data availability	Effect size = .45	Effect size = .15
Pretest data available (R^2 =.45)		
Number of students per school = 10	7	62
Number of students per school = 20	4	31
Number of students per school = 30	3	21
Pretest data not available (R^2 =0)		
Number of students per school = 10	13	112
Number of students per school = 20	7	56
Number of students per school = 30	5	37

SOURCE: Calculations of the minimum detectable effect size (MDES) at 0.05 significance levels, with 80% power, for a one-tailed test =

$$2.5\sqrt{\frac{\left(1-R^2\right)}{(P)(1-P)nM}}$$ where,

R^2 = the explanatory power of the pretest,
P = the proportion of sample members in the program group,
n = the total number of sample members per school, and
M = the total number of schools.

get," and 20 as an in-between figure. The second variable is whether or not pretest scores (e.g., measures of children's literacy in the prior year) are or are not available.

The table indicates that to detect an effect size of .45 when pretest scores are available, a total sample of about 70 students would be needed. Assuming that each school randomly assigns 10 students to program and control groups, seven schools would need to be included in the research sample. Four schools would have to participate if the number randomly assigned were increased to 20, and three schools if the number were 30.

If, on the other hand, program impacts are likely to be quite small (an effect size of .15), then it seems unwise to conduct an impact evaluation, for two main reasons. First, it will take an unrealistically large sample to ascertain that the effects are statistically significant. With 10 students in each school participating in the experiment, 62 schools would have to participate — and there are only 45 schools slated to participate in the demonstration altogether! Even if 30 students per school were randomly assigned, 21 schools would still have to take part in the evaluation. It seems doubtful that a very large number of schools will assent to random assignment, and the cost of mounting and monitoring an evaluation at so many sites would likely be prohibitive. Second, aside from logistical considerations, NCFL officials will need to decide whether an effect size of .15 is policy-relevant.

We note the critical importance of pretest data for keeping sample sizes manageable. While a statistical explanation of the role of pretests is beyond the scope of the paper, pretest data are essential because they reduce the amount of error associated with estimating impacts. As the bottom panel of Table 3 shows, without pretest data it would be necessary to increase substantially the number of schools participating in the study to detect effect sizes of the magnitudes shown.

VIII. The TFS Program Model, Participation, and Program Impacts

A critical question is whether the TFS program, as it currently operates, is likely to achieve impacts that are both statistically significant and large enough to be policy-relevant. While this question cannot be answered conclusively, we will address the evidence in two ways. First, we will review the effect sizes achieved in several studies that included child outcomes. Then, we will examine more closely the impacts achieved by two interventions that resemble TFS in a number of respects.

A. Effect Sizes Achieved by Several Interventions

Table 4 shows the effect sizes achieved on a range of child outcomes in several demonstration programs. Some of the interventions studied aimed primarily at improving the parents' human capital, while others were directed mostly toward the children. All these programs were evaluated using rigorous random assignment designs.

It is immediately apparent that none of the programs achieved what would be considered "large" impacts. In fact, the effects registered by the Perry Preschool Program and the Abecedarian Project — two of the best-known and most highly regarded programs targeted toward preschoolers — could be characterized only as only "medium-sized." (It is notable, moreover, that a

Table 4

Effect Sizes Achieved by Demonstration Programs

Program name and outcome variable	Effect Size
Programs aimed at improving parents' human capital	
Teenage Parent Demonstration	
Adaptive Social Behavior Inventory	
Site 1	0.10
Site 2	0.07
Site 3	0.11
Peabody Picture Vocabulary Test	
Site 1	0.07
Site 2	0.02
Site 3	-0.06
National Evaluation of Welfare to Work Strategies	
Total Behavior Problems	
Site 1	0.00
Site 2	0.10
Site 3	0.00
Bracken School Readiness Subscale	
Site 1	0.08
Site 2	-0.01
Site 3	0.11
Programs aimed at improving outcomes for preschoolers	
Perry Preschool Program effects at age 7	
Total school achievement	0.34
Stanford-Binet IQ	0.45
Non-verbal intellectual performance	0.12
Peabody Picture Vocabulary Test	0.26
Carolina Abecedarian Project effects at age 12	
Preschool group:	
Wechsler Intelligence Scale for Children	
Verbal	0.56
Performance	0.22
Woodcock-Johnson Psychoeducational Battery	
Reading	0.48
Math	0.35
Written language	0.41
Knowledge	0.61

(continued)

Table 4 (continued)

Program name and outcome variable	Effect Size
School-age group:	
Wechsler Intelligence Scale for Children, age 12 years	-0.09
Verbal	0.22
Performance	-0.36
Woodcock-Johnson Psychoeducational Battery	
Reading	0.17
Math	0.08
Written language	0.36
Knowledge	0.06

SOURCES: Campbell, Frances A. and Craig T. Ramey. (1994). Effects of Early Intervention on Intellectual and Academic Achievement: A Follow-up Study of Children from Low-Income Families; Freedman, Steven, et al. (1999). *National Evaluation of Welfare-to-Work Strategies: Evaluating Alternative Welfare-to-Work Approaches: Two-Year Impacts for Eleven Programs;* Granger, Robert and Rachel Cytron. (1998). Teenage Parent Programs: A Synthesis of the Long-Term Effects of the New Chance Demonstration, Ohio's Learning, Earning, and Parenting (LEAP) Program, and the Teenage parent Demonstration (TPD); Kisker, Ellen, Anu Rangarajan, and Kimberly Boller. (1998). *Moving Into Adulthood: Were the Impacts of Mandatory Programs for Welfare-Dependent Teenage Parents Sustained After the Program Ended?* ; Schweinhart, Lawrence J., Helen Barnes, and David P. Weikart. (1993). *Significant Benefits: The High/Scope Perry Preschool Study Through Age 27.*

version of the Abecedarian project that was directed toward children in elementary school rather than preschoolers achieved much smaller impacts.) The programs aimed at increasing the educational levels of the parents achieved effects that could be described as negligible to small. While these data cannot speak to the impacts of TFS, they do suggest that even medium-sized effects are uncommon and difficult to achieve.

B. TFS and Other Programs Aimed at Both Parents and Children

The evaluations of the Even Start Family Literacy Program (St. Pierre et al., 1996) and of the New Chance Demonstration (Quint, Bos, and Polit, 1997) provide suggestive comparisons with TFS. All three programs were voluntary initiatives that were directed to low-income families with young children (although the target populations differed in several respects).[20] All included adult education and parenting education as core components; Even Start included early childhood education as well. The impacts of both Even Start and New Chance were rigorously measured using a random assignment methodology. (This part of the Even Start evaluation is known as the "In-Depth Study," and it provides the most reliable evidence about the program's effectiveness.)

Table 5 presents information on the extent of participation in program services for all three demonstrations and summarizes the impacts of Even Start and New Chance on adults and children. It indicates that, on average, Even Start and New Chance participants received very similar amounts of adult education (107 and 101 hours, respectively), but Even Start enrollees received a much greater amount of parenting education (58 hours versus 18 hours).

The third column of the table presents the participation information available from TFS. This information is incomplete in four ways. First is the missing data problem we have noted, and the resulting imprecision of the estimates that are shown. Secondly, we recognize that the first-year experience of enrollees was truncated in that the program began well into the 1998–1999 school year; we assume that some TFS participants would have joined the program earlier in the year, and would have remained longer, had they been able to do so. Third, we do not know how many first-year enrollees remained in the program during the 1999–2000 academic year. (We suspect that some did so, while others found jobs, lost interest, or left the program for other reasons.) For this reason, we do not know the *total* number of hours of adult education and parenting education that these TFS enrollees will *ever* receive throughout their stay in the program. Finally, PACT time is excluded from the picture.

---

[20]TFS parents (at least those enrolled during the program's first year at the three study sites) could be characterized as more disadvantaged than their Even Start and New Chance counterparts on some counts, less disadvantaged on others. For example, about one in five Even Start adults had completed high school, compared with 6 to 7 percent of adults in TFS. On the other hand, only 14 percent of TFS adults reported that their primary source of income was government assistance, while 49 percent of Even Start families received most of their support from welfare. TFS parents were much better off economically than the young mothers in New Chance, all of whom received public assistance. But only 13 percent of the New Chance mothers had completed eight or fewer years of schooling, compared with between 39 and 42 percent of the TFS parents.

To participate, children in Even Start households had to be less than eight years old at baseline. Children in New Chance households were considerably younger, averaging about 18 months old when their mothers entered the program. At the 36-month follow-up, the New Chance focal children were just under five years old, on average, but ranged in age between 3 ½ and 10 years old.

Table 5

Participation and Effects of Three Programs Directed Toward Disadvantaged Families

Outcome	Even Start	New Chance	Toyota Families in Schools
Participation			
Average number of hours per month participants engaged in:			
Adult education	13.5	--	24-33
Parenting education	6.5	--	4-6
Average number of total hours participants engaged in:			
Adult education	107	101	65 - 122
Parenting education	58	18	10 - 22
Effects			
Effects on children	Transient positive effects on school readiness; no effect on receptive vocabulary	No effect on cognitive or socio-emotional development (Bracken, ES = -.07); unexpected small but negative effects on mothers' ratings of children's behavior (BPI, ES = .12)	N/A
Effects on adults	Higher rate of GED attainment; no effects on functional literacy, employment status, income, or psychological variables	Higher rate of GED attainment; higher rates of depression, reported stress; no effects on reading test scores, skills, training attainment, employment, welfare, or income	N/A
Effects on home environment	Increased printed matter in home, otherwise no effects	Positive effects for mothers not at risk of depression; no effects overall	N/A
Other comments	Increased time in parenting education was associated with children's higher receptive vocabulary	Women who had been out of school longer and were at high risk of depression experienced especially adverse outcomes, as did their children	N/A

SOURCES: National Center for Family Literacy Toyota Families in Schools database; Quint, Bos and Polit. (1997). *New Chance: Final Report on a Comprehensive Program for Young Mothers in Poverty and Their Children*; St. Pierre et al. (1996). *Improving Family Literacy: Findings from the National Even Start Evaluation.*

NOTES: N/A = not applicable.
ES = effect size.
Bracken = Bracken Basic Concept Scale.
BPI = Behavior Problems Index.

With these data limitations acknowledged, the findings nonetheless indicate that the extent of participation in adult education was fairly similar among enrollees in all three programs. TFS enrollees participated in parenting education about as much as their New Chance counterparts, but much less than did Even Start enrollees.

What, then, can be said about the effects of the two earlier interventions? In brief, neither intervention produced many durable positive effects on the children, their parents, or the home environment.

Looking first at child outcomes, Even Start participation initially increased children's school readiness, but by 18 months after random assignment, control group children, who had by then entered preschool or kindergarten, had caught up with the Even Start group. In other words, Even Start had a statistically significant but transient effect on school readiness: over time, the combination of early childhood education with adult-focused services did not have a stronger effect on school readiness than early childhood education alone.

New Chance also did not have the hoped-for positive effects on children's cognitive and socioemotional development. Indeed, there were small but statistically significant differences in the direction opposite to what was expected and desired: mothers in the program group rated their children as having *more* behavior problems than those in the control group. This may reflect the higher level of depression and stress found among the program group mothers.

Participation in both Even Start and New Chance led to a substantial increase in the proportion of adults attaining a GED. There is no evidence, however, that either program improved adults' functional literacy: Even Start and control group members obtained similar gains on the CASAS, and New Chance and control group members scored similarly on the TABE. Neither program affected adults' employment status or the level or sources of household income. Even Start did not alter participants' perceptions of social support, their levels of depression, or their sense of mastery. Participation in New Chance, as noted above, led to increased levels of stress and depression. Finally, neither program had multiple or lasting effects on participants' home environments.

There is some evidence available from both evaluations that more participation is better than less, although this conclusion, because it is not grounded in the random assignment research design, is much less certain than the findings cited above. While in general the Even Start program did not have any effect on children's receptive vocabulary, additional analyses indicated that the more time that parents participated in parenting education, the greater the gains in their children's vocabulary. Receiving more than 18 weeks of instruction in New Chance adult basic education and GED classes was associated with a higher rate of GED attainment, and members of both the program and control groups who received a GED (or who participated in skills training or college) had higher earnings than they would have otherwise.[21]

[21]These analyses are not wholly credible because they could not completely control for selection bias — the possibility that adults who elect to spend more time in parenting education classes would also be more likely to engage in other activities that would improve their children's vocabulary, or that those who exhibit greater perseverance in education classes would do better in the labor market in any event.

The Even Start and New Chance results indicate how hard it is for programs to make a measurable and lasting difference in participants' lives. They suggest that if TFS is to be more effective than the other programs, it will need to provide services that are far stronger, more intensive, and of longer duration than its predecessors.

IX. Conclusion

In laying out the conditions for a rigorous experimental study of TFS earlier in this paper, we noted as the first condition that the program must be capable of producing a medium-sized effect on enrollees. Later we noted that if the effect is a small one, it is highly unlikely that we could detect it given the small number of participants at each of the schools involved in the demonstration. Judging from the performance of the strongest local programs during the TFS start-up year and comparing that performance with those of previous programs, we do not believe that TFS is likely to produce medium-sized effects at this time.

We began by saying that the family literacy concept holds promise. We still very much believe that this is true. But it is very hard for any program to realize that promise when it is just getting under way. We believe that at this juncture it is important to focus on strengthening the programs at the TFS sites — ensuring that the services offered are plentiful, substantive, and meaningful and that absenteeism and attrition are kept to a minimum. NCFL, through its strong technical assistance capacity, is well equipped to guide this effort to strengthen and consolidate program operations and may wish to do so before undertaking an evaluation of the initiative's effects. In the meantime, ongoing process evaluation can provide much-needed information about how to boost demand, sustain participation, and enhance service quality.

References

Campbell, F. A., and C. T. Ramey. 1994. "Effects of Early Intervention on Intellectual and Academic Achievement: A Follow-Up Study of Children from Low-Income Families." *Child Development,* 65: 684-698.

Freedman, S., D. Friedlander, G. Hamilton, J. Rock, M. Mitchell, J. Nudelman, A. Schweder, and L. Storto. 2000. *National Evaluation of Welfare-to-Work Strategies: Evaluating Alternative Welfare-to-Work Approaches: Two-Year Impacts for Eleven Programs.* New York: Manpower Demonstration Research Corporation.

Granger, R., and R. Cytron. 1998. "Teenage Parent Programs: A Synthesis of the Long-Term Effects of the New Chance Demonstration, Ohio's Learning, Earning, and Parenting (LEAP) Program, and the Teenage Parent Demonstration (TPD)." New York: Manpower Demonstration Research Corporation.

Kisker, E., A. Rangarajan, and K. Boller. 1998. *Moving into Adulthood: Were the Impacts of Mandatory Programs for Welfare-Dependent Teenage Parents Sustained After the Program Ended?* Princeton, NJ: Mathematica Policy Research, Inc.

Lipsey, M. W. 1990. *Design Sensitivity: Statistical Power for Experimental Research.* Newbury Park, CA: Sage Publications.

National Center for Educational Statistics. 1993. *Adult Literacy in America.* Washington, DC: Office of Educational Research and Improvement.

Philliber, W. W., R. E. Spillman, and R. E. King. 1996. "Consequences of Family Literacy for Adults and Children: Some Preliminary Findings." *Journal of Adolescent and Adult Literacy,* 39 (7): 558-565.

Quint, J. C., J. M. Bos, and D. F. Polit. 1997. *Final Report on a Comprehensive Program for Young Mothers in Poverty and Their Children.* New York: Manpower Demonstration Research Corporation.

Schweinhart, L. J., H. Barnes, and D. P. Weikart. 1993. *Significant Benefits: The High/Scope Perry Preschool Study Through Age 27.* Ypsilanti, MI: High/Scope Press.

Snow, C. E., W. .S. Barnes, J. Chandler, I. F. Goodman, and L. Hemphill. 1991. *Home and School Influences on Literacy.* Cambridge, MA: Harvard University Press.

Sticht, T. G., and B. A. McDonald. 1990. *Teach the Mother and Reach the Child: Literacy Across Generations. Literacy Lessons.* Geneva: International Bureau of Education.

St. Pierre, R. G., J. P. Swartz, S. Murray, and D. Deck. 1996. *Improving Family Literacy: Findings from the National Even Start Evaluation.* Cambridge, MA: Abt Associates Inc.

Appendix Table 1

Program-Related Characteristics of Individual Schools in the Toyota Families in Schools Program

Site	Month first family enrolled	Number of households	Number of children	Total number of students housed in school
Site 1				
School 1	February	8	9	300-500
School 2	February	11	13	300-500
School 3	February	16	26	>500
Total	N/A	35	48	N/A
Site 2				
School 4	November	17	18	>500
School 5	November	8	8	300-500
School 6	November	11	16	>500
Total	N/A	36	42	N/A
Site 3				
School 7	March	7	8	>500
School 8	November	6	8	300-500
School 9	January	9	11	<300
Total	N/A	22	27	N/A

SOURCE: National Center for Family Literacy Toyota Families in Schools database.

NOTE: N/A = not applicable.

About MDRC

The Manpower Demonstration Research Corporation (MDRC) is a nonprofit, nonpartisan social policy research organization. We are dedicated to learning what works to improve the well-being of low-income people. Through our research and the active communication of our findings, we seek to enhance the effectiveness of social policies and programs. MDRC was founded in 1974 and is located in New York City and San Francisco.

MDRC's current projects focus on welfare and economic security, education, and employment and community initiatives. Complementing our evaluations of a wide range of welfare reforms are new studies of supports for the working poor and emerging analyses of how programs affect children's development and their families' well-being. In the field of education, we are testing reforms aimed at improving the performance of public schools, especially in urban areas. Finally, our community projects are using innovative approaches to increase employment in low-income neighborhoods.

Our projects are a mix of demonstrations — field tests of promising program models — and evaluations of government and community initiatives, and we employ a wide range of methods such as large-scale studies to determine a program's effects, surveys, case studies, and ethnographies of individuals and families. We share the findings and lessons from our work — including best practices for program operators — with a broad audience within the policy and practitioner community, as well as the general public and the media.

Over the past quarter century, MDRC has worked in almost every state, all of the nation's largest cities, and Canada. We conduct our projects in partnership with state and local governments, the federal government, public school systems, community organizations, and numerous private philanthropies.