Contract No.: 03Y00416401D/SIN 874-3
MPR Reference No.: 6030-180/260/370

MATHEMATICA
Policy Research, Inc.

Design Options for the Assessment of Head Start Quality Enhancements

Final Report
Volume I

September 30, 2005

Christine Ross
Gretchen Kirby
Peter Schochet
John Hall
Susan Sprachman
Kimberly Boller
Diane Paulsell
Sheena McConnell

Submitted to:

Office of Planning, Research and Evaluation
Administration for Children and Families
370 L'Enfant Promenade
Seventh Floor West
Washington, DC 20447

Project Officer:
Anne Bergan

Submitted by:

Mathematica Policy Research, Inc.
P.O. Box 2393
Princeton, NJ 08543-2393
Telephone: (609) 799-3535
Facsimile: (609) 799-0005

Project Director:
Christine Ross

CONTENTS

Chapter **Page**

Chapter **Page**

Chapter **Page**

Chapter **Page**

TABLES

Table **Page**

Table **Page**

FIGURES

CHAPTER I

INTRODUCTION

Head Start, the largest federally funded preschool program, provides comprehensive services to economically disadvantaged children and their families so that children can enter kindergarten ready to succeed in school. Performance standards for the program include requirements for the intensity and quality of a broad range of services for children and families. Head Start programs must offer education, health, and nutrition services to children, offer social services to their families, and provide opportunities for parents' involvement in the programs. Head Start is designed to enhance children's cognitive skills, social development, physical and mental health, and good nutrition. Programs also are expected to support the parent as the child's economic provider, first teacher, and primary advocate for education and health services. Some tailoring of program services is expected to meet the needs of diverse communities. Education services must be appropriate to children's linguistic backgrounds and developmental needs, and family services must be individualized to meet parents' goals and needs.

Head Start has long emphasized the importance of continuous program improvement and, in keeping with this emphasis, has invested significant resources in strategies to enhance the quality of program services. Since its early years, Head Start has considered itself a national laboratory for developing good early childhood practice through the development of innovative approaches and the honing of "best practices" that are based on professional wisdom. In many cases, research partnerships with universities have been a part of these efforts. Particularly during the past decade, policymakers and program administrators have focused on devising strategies to enhance the quality of Head Start services that can improve children's readiness to enter kindergarten.

A focus on program improvement and the development of innovative strategies for meeting the needs of children and families has occurred throughout the Head Start community, on both a large and a small scale. During the past several years, for example, the Head Start Bureau has initiated several large-scale quality enhancement efforts, including a series of national teacher training conferences on the subject of fostering early literacy development, a technical assistance initiative to help programs to implement mentor-coaching strategies designed to support teachers in the classroom, and regional training conferences on fostering children's social-emotional development and on addressing difficult

behaviors. Regional and state collaboration offices of the Administration for Children and Families (ACF) have launched statewide efforts to improve children's access to health care, often through the development of partnerships with health care providers and dentists. On a smaller scale, individual Head Start programs have implemented new curricula, services to address the needs of English-language learners, and strategies for using child assessment data to improve services.

Despite this significant focus on quality improvement and the wide range of enhancement ideas that have been developed, little is known about the effectiveness of these national, state, regional, and locally initiated strategies. To more fully realize Head Start's efforts to improve program quality and children's readiness for school, the effectiveness of these quality enhancement strategies should be rigorously evaluated.

The Advisory Panel for the Head Start Evaluation Design Project recommended the following question as one of the central ones for research: "Which Head Start practices maximize benefits for children and families with different characteristics under what types of circumstances?" (U.S. Department of Health and Human Services 1990). This question would naturally call for a research design that compares the effectiveness of different Head Start programs that use different practices, and that have different characteristics. More recently, the Advisory Committee on Head Start Research and Evaluation considered the possibility that studying promising approaches in Head Start (by comparing some Head Start programs with others) might be a valuable complement to studying the contribution of Head Start generally, as the current Head Start Impact Study is doing by comparing children randomly assigned to participate or not participate in Head Start (Advisory Committee on Head Start Research and Evaluation 1999).

A study design that rigorously tests enhancements against current practice and against each other would provide valuable information about whether these enhancement initiatives improve children's readiness for school beyond the benefit that may be gained from participation in typical Head Start services. The Head Start community also could learn more about how to invest resources strategically in quality enhancement activities by testing enhancement ideas that vary according to the intensity and duration of teacher training, ongoing technical assistance, or services provided to children and families. Knowledge gained from these studies could help the Head Start Bureau to decide whether to implement system-wide quality improvement efforts, determine where and how to target technical assistance, and determine which resources are required to achieve the desired results.

In recognition of the need for rigorous evaluation of Head Start quality enhancements, ACF's Office of Planning, Research, and Evaluation contracted with Mathematica Policy Research, Inc. and its subcontractor, Xtria, LLC, to prepare design options for potential future evaluations. The purpose of the project, known as the Design Options for the Assessment of Head Start Quality Enhancements project, is to develop a framework for rigorously evaluating enhancement initiatives and program variations in Head Start. This framework is intended to provide flexible guidelines for designing research to evaluate the

effectiveness of specific quality enhancements in Head Start across a range of different settings, on either a small or a large scale.[1] Previous concept papers developed for this project have focused on key issues in developing research designs, implementing enhancements and measuring implementation and fidelity, and measuring intermediate and child outcomes (Ross et al. 2004; Paulsell et al. 2004; Boller et al. 2004). This final report identifies specific quality enhancement ideas and describes appropriate evaluation plans for each.

Evaluations of quality enhancement ideas in Head Start will be based on a comparison of the outcomes of children in programs that have implemented the quality enhancement with the outcomes of children in regular Head Start programs, or with those of children in programs that implemented different quality enhancements. In most cases, rigorous evaluations of quality enhancement strategies in Head Start will rely on random assignment of groups of children (classrooms, entire centers, or entire grantees), rather than on the random assignment of individual children, which has been the traditional approach in program evaluations.[2] Because the outcomes of children within the same group generally are correlated, random assignment of groups of children generates design effects that considerably reduce the precision of the impact estimates relative to the random assignment of individual children. As we discuss in this report, the greater imprecision of impact estimates resulting from the random assignment of classrooms, centers, or grantees necessitates increases in the size of the samples of children for an evaluation.[3] Accordingly, the evaluation may have to include more children, classrooms, and centers than are traditionally included as part of program evaluations. As a result, the costs of evaluating quality enhancement strategies relative to traditional program evaluations will generally increase as well.

A. POTENTIAL QUALITY ENHANCEMENT IDEAS

Many ideas for improving Head Start practices are currently under development by Head Start programs, by program-researcher partnerships, through federal initiatives, and through a variety of other projects. For example, some Head Start programs regularly conduct self-assessments to identify areas for improvement, and to develop strategies to make those improvements. For example, the programs might consider improved

[1] Head Start programs serving preschool-age children are the focus of this research design effort; extensions of this framework to Early Head Start programs are likely to be straightforward but are not part of the current design effort. In addition, because Migrant/Seasonal Head Start programs pose unique design issues currently under consideration as part of a separate research design project, we do not address them as part of this research framework.

[2] Random assignment is not the only approach to evaluation; for most questions, however, it is the most rigorous one. Shadish et al. (2002) discuss this issue in greater detail. We discuss a variety of designs in this report, but most involve random assignment of groups.

[3] This assumes researchers have selected a target minimum detectable effect size, or the minimum impact as a percentage of the standard deviation of the outcome measure that can be detected under the sampling and random assignment design.

management approaches, specific teacher training, or adoption of a specialized curriculum to enhance children's performance. Several Head Start programs are working with university or research partners to implement and test strategies to improve program practices. The Head Start-University partnerships that were funded through 2004 included five universities and their Head Start partners investigating strategies to promote language development, pro-social behavior, relationships, and school readiness. The Head Start Quality Research Center (QRC) Consortium includes eight university or private non-profit research teams and their Head Start program partners investigating the effects of interventions in the areas of early literacy, social-emotional development, and other domains of school readiness. Intervention approaches include enhancements to curriculum, teacher training and mentoring, parent involvement, and assessment practices. Federal initiatives include the Strategic Teacher Education Program (STEP) training to enhance early language and literacy practices in the classroom (provided to all Head Start teachers in 2002-2003) and Parent Mentor Training to develop parent leaders who could encourage other parents to read to their children, and to promote early language and literacy in the home (offered in 2004). Local Head Start programs and local organizations that partner with Head Start programs are using Innovation and Improvement grants, awarded by the Head Start Bureau in 2003, to develop ideas that will enhance early literacy, strengthen families and fatherhood, support children's mental health and social-emotional development, promote youth development, promote coordinated social services within communities, and improve technical assistance to programs.

As part of the process of developing the research designs and implementation plans for this project, staff conducted telephone discussions or held in-person meetings with a range of stakeholders in the Head Start community to identify innovative practices and ideas that merit evaluation.[4] The group of informants included federal Head Start Bureau staff, representatives of the 10 geographic ACF regions, local Head Start program staff, state Head Start collaboration officials, staff of the National Head Start Association, selected principal investigators in the Head Start QRC Consortium, and other researchers. These informants identified several areas in need of quality enhancement, including program management and leadership, access to child health care services, staff recruitment and retention, parents' involvement, the use of data to improve programs, and strategies to address challenging behaviors in the classroom.

These discussions and a review of Head Start quality initiatives and research suggest that enhancement ideas in the following areas are the most prominent:

- **Enhancing Children's Language and Emergent Literacy Skills.** Curricula or classroom interventions to promote children's early literacy, including two that focus specifically on the needs of English-language learners, represented more than half of the enhancements identified.

[4] Additional information about the discussions and the information obtained from them can be found in Paulsell et al. (2004).

- **Program Management and Continuous Program Improvement.** Several enhancements had the goal of improving program management, either through training for program directors or strategies for using data to support continuous program improvement.
- **Supporting Children's Social-Emotional Development.** Some enhancements were designed to foster children's social-emotional development and to support teachers in fostering positive classroom environments by helping to address challenging classroom behaviors.
- **Increasing Access to Health Services.** Two of the suggested enhancements were designed to increase families' access to health and dental health services for their young children.

These ideas were most often mentioned as perceived quality enhancement needs, are currently under development, and, in some cases, are being evaluated in Head Start programs. Even so, as policies change and as more is learned about Head Start programs, many other quality enhancements will be considered and may even become more prominent. Accordingly, the research ideas discussed in this report are based in part on this list but also have been chosen to illustrate a range of research approaches so that they provide examples that will remain relevant even as the most important quality enhancement ideas change over time.

B. CRITERIA FOR ENHANCEMENTS

Given the large number of quality enhancement ideas that might be examined, some criteria are necessary to focus the evaluation efforts on enhancements that are the most promising in terms of both their appeal to the Head Start community and their readiness for evaluation. We propose five criteria to select a more focused set of enhancements for consideration:

1. **Relevance.** A theory has been developed to explain how and why the enhancement is expected to affect outcomes critical to children's development, some evidence links the enhancement strategy to children's outcomes, and the problem that the enhancement addresses is of particular importance to Head Start staff or Head Start children.
2. **Distinctiveness.** The enhancement is measurably different in key program dimensions from current Head Start practice.
3. **Clarity.** The enhancement as it is fully implemented is clearly described, and the activities required to fully implement the quality enhancement strategy are clearly documented.
4. **Replicability.** The enhancement can be implemented to a high degree of fidelity in programs beyond the original sites and on a larger scale.
5. **Feasibility.** The enhancement can be implemented and rigorously evaluated.

Meeting the relevance and distinctiveness criteria are prerequisites for considering a quality enhancement for evaluation. If an enhancement idea meets these criteria, then we have reason to believe that child outcomes will be influenced to a meaningful degree. Relevant enhancements not only address a developmental domain or program process of central importance to the Head Start community but also are based on a strong intervention theory of change. In other words, evidence suggests that the enhancement strategy can affect the outcomes that are the target of the intervention, and that the enhancement will be delivered with sufficient intensity to produce meaningful change during the Head Start program year.

To assess distinctiveness, careful consideration must be given not only to the strength of the enhancement itself, but also to system-wide training and technical assistance efforts that are under way or in the planning stage that will affect the level of typical Head Start services at the time the evaluation is launched. For example, the Head Start Bureau has invested considerable resources in providing training to all programs about techniques to promote early literacy, and it plans to launch a new, nationwide training and technical assistance effort in the near future. Any early literacy enhancement would have to be distinct from program services that already have been strengthened through these efforts.

In addition to being relevant and distinctive, enhancements must be clear or well documented. Documentation must include clear information about how the fully implemented enhancement changes practice (documented through photographs, videos, or very precise explanations). It also must include implementation manuals, plans for classroom activities, parent-training materials, teacher-training protocols, and other materials that provide the foundation necessary to implement the quality enhancement with high fidelity to the enhancement model in a large number of Head Start programs. Documentation also should include measures of fidelity, so that the progress of implementing the quality enhancement can be monitored. Finally, documentation should include both expected interim outcomes (changes in teacher behaviors and beliefs and in observed classroom practice), and in children's outcomes.

Enhancements that prove replicable in a small-scale study may need further refinement if they are to be replicable on a larger scale, particularly if national implementation is anticipated. Procedures, manuals, and training protocols may have to be adjusted or refined in order to ensure that an enhancement can be replicated well in a broad range of Head Start programs, and that it can be supported by the Head Start training and technical assistance system.

Feasible enhancements are those that can be implemented in a way that provide opportunities for rigorous evaluation. For example, it should be possible to implement the intervention without its effects "spilling over" to other programs or classrooms that are serving as the counterfactual, and thus are not supposed to implement the enhancement. Adjustments to implementation or evaluation designs can be made to address this issue, such as randomization carried out at the center rather than the classroom level.

C. A STAGED APPROACH TO EVALUATION

In light of both the need for larger evaluations that are potentially more expensive to deploy and the concomitant need for enhancements to meet the five evaluation criteria before large-scale evaluation is attempted, previous concept papers have described a staged evaluation approach. Such an approach first builds evidence of implementation feasibility, then of positive impacts in a best-case scenario, and finally of the strength of the idea in a broader field test. The approach thus includes the following three stages, described further in this section: (1) development (early implementation and documentation), (2) evaluation, and (3) field testing.

Development (Stage 1)

Many quality enhancement ideas could benefit from additional development during a pilot phase to ensure that implementation is well documented and well understood before the interventions are evaluated. In some cases, an enhancement may already be well developed, and lessons learned from previous implementation experiences may make this initial stage unnecessary. Nevertheless, for enhancements that are in need of further development, a pilot phase—most likely, lasting for one year—would be an important first step in the evaluation process.

During the development phase, an enhancement would be implemented in several programs. The goals of this phase would be to ensure that the theory of change underlying the intervention is well understood, implementation procedures are fine-tuned, manuals and other documentation are developed, and measures of fidelity are established. The development phase also would provide an opportunity to determine whether the enhancement can be implemented successfully with a reasonable level of resources.

Given the diversity of Head Start programs and children, the quality enhancement might have to be tailored to the specific circumstances of particular programs while maintaining the essential elements of the enhancement model. For example, the model may have to be altered due to constraints on physical space or because the training materials need modification to be more culturally appropriate. Because some amount of tailoring across Head Start settings is likely to be necessary, a critical task during the development phase is to identify the aspects of the enhancement's implementation that are necessary to maintain a high degree of fidelity to the enhancement model. Measures of fidelity to the enhancement model should be developed, and teachers and programs should be given feedback based on classroom observations. Measures of key outcomes of the enhancement might have to be refined, or developed and pilot tested.

Evaluation (Stage 2)

After the enhancement has been fully developed and the techniques for implementing it with a high degree of fidelity are well understood, it can be implemented as part of a rigorous evaluation. The evaluation phase would measure effectiveness under favorable conditions (for example, in programs that volunteer for the study). Some enhancement strategies, such as the ones that the Head Start QRC Consortium is studying, may already

have been evaluated on a small scale. Others have not been rigorously evaluated. Before investing the resources required for a full-scale field test, these enhancements should first be evaluated on a small scale to determine whether they show sufficient promise. Cost considerations and practicality suggest that an evaluation at this stage would likely focus on a few ACF regions or on other limited geographic areas, and, within those areas, on Head Start programs that volunteer to participate. However, ideally, programs selected for participation at this stage would encompass diversity in key programmatic characteristics such as part- or full-day schedules, and in key population characteristics, such as ethnicity and home language.

The evaluation schedule should allow time to implement the enhancement before children in the research sample begin their program year. For example, the enhancement could be implemented at some point during a program year, but not evaluated until the following program year. In addition to measuring impacts on children's outcomes, an important component of the evaluation is the assessment of the quality of implementation, and of the degree of fidelity to the enhancement model programs. Knowing that the enhancement was implemented and the degree to which implementation in particular sites was true to the model will be critical for understanding impacts, and, if the enhancement has positive impacts on children, for understanding how the initiative can be replicated more broadly. If the evaluation indicates that the initiative is well implemented and has favorable impacts on children, then planners can consider either broadening the evaluation to include more programs that are willing to participate or moving to a field test of the enhancement.

Field Test (Stage 3)

Ultimately, enhancement ideas should be field tested in a representative sample of Head Start programs to determine whether the enhancement continues to be effective when implemented on a wide scale across a variety of Head Start programs. Support for implementing quality enhancements strategically (to permit evaluation) and on a wide scale could be provided through the Head Start technical assistance and training system at a level of intensity that is typical for broad, new initiatives. A field test would be most appropriate for an initiative that is under consideration for full-program implementation, regardless of whether the initiative has passed through the development and evaluation stages.

The Value of a Staged Approach to Evaluation

This staged approach to evaluation is one that allows enhancements to be well developed and well implemented prior to evaluation, and that allows for the demonstration of evidence of effectiveness before investing resources in a full-scale field test. Nevertheless, following these steps sequentially would take considerable time, which is not always necessary or feasible. For example, because potential quality enhancements are at various stages of development, the lessons learned from prior implementation experiences or from small-scale evaluation efforts may preclude the need for a development phase or even smaller-scale evaluation. Moreover, in some cases, it may not be possible to implement a staged approach to evaluation, as policy mandates may require more immediate change in program practices, teacher qualifications, or other aspects of Head Start.

Yet, despite the time and expense associated with the staged approach, the steps are often necessary for any particular enhancement idea in order to avoid implementation disasters and to gain the confidence of skeptics. In fact, as we researched enhancement ideas, we found that those achieving public recognition had gone through a development phase and often, a small-scale evaluation, although in many cases, the evaluations were not rigorous. The development and evaluation activities had been conducted over many years, and often we found that ideas were abandoned before they had fully emerged from this process.

A more systematic and planful approach to incubating new ideas and evaluating them using rigorous designs could put public research and development resources to more efficient use and in some cases, reduce the amount of elapsed time to develop and evaluate new ideas. The staged approach we have described provides a useful framework for learning about the effectiveness of enhancement ideas at different stages of development. Fully developing enhancements in a pilot phase (by setting expectations for documentation to be produced at the end of an implementation period) and launching small-scale, rigorous evaluations to identify promising practices can provide program administrators with a pool of potentially valuable ideas to be considered when broad policy mandates require a prompt response. A system for reporting and maintaining an archive of reports and other information about these ideas is critical to ensure that ideas progress through the evaluation stages, and that information is available when broad policy mandates are under consideration.

D. PLAN FOR THE REPORT

This report describes the goals and activities associated with each of the three stages of research through the use of specific examples of potential Head Start quality enhancements that are good illustrations of the particular stage of research. To provide a context for the discussion of research designs in this report, Chapter II describes the Head Start program and research. Subsequent chapters are organized according to the three stages of research described here. Chapter III discusses the development phase and the information that should emerge from early efforts to establish the essential features of a quality enhancement idea and the requirements for implementing the initiative. Chapters IV through VI describe research designs for specific quality enhancement initiatives that are ready for evaluation. Chapter VII describes field testing of broad program initiatives. Our selection of quality enhancement ideas to be discussed in this report was influenced in part by the ideas that seem to be of most interest to the Head Start community, with the addition of one area—early mathematics—that was not discussed, but that is likely to become more prominent after programs have reviewed the results of the annual child assessments conducted as part of the Head Start National Reporting System.

CHAPTER II

THE HEAD START PROGRAM AND RELATED RESEARCH

Local agencies operate thousands of Head Start programs in communities across the country. Although all of the programs are required to meet the Head Start program performance standards, services are tailored to the needs of the local communities. To provide context for the evaluation design issues discussed in this report, we begin by describing the major characteristics and features of Head Start programs, highlighting the ones that are common to most programs, as well as the ones that generally vary among programs. We then describe the diverse demographic characteristics of the children and families served.

Research has been an integral part of Head Start over the years. Recent studies have examined several questions in an effort to understand better which Head Start practices work best for children. We discuss the goals and approaches of these in order to identify the unique contribution that the current study will make to the recent Head Start research efforts. We also discuss the Head Start Planned Variation study that was conducted from 1967 to 1977 and the lessons from that effort that inform our design today.

A. OVERVIEW OF THE HEAD START PROGRAM

Since 1965, the Head Start program has provided comprehensive child development services to approximately 21 million low-income children and their families. In an effort to break the cycle of poverty the Office of Economic Opportunity launched Head Start as an eight-week summer program for children about to enter kindergarten. Head Start soon expanded to offer comprehensive services to four-year-old children in half-day, center-based programs that operated during the school year.

Head Start now is the largest federally funded early childhood program. In fiscal year (FY) 2004, the Administration for Children and Families (ACF) provided $6.6 billion to local grantees across the country to offer comprehensive child development services,

predominantly to three- and four-year-old children.[1] Grantees operated approximately 18,269 centers and served 976,013 children and their families.[2] The program provides half-day and full-day program options, two-year programs (for three- and four-year-old children), and other service variations for children in rural areas. Early Head Start, begun in 1995, extends Head Start services to younger children by providing child development and family development services to pregnant women and to children up to age three in home-based and/or center-based settings. In 2003, Early Head Start programs served approximately 71,000 children and 9,500 pregnant women.

Head Start is administered by the Head Start Bureau, which is part of the Administration on Children, Youth, and Families, ACF, U.S. Department of Health and Human Services (DHHS). Grants to operate Head Start programs at the community level are awarded directly from the federal government to community action agencies, public or private school systems, nonprofit organizations, local government agencies, private or public for-profit agencies, and Tribal or Alaska Native government organizations. Staff in the 12 regional ACF offices work closely with Head Start Bureau staff and serve as a direct link to local grantee programs.[3] The regional offices monitor the operation of local programs.

The Head Start program offers a comprehensive set of services to low-income children and their families that focus on preparing economically disadvantaged children for school. The services fall into two main domains:

1. ***Early childhood development and health services*** cover the areas of child health and development, education and early childhood development, child health and safety, child nutrition, and child mental health.
2. ***Family and community partnerships*** include family goal-setting, access to community services and resources, services for pregnant women, involvement of parents, formation of partnerships in the community, advisory committees, and transition services.

Head Start programs are charged with providing services that promote children's development and school readiness in culturally appropriate ways that meet the needs of the children and their families. Grantees must conduct periodic community needs surveys to update their knowledge about the needs of families in the community. In the following section, we discuss the diversity of Head Start program models and characteristics and then describe the diversity of families served by Head Start programs. This discussion is based on

[1] Budget figure represents appropriations for local Head Start projects in fiscal year 2004 (U.S. Department of Health and Human Services 2004a).

[2] Information about the numbers and characteristics of children served and about programs in 2002-2003 is based on analyses of the Head Start Program Information Report (PIR) data submitted by 1,969 Head Start programs and processed by Xtria, LLC, on February 12, 2004.

[3] Regional ACF offices include 10 geographically based offices and 2 additional offices working with American Indian and Alaska Native programs and with Migrant and Seasonal Head Start programs.

analyses of data from the Head Start PIR for the 2002-2003 program year. PIR data are collected annually from Head Start grantees.

B. CHARACTERISTICS OF HEAD START PROGRAMS

A Head Start grantee is a public or private agency or organization that receives funds to provide Head Start services, and that is responsible for meeting the conditions of the grant. Grantees may select delegate agencies to help provide Head Start services. A delegate agency is a public or private nonprofit organization or agency to which a grantee has delegated, by written agreement, responsibility for carrying out all or part of its Head Start program. A Head Start program can be either a grantee or a delegate agency. Many grantees (65 percent of all programs operating during the 2002-2003 program year) directly operate Head Start programs without any delegate agencies. A few grantees (five percent of all programs) both directly operate programs and have delegate agencies that operate programs. A very small proportion of grantees (one percent of all programs) maintains only central office staff and delegates all program operations. In total, 28 percent of all Head Start programs in the 2002-2003 program year were delegate agencies.[4] When working with Head Start programs, researchers will have to approach grantees and, if applicable, delegate agencies to obtain permission to conduct the evaluation, and to obtain contact information about center directors.

Private or public nonprofit agencies and community action agencies were the most common types of agencies running Head Start programs during the 2002-2003 program year, operating 35 percent and 32 percent of all Head Start programs, respectively. School systems ran a substantial 20 percent of all Head Start programs.[5] Head Start programs operated in 18,269 centers, which ran 44,656 classes. In addition, Head Start programs operated with 3,096 child care partnerships, 1,374 family child care homes, and 3,824 home-based groups.

The program options and intensity of services offered to families by Head Start programs are driven largely by the programs' community resource and needs assessments. In recent years, welfare-related work requirements have led to an increasing need for child care while parents work. ACF has recognized this need by increasing funding in FY 2000 for full-day, full-year services designed to help working families that are moving from welfare to work. One of the ways in which programs provide full-day, full-year services to their enrolled children is by partnering with other child care centers or family child care homes. Services provided by the child care partners must meet the Head Start performance standards, including requirements for maximum child-adult ratios and the quality of the educational program.

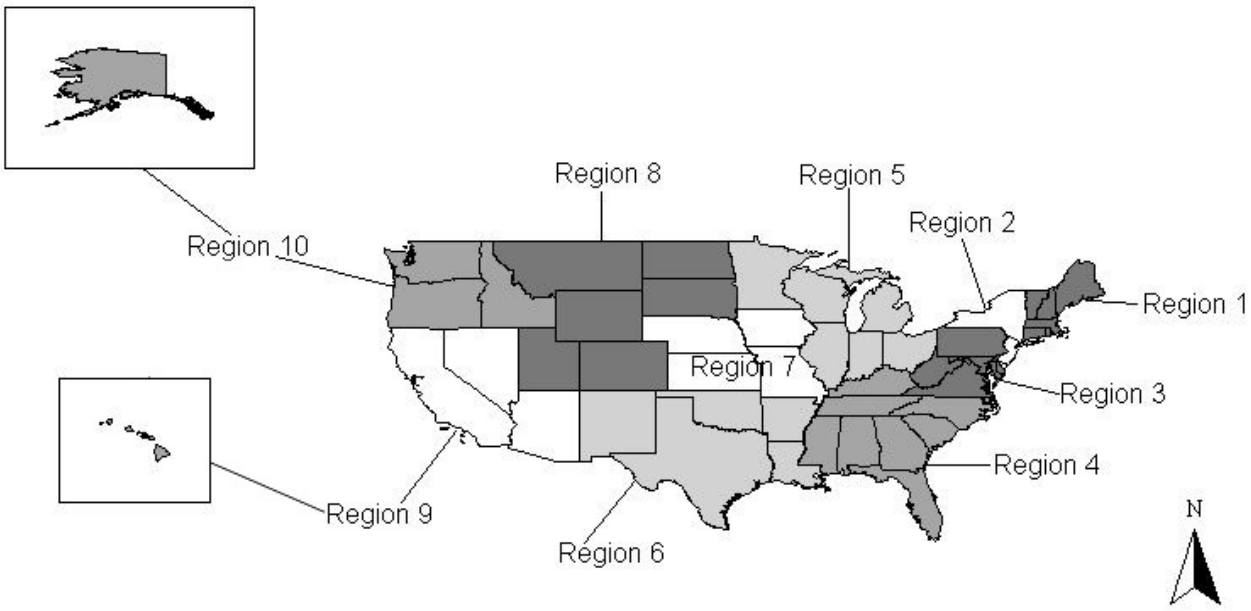
[4] Approximately two percent of programs were grantees that delegated all programs and maintained no central office staff.

[5] Approximately 13 percent of Head Start programs were run by private or public for-profit agencies, government agencies, or tribal governments or consortia.

During the 2002-2003 program year, nearly all Head Start children (93 percent) were enrolled in a part- or full-day center-based program. Almost half the children (49 percent) were enrolled in the full-day option (four or five days per week); 45 percent were enrolled in the part-day option (also four or five days per week); 3 percent were in the home-based option; and the remaining 4 percent were in combination, family child care, and locally designed options. The proportion of children served in a full-time, center-based program has increased substantially since the 1996-1997 program year, when only 24 percent of Head Start children received Head Start services in full-day, full-week, center-based care settings (Schumacher and Rakpraja 2003). Part-day programs offer a center-based program for fewer than six hours per day. Some of these programs are "double-session" classes in which a teacher provides center-based services to two different groups of children by offering morning and afternoon sessions.

Head Start programs across the country differ in important ways, reflecting regional and state differences in demographics, economic well-being, the states' early childhood program contexts, and other characteristics. To some degree, we can illustrate the diversity of Head Start programs by comparing the characteristics of programs from different regions. Information about Head Start programs, by region, also is useful because evaluations of quality enhancement options are likely to be conducted in a few sites with sufficient geographic concentration to enable technical assistance staff to reach them easily. The characteristics of Head Start programs in specific regions might indicate that the region is a good (or a more challenging) place to try a particular innovative idea. To simplify the tables, we have collapsed the 10 geographically based regions (see Figure II.1) to form five regional

Figure II.1. ACF Regions and Regional Offices



Source: ACF Office of Regional Operations (http://www.acf.hhs.gov/programs/oro)

groups: (1) Northeast (Regions 1, 2, and 3); (2) the South (Regions 4 and 6); (3) North Central (Region 5); (4) Mountain/Plains (Regions 7 and 8); and (5) the West (Regions 9 and 10). We also show separately the Migrant/Seasonal and American Indian/Alaska Native programs, which may be located anywhere in the country.

Table II.1 indicates that the mix of program options varies substantially across regions. During the 2002-2003 program year, full-day, full-week, center-based programs predominated in the South (where center-based child care is more available than in other regions and tends to be chosen more often than other forms of care) and among Migrant/Seasonal programs (because parents typically work full-time when seasonal jobs are available). Full-day, center-based care programs constituted half the program options offered in the Northeast. Half-day, center-based programs operating four or five days per week were the most common program options in the North Central, Mountain/Plains, and West regions, and in the American Indian/Alaska Native programs. The least common program options across all regions were the full-day, center-based programs operating four days per week and the home-based, combination, and family child care programs.

Table II.1. Percentage of Children Enrolled in Each Program Option, 2002-2003 Program Year

Program Option	All	Northeast (Regions 1, 2, and 3)	South (Regions 4 and 6)	North Central (Region 5)	Mountain and Plains (Regions 7 and 8)	West (Regions 9 and 10)	Migrant and Seasonal Programs	American Indian and Alaska Native Programs
Center-based: Five days/week, full-day	47	51	72	24	22	19	96	22
Center-based: Five days/week, half-day	22	28	20	9	11	42	1	13
Center-based: Four days/week, full-day	2	2	2	4	4	1	0	10
Center-based: Four days/week, half-day	23	13	4	54	53	27	0	47
Home-based	3	3	1	4	5	4	0	5
Combination	1	2	0	2	2	3	0	2
Family child care	1	1	0	1	1	1	3	0
Locally designed options	2	1	1	2	3	3	0	1

Source: Calculations based on 2002-2003 Head Start PIR data submitted by 1,969 Head Start programs and processed by Xtria, LLC, on February 12, 2004. Early Head Start programs were excluded from these calculations.

The number of children per adult in Head Start programs was low relative to other early childhood settings. Head Start Program Performance Standards require a ratio of 1 adult caring for no more than 7.5 to 8.5 children aged three years, and no more than 10 children aged four years. This ratio is comparable to the National Association for the Education of Young Children accreditation standards of 8 or fewer three-year-olds or 10 or fewer four-year-olds for each adult. The PIR data indicate that, on average, 7.8 children were cared for

by each adult. Table II.2 indicates that the numbers of children per adult in the South and West were somewhat higher than the national average, but still within the limits set by the performance standards. The number of children per adult was lower in Migrant/Seasonal Head Start programs, which serve children younger than age three as well as preschool-age children.

Table II.2. Number of Children per Adult in Head Start Classrooms, 2002-2003 Program Year

All	Northeast (Regions 1, 2, and 3)	South (Regions 4 and 6)	North Central (Region 5)	Mountain and Plains (Regions 7 and 8)	West (Regions 9 and 10)	Migrant and Seasonal Programs	American Indian and Alaska Native Programs
7.8	7.7	8.5	7.6	7.0	8.2	5.1	7.0

Source: Calculations based on 2002-03 Head Start PIR data submitted by 1,969 Head Start programs and processed by Xtria, LLC, on February 12, 2004. Early Head Start programs were excluded from these calculations.

Note: The child-to-adult ratio is calculated as the total funded enrollment (excluding home-based and family day care enrollment) subtracting one-half of the children enrolled in double sessions (to ensure that children are counted only once) and divided by the total number of classroom staff (teachers and assistant teachers).

Teachers' education levels have increased in recent years in response to a requirement of the 1998 Head Start reauthorization law that, by 2003, at least half of all Head Start teachers in center-based programs must have an Associate in Arts degree (A.A.), a Bachelor of Arts degree (B.A.), or advanced degree in early childhood education (ECE) or in a related field. Table II.3 shows that 59 percent of the teachers had an associate's, bachelor's, or graduate degree in the 2002-2003 program year, thereby exceeding the teacher qualification mandate. A larger proportion of teachers in the Northeast relative to other regions had a bachelor's or graduate degree.

The availability of direct child development staff (teachers, assistant teachers, home visitors, and family child care providers) who speak languages other than English to meet the needs of the linguistically diverse Head Start population is particularly significant. Overall, 28 percent of direct child development staff across all Head Start programs spoke languages other than English (see Table II.3). Migrant and American Indian/Alaska Native programs and programs in the West and Northeast far exceeded the national percentage in this respect, consistent with the substantial proportion of families in those regions who speak languages other than English. (This issue is discussed in greater detail in the next section.)

Table II.3. Percentage of Teachers by Education Levels and Languages Spoken, 2002-2003 Program Year

Education Levels/ Languages Spoken	All	Northeast (Regions 1, 2, and 3)	South (Regions 4 and 6)	North Central (Region 5)	Mountain and Plains (Regions 7 and 8)	West (Regions 9 and 10)	Migrant and Seasonal Programs	American Indian and Alaska Native Programs
Child Development Associate credential/ state equivalent	26	15	35	20	25	21	42	38
Associate's degree, ECE/related	27	20	28	33	22	40	15	21
Bachelor's degree, ECE/related	28	47	23	29	35	19	11	9
Graduate degree, ECE/related	4	10	3	4	4	2	1	0
Child development staff speaking languages other than English	28	32	17	10	12	49	65	38

Source: Calculations based on 2002-2003 Head Start PIR data submitted by 1,969 Head Start programs and processed by Xtria, LLC, on February 12, 2004. Early Head Start programs were excluded from these calculations.

ECE = early childhood education.

C. CHARACTERISTICS OF HEAD START CHILDREN AND FAMILIES

The 869,570 Head Start slots funded during the 2002-2003 program year were distributed unevenly across the geographic regions (see Table II.4). The South contained the largest proportion of Head Start-funded slots (nearly one-third of all funded slots), followed by the Northeast and North Central regional groups, each of which had approximately one-fifth of the funded slots. The Migrant/Seasonal programs and the American Indian/Alaska Native programs had the smallest proportions of funded slots.

The number of children enrolled in Head Start (976,013) during the 2002-2003 program year exceeded the number of funded slots because of normal turnover during the program year. Children are counted as enrolled in Head Start if they have attended at least one class session or have received at least one home visit. Head Start staff try to fill the slots of children who leave the program, although they cannot always do so. During the 2002-2003 program year, 16 percent of children who had enrolled left during the program year, and approximately 5 percent of the slots could not be filled. The proportion of enrolled children who leave during the program year is important because it lies in the fact that children in an evaluation study who leave the program during the program year still must be located and assessed at the follow-up points, even though this might be more difficult to do so if the children have left the program. The proportion of children who left Head Start during the 2002-2003 program year was lowest (12 percent to 13 percent) among children in the Northeast and South regional groups and among children in the American Indian/Alaska Native programs. Turnover was higher (17 percent to 23 percent) in the other regional groups.

Table II.4. Head Start Program Funded Enrollment, 2002-2003 Program Year

Enrollment	All	Northeast (Regions 1, 2, and 3)	South (Regions 4 and 6)	North Central (Region 5)	Mountain and Plains (Regions 7 and 8)	West (Regions 9 and 10)	Migrant and Seasonal Programs	American Indian and Alaska Native Programs
Total funded enrollment	869,570	187,296	271,282	159,800	57,420	138,599	33,721	21,452
% funded enrollment, by region	100	22	31	18	7	19	4	3
Total actual enrollment	976,013	205,354	303,534	185,145	65,169	161,393	33,188	22,230
% who left during the program year	16	13	13	18	20	17	23	12
Number of programs	1,969	513	498	337	175	232	65	149

Source: Calculations based on 2002-2003 Head Start PIR data submitted by 1,969 Head Start programs and processed by Xtria, LLC, on February 12, 2004. Early Head Start programs were excluded from these calculations.

Note: Funded enrollment is the number of children that the Head Start program is funded to serve. The actual enrollment, or the number of children enrolled who attended at least one class or who had at least one home visit, is larger than the funded enrollment because of normal attrition and replacement of children during the program year.

To be eligible for Head Start, a child must meet the program's age requirements, and the family's income must be lower than the official poverty guidelines. Up to 10 percent of the children enrolled in a Head Start program may be from families that exceed the poverty guidelines. A child from a family that receives public assistance or a child in foster care is eligible even if the family income exceeds the income guidelines. As shown in Table II.5, the majority of Head Start children and families (73.9 percent) were eligible strictly on income criteria, 18 percent were eligible because they received public assistance, 7 percent were from families that exceeded the income guidelines, and 1 percent were foster children. The

Table II.5. Percentage of Children Enrolled in Head Start, by Type of Eligibility, 2002-2003 Program Year

Eligibility for Head Start	All	Northeast (Regions 1, 2, and 3)	South (Regions 4 and 6)	North Central (Region 5)	Mountain and Plains (Regions 7 and 8)	West (Regions 9 and 10)	Migrant and Seasonal Programs	American Indian and Alaska Native Programs
Receipt of public assistance	18	23	13	22	16	18	2	22
Income-eligible (<100% of the federal poverty line)	74	67	81	69	74	73	93	57
Over-income (>100% of the federal poverty line)[a]	7	8	6	8	8	7	5	19
Foster child status	1	2	1	1	2	2	0[b]	2

Source: Calculations based on 2002-2003 Head Start PIR data submitted by 1,969 Head Start programs and processed by Xtria, LLC, on February 12, 2004. Early Head Start programs were excluded from these calculations.

[a]Up to 10 percent of children enrolled in a Head Start program can be from families with income above 100 percent of the federal poverty line.

[b]One-tenth of one percent of children in Migrant/Seasonal programs have foster child status.

highest percentage of children eligible by the public assistance criterion were from programs in the Northeast and North Central regional groups and from American Indian/Alaska Native programs; in general, however, the regional groups differed little in enrollment by eligibility status. Migrant/Seasonal programs served the highest number of children who met the federal poverty income guidelines.

The majority of children in Head Start were four-year-olds (56 percent during the 2002-2003 program year), and more than one-third (36 percent) were three-year-olds (see Table II.6). This age distribution was similar in all regions except Migrant/Seasonal programs, which serve a somewhat more even distribution of children across the age range from birth to six years.

Seventy percent of the children enrolled in Head Start programs were in their first year of Head Start, and 28 percent were enrolled for a second year (see Table II.6). The composition of children according to the number of years in Head Start was fairly consistent across the geographic regions, but the Migrant/Seasonal programs, which traditionally serve infants, toddlers and preschool-age children, served a substantial percentage of children (14 percent) for a third year or longer.

Table II.6. Percentage of Children Enrolled in Head Start, by Age and Years of Enrollment, 2002-2003 Program Year

Characteristic	All	Northeast (Regions 1, 2, and 3)	South (Regions 4 and 6)	North Central (Region 5)	Mountain and Plains (Regions 7 and 8)	West (Regions 9 and 10)	Migrant and Seasonal Programs	American Indian and Alaska Native Programs
Age								
Younger than one year	0	0	0	0	0	0	12	0
One year	1	0	0	0	0	0	15	0
Two years	1	1	0	2	1	1	18	1
Three years	36	36	39	38	38	32	23	37
Four years	56	58	57	53	58	64	21	49
Five years	5	5	4	7	3	4	11	14
Years in Head Start								
One	70	70	70	68	72	75	58	62
Two	28	28	30	30	27	24	28	34
Three or more	2	2	1	2	2	2	14	4

Source: Calculations based on 2002-2003 Head Start PIR data submitted by 1,969 Head Start programs and processed by Xtria, LLC, on February 12, 2004. Early Head Start programs were excluded from these calculations.

Slightly more than 10 percent of children enrolled in Head Start had a diagnosed disability. The majority of these children had speech and language disorders. A large proportion of those with disabilities were classified as having non-categorical/developmental delays. The next most common disabilities reported were health impairments, emotional/behavioral disorders, and multiple disabilities.

Children enrolled in Head Start programs were ethnically diverse, with fairly equal proportions of black or African American children (32 percent), Hispanic or Latino children (31 percent), and white children (27 percent) enrolled. Children's ethnic backgrounds varied substantially by geographic region (see Table II.7). The Northeast's programs served the three major ethnic groups in fairly even proportions. Black or African American children predominated in the South and North Central regional groups; white children predominated in the Mountain/Plains regional group; and Hispanic or Latino children predominated in the West.

Table II.7. Percentage of Children Enrolled in Head Start by Ethnicity and Primary Home Language, 2002-2003 Program Year

Characteristic	All	Northeast (Regions 1, 2, and 3)	South (Regions 4 and 6)	North Central (Region 5)	Mountain and Plains (Regions 7 and 8)	West (Regions 9 and 10)	Migrant and Seasonal Programs	American Indian and Alaska Native Programs
Ethnicity								
Black or African American	32	30	49	39	17	11	1	1
Hispanic or Latino origin	31	35	23	13	22	56	97	2
White	27	28	23	40	52	17	1	8
American Indian or Alaska Native	3	1	1	1	3	2	0	86
Native Hawaiian/other Pacific Islander	1	1	0	0	0	5	0	1
Asian	2	2	1	2	1	5	0	0
Biracial/multiracial	3	3	2	4	5	4	0	2
Other/unspecified	1	1	1	1	1	2	1	1
Primary Language Spoken at Home								
English	73	63	86	87	84	49	9	92
Spanish	23	31	12	9	13	42	86	1
Native Central American, South American, and Mexican	1	1	1	0	0	0	4	0
Caribbean	0	1	1	0	0	0	1	0
Middle Eastern and South Asian	1	1	0	1	0	1	0	0
East Asian	1	1	0	1	1	3	0	0
Native North American/Alaska Native	0	0	0	0	0	0	0	7
Pacific Island	1	0	0	0	0	4	0	0
European and Slavic	1	2	0	0	0	1	0	0
African	0	1	0	1	1	0	0	0
Other	0	0	0	0	0	0	0	0

Source: Calculations based on 2002-2003 Head Start PIR data submitted by 1,969 Head Start programs and processed by Xtria, LLC, on February 12, 2004. Early Head Start programs were excluded from these calculations.

More than one-quarter of Head Start children spoke a language other than English at home. Spanish was the most common non-English language, spoken by 23 percent of the Head Start children at home (see Table II.7). Migrant/Seasonal programs and programs in the West and Northeast had substantial percentages of Spanish-speaking children relative to the national average. The number of children and families speaking non-English languages in Head Start has implications for the hiring of bilingual/multilingual program staff, the program materials printed for families, translation needs at policy council and other parent-staff meetings, and the child assessments conducted by programs as part of their self-monitoring.

Slightly more than half of all Head Start families (56 percent) were headed by single parents (see Table II.8). Single-parent families were the predominant type in the Northeast, South, and North Central regional groups, comprising 56 percent to 64 percent of all families. In the West and among the Migrant/Seasonal and American Indian/Alaska Native programs, two-parent families were more common than were single-parent families. In the Migrant/Seasonal programs in particular, 80 percent of the families were headed by two parents.

Table II.8. Percentage of Families Enrolled in Head Start, by Household Composition and Employment Status, 2002-2003 Program Year

Characteristic	All	Northeast (Regions 1, 2, and 3)	South (Regions 4 and 6)	North Central (Region 5)	Mountain and Plains (Regions 7 and 8)	West (Regions 9 and 10)	Migrant and Seasonal Programs	American Indian and Alaska Native Programs
Household Composition								
Two-parent families[a]	44	44	37	38	49	58	80	59
Single-parent families	56	56	64	62	52	42	20	41
Employment Status of Two-Parent Families								
Both parents employed	34	31	34	32	35	28	76	42
One parent employed	52	53	54	53	53	59	17	42
Neither parent employed	14	16	13	14	12	14	7	16
Unknown	0	0	0	1	0	0	0	0
Employment Status of Single-Parent Families								
Parent employed	61	58	62	64	60	60	82	58
Parent not employed	39	42	38	36	40	40	18	42
Unknown	0	0	0	1	0	0	0	0

Source: Calculations based on 2002-2003 Head Start PIR data submitted by 1,969 Head Start programs and processed by Xtria, LLC, on February 12, 2004. Early Head Start programs were excluded from these calculations.

[a]Two-parent families may be married or unmarried.

Approximately one-third of two-parent Head Start families were families in which both parents worked; slightly more than half were families in which one of the two parents worked. Neither parent worked in just over one-tenth of the families. Among one-parent Head Start families, 61 percent of the parents worked and 39 percent did not. These proportions were fairly consistent across geographic regions (see Table II.8), although employment rates were higher among parents of children enrolled in Migrant/Seasonal programs than in other programs.

The majority (68 percent) of Head Start parents had at least a high school diploma or GED (high school equivalence certificate), although a substantial proportion (one-third) did not (see Table II.9). The distribution of education levels was similar across most regional groups, with three exceptions. Parents from programs in the West and from the Migrant/Seasonal programs were less likely than parents in the other regional groups to have completed high school or a GED. Nearly 82 percent of parents in the Migrant/Seasonal programs had not completed high school or a GED. In contrast, parents in the American Indian and Alaska Native programs had higher levels of schooling compared with parents in other regions, as 77 percent of the parents had completed high school, a GED, or higher levels of education.

Table II.9. Percentage of Children Enrolled in Head Start by Highest Education Level of Parents, 2002-2003 Program Year

Education Level	All	Northeast (Regions 1, 2, and 3)	South (Regions 4 and 6)	North Central (Region 5)	Mountain and Plains (Regions 7 and 8)	West (Regions 9 and 10)	Migrant and Seasonal Programs	American Indian and Alaska Native Programs
Less than high school graduate	32	27	32	27	29	42	82	23
High school graduate or GED	45	50	46	47	41	36	15	50
Some college, vocational school, or associate's degree	19	18	19	21	25	19	3	22
Bachelor's or advanced degree	4	5	3	4	5	3	0	5

Source: Calculations based on 2002-2003 Head Start PIR data submitted by 1,969 Head Start programs and processed by Xtria, LLC, on February 12, 2004. Early Head Start programs were excluded from these calculations.

D. PROGRAM MONITORING AND OVERSIGHT ACTIVITIES

As funding for Head Start grew throughout the 1990s, policymakers sought not only to expand enrollment, but also to enhance the quality of Head Start services and to increase program accountability (U.S. Department of Health and Human Services 1993). For example, during the 1990s, the Head Start Program Performance Standards, which establish minimum levels of service provision and quality for Head Start programs, were revised and strengthened. Some of the initiatives, for example, the National Reporting System assessments of all four-year-old children in Head Start, have involved collecting data that has the potential to inform ideas for quality enhancements, or that can be used as part of an evaluation. Other initiatives have demonstrated the capacity of the Head Start Bureau to implement quality enhancement ideas on a national scale. In 2002, the Head Start Bureau sponsored the Strategic Teacher Education Program (STEP), an initiative to impart to every Head Start teacher techniques and activities designed to promote children's emergent literacy skills. Head Start's training and technical assistance system has been redesigned; it could play a role in implementing and monitoring initiatives identified for field testing. This section discusses the initiatives and their potential applicability to evaluations of quality enhancement initiatives.

In response to the Government Performance and Results Act (GPRA; P.L. 103-62) and the Head Start Act reauthorization in 1994 (42 USC 9831 et seq.), Head Start also began developing specific performance indicators in 1995. In 1997, the Head Start Bureau launched the Family and Child Experiences Survey (FACES) to collect data on the indicators (described in detail in the next section). These indicators cover all the major Head Start program objectives, including program management and parents' involvement; the provision of educational, health, and nutritional services for children; the ability to link children and families with community services; family economic well-being, health, and parenting; and children's school readiness (U.S. Department of Health and Human Services 1995; 2001; 2003a). As we discuss in Section E, FACES has provided extensive data on a random sample of programs. Although they are purely descriptive, the data provide insights into program characteristics associated with higher classroom quality and more favorable child outcomes. These insights suggest potential ideas for quality enhancements to be developed and tested.

The Coats Human Services Reauthorization Act of 1998 (P.L. 105-285) amended the Head Start Act with new provisions designed to further strengthen the program's quality and accountability. The new law required the Secretary of Health and Human Services to establish new educational performance measures related to school readiness, and to ensure that the measures were "results-based" and adaptable for use in peer review and program evaluation. Moreover, program monitoring was to be expanded to include determining whether programs are meeting results-based performance measures.

To guide the development of program assessments of children, the Head Start Bureau convened the Technical Work Group on the Assessment of Program Outcomes, which advised ACF on the development of the Head Start Child Outcomes Framework (U.S. Department of Health and Human Services 2003b). This framework consists of eight

general domains of early learning and child development, 27 domain elements, and examples of specific indicators of children's skills, abilities, knowledge, and behavior—including the four specific congressionally mandated domains and nine mandated indicators of early literacy, language, and mathematics skills (see Appendix A). Since 2003, each Head Start program has been required to assess children throughout the Head Start program year on the congressionally mandated domain elements and on some aspect of each of the eight domains of child learning and development (U.S. Department of Health and Human Services 2000a and 2000b). Programs select their own measures for conducting these assessments. Programs use this information on child outcomes in their self-assessment and continuous program improvement activities. Although the types of assessments that programs choose typically vary, some states (for example, California) have begun to require programs to use specific measures, which could help to make programs a source of outcome data for evaluations.

The 1998 Head Start reauthorization also required that, by September 2003, at least 50 percent of Head Start teachers in center-based programs have an associate's, bachelor's, or advanced degree in ECE, child development, or a related field, as well as experience teaching preschool children. A General Accounting Office (GAO) study (2003) found that, based on Head Start PIR data, at the end of the 2002 program year, 52 percent of Head Start teachers nationwide had at least an associate's degree in ECE or a related field. GAO could not determine whether each classroom had a teacher with minimum credentials in 2002; on average, however, the percentage of teachers with minimum credentials across the 12 ACF regions increased by 14 percentage points during the period from 1999 to 2002.

In recent years, the Head Start Bureau has used several key tools to monitor programs' compliance with the program performance standards, help programs to meet the standards, and help them to improve the quality of services. All Head Start programs participate in an on-site monitoring review at least every three years to assess compliance with the program performance standards. Monitoring teams use the Program Review Instrument for Systems Monitoring (PRISM), are led by staff from the ACF regional offices, and consist of peer and consultant reviewers with relevant expertise. The current PRISM form includes very few detailed, well-defined data items. Moreover, because of the three-year span between assessments, this data collection effort is less helpful for evaluation work than are other Head Start data sources.

In addition to monitoring, the Head Start Bureau provides training and technical assistance to help programs to meet standards and improve quality. Head Start's training and technical assistance system has been redesigned, and implementation is under way. The new system, initiated on September 1, 2003, consists of 12 contracted centers that are managed by the ACF regional offices. In the 10 geographically based regions, two or three training and technical assistance managers and several content experts (in such areas as early literacy, disabilities, health, and administration) work out of the regional offices, with the goal of facilitating closer communication and coordination. In addition, training and technical assistance specialists work directly with a group of 12 to 15 Head Start grantees, under the supervision of the training and technical assistance managers. According to Head Start Bureau staff, the new training and technical assistance managers and other staff are a

mix of experts who were part of the previous training and technical assistance system, former Head Start program staff, and other experts in priority areas. The training and technical assistance staff provide a potential source of assistance in implementing initiatives that are ready for a broad field test.

In addition to the new staffing structure for training and technical assistance, the Head Start Bureau plans to launch an extensive on-line network of resources for training and technical assistance staff and programs. The intent of this on-line network is to ensure that all staff members providing support to programs have access to the same set of high-quality resources.

The Head Start National Reporting System (NRS), implemented for the first time in fall 2003, is a child assessment system to collect uniform outcome data on all four- and five-year-olds in Head Start who are expected to enter kindergarten the following year. Under the NRS, all Head Start programs assess their four- and five-year-olds at the beginning and end of the program year on a selected set of language, literacy, and mathematics indicators, using a standard set of assessments. The Head Start Bureau and regional offices will use the findings to identify areas in which programs need additional training and technical assistance to more effectively support the performance measures in the 1998 reauthorization.

The NRS potentially could provide baseline data for small- or large-scale evaluations, as well as follow-up data, when relevant, for large-scale field tests of quality enhancement initiatives. However, NRS data are not currently available for research purposes, nor are they available below the grantee level. Obtaining center- or classroom-level NRS data is likely to be critical for evaluating quality enhancement initiatives for a field test (Stage 3), and child-level data will be necessary to compute impact estimates in smaller-scale Stage 2 evaluations. A study of the implementation and system improvement of the NRS is ongoing, including a Technical Working Group that reviews current progress and recommends system improvement activities. This group is currently considering a field test of social-emotional measures, potential sampling options, and methods for assisting Head Start programs in interpreting and using the results of the NRS and other assessments. It is worth noting that a new DHHS Secretary's Advisory Committee has recently been appointed, focusing particularly on assessing progress in the development and implementation of the NRS. This group is also charged with considering ways to integrate the NRS into broader assessments of early childhood learning found in FACES, the Head Start Impact Study, and the Child Outcomes Framework.

E. CURRENT HEAD START RESEARCH AND DATA

Rigorous evaluation of quality enhancement ideas and program variations in Head Start will build in important ways on the existing frameworks for Head Start research—the blueprint for Head Start research and the plan for evaluating Head Start impacts, set forth by two advisory committees (U.S. Department of Health and Human Services 1990 and 1999). In addition to emphasizing the importance of studying overall impacts of Head Start, the two advisory committees also stressed the importance of determining which program practices work best to support children's development, and for which families and children

the programs are most effective. In this section, we review recent and ongoing major research efforts on Head Start to illustrate the scope and purpose of existing research efforts, and to explain how the current design effort would help to fulfill the vision for Head Start research described by the two research advisory committees.

Our review is based on a discussion of four research efforts, each of which focuses on a subset of Head Start center-based programs. Given that many of these research efforts still are under way, the design for a study of Head Start quality enhancement and program variations will have to take into account the potential for "research fatigue" among programs selected for multiple studies. Ongoing research efforts could complicate the task of identifying programs and centers willing to participate in a study of Head Start quality enhancement. Nevertheless, in light of the large number of Head Start centers and the potential attractiveness of the offer of targeted assistance to implement a quality enhancement initiative, this challenge should not be an insurmountable one.

1. Family and Child Experiences Survey

FACES, launched in 1997, is a national longitudinal study of the characteristics, experiences, and outcomes of Head Start children and families. The survey was designed to provide information for the Head Start Program Performance Measures Initiative. The performance measures define the type and quality of services that programs must provide in order to successfully meet the goals and objectives of Head Start. Measures cover all of the major program objectives, including program management and parents' involvement; the provision of educational, health, and nutritional services for children; the ability to link children and families with community services they need; families' economic well-being, health, and parenting; and children's school readiness (U.S. Department of Health and Human Services 2003a). Data collection includes assessments of the quality of the classroom environment and measures of children's emergent literacy, mathematics skills, social skills, and problem behaviors.

The FACES study of Head Start programs and children is based on a nationally representative sample of center-based programs from the 10 ACF regions.[6] The sample is drawn in two stages. For the fall 2000 cohort, a sample of programs stratified by Census region, percentage minority, and urban/rural status was drawn. This selection yielded 43 programs. Within those programs, a sample of classrooms was drawn, yielding 286 classrooms. All 2,790 children in the selected classrooms were eligible for the study. Sample selection for the earlier FACES cohort (1997) and for the most recent FACES cohort (2003) was conducted in a similar fashion. Forty programs were included in the sample for the 1997 cohort, and 60 programs were included in the 2003 cohort. The previous FACES study followed the 1997 cohort of three- and four-year-old Head Start children from program entry through first grade. Children from the 2000 cohort were followed through

[6] Migrant/Seasonal, American Indian/Alaskan Native, Early Head Start, and programs in the territories were excluded from the sample frame. The sample frame included 1,675 programs.

kindergarten (U.S. Department of Health and Human Services 2001 and 2003a), and children from the 2003 cohort are also being followed through kindergarten.

The FACES battery includes a child assessment, parent interview, teacher and staff interviews, and classroom observations. Children are directly assessed on vocabulary, emergent literacy and early mathematical skills. In addition, their social skills and problem behaviors are assessed through rating scales completed by parents and teachers. Parents are interviewed about parenting behaviors, the socioeconomic characteristics of the family, and parental health. Interviews with classroom teachers, center directors, program directors, and coordinators gather data on the staffs' experience, education, and training, as well as on their attitudes about early childhood education practices and activities with children and parents. Classroom observations collect data on the structure of the classroom, the overall quality of the classroom environment, and teacher-child interactions (U.S. Department of Health and Human Services 2001 and 2003a).

The FACES studies provide a nationally representative, descriptive picture of center-based Head Start programs, and a longitudinal description of children's development during the Head Start year and into kindergarten. The majority of Head Start programs were rated as good on a global measure of the quality of the classroom environment. Children achieved substantial gains (of one-third to more than one-half a standard deviation) in vocabulary, mathematics, and writing skills during the Head Start year. Moreover, scores at the end of the Head Start year and the size of the gains in scores during the year predicted achievement at the end of kindergarten. Although the study has not been designed to provide estimates of the contribution of staff qualifications or aspects of program quality to improvements in children's achievement, it does suggest a relationship among staff qualifications, program quality, and children's development.

2. The National Head Start Impact Study

In 1998, as part of Head Start's reauthorization, Congress instructed DHHS to conduct a national study to measure the impact of Head Start on the children served by the program and took into account the recommendations of an advisory panel in its design (U.S. Department of Health and Human Services 1999). The National Head Start Impact Study includes nearly 5,000 three- and four-year-old preschool children who applied to 383 centers from 84 grantee/delegate agencies across the country. The children in the study have been randomly assigned to receive Head Start services (the treatment group) or not receive Head Start services (the control group).[7]

The impact study has two main objectives. The first is to estimate the impact of Head Start on the school readiness of children receiving Head Start services compared with children who are not enrolled in Head Start. The second objective is to investigate the conditions under which Head Start works best, and the children for whom the program has

[7] Only centers that had more eligible children and families than they were able to serve participated in the study.

the greatest impacts. The impact study will examine various factors that may differentially contribute to Head Start's impacts, including differences among children attending Head Start, differences in children's home environments, variations in the types of Head Start programs available, and the availability and quality of other child care and preschool programs in a particular area that could influence outcomes for control-group children (Lopez et al. 2002).

Data collection for the study began in fall 2002 and continues through 2006. Data sources include parent interviews, child assessments, surveys of child care providers and teachers, direct observations of care settings, and teachers' ratings of children. Children will be followed through the spring of their first grade year. The study will provide a nationally representative picture of the quality of Head Start classroom settings for three- and four-year-old children and a similar picture of the quality of early childhood settings (from other preschools to home-based child care) used by children who did not attend Head Start. The study will yield estimates of the impact of Head Start on children's achievement into the early school years relative to the impact of other early childhood care experiences that children would have in the absence of Head Start. Subgroup impact estimates will indicate whether the impacts of Head Start are different for key subgroups and for Head Start programs with different characteristics.

The random assignment design can yield strong evidence about the impacts of Head Start on children's achievement relative to other early childhood care options in the community in the absence of Head Start; however, evidence about the impacts of different program approaches and practices (based on subgroup analyses) is weaker. The subgroups are formed by grouping programs that have particular characteristics in common, and, even if differential impacts are found, they could stem from some other factor common to the group of programs that was not identified but has driven the results. Consequently, to rigorously examine the question of whether a particular promising program practice has stronger impacts on children's development than do the basic Head Start program services, it is necessary to use a design such as we describe in this report, in which centers or classrooms implementing a new approach are compared with centers or classrooms offering regular Head Start Services.

3. Head Start Quality Research Center Consortium and Data Coordinating Center[8]

The Head Start Quality Research Center (QRC) Consortium originally was funded to explore research questions related to program quality and program outcomes in Head Start. The first QRC Consortium (funded from 1995 through 2000) included four teams of university-based researchers or consultants who worked with one or more Head Start program partners to study the relationships between aspects of program quality and child and family developmental outcomes. QRC Consortium members also served as technical advisors on the design, development, and implementation of program performance measures, including the first Head Start Family and Child Experiences Survey (FACES).

[8] Information was compiled from the QRC meeting minutes and FACES reports.

The current QRC Consortium was funded for five years beginning in March 2001 with the goal of developing interventions to promote the school readiness of preschool children in Head Start. QRC Consortium members include the same researchers and programs involved in the first QRC Consortium, plus four additional partnerships. Many of the researchers expanded their partnerships to include more Head Start programs. Interventions include the types of quality improvements and promising practices about which the Head Start community is eager to learn more, including a stronger implementation of a general classroom curriculum, approaches to enhancing children's language and emergent literacy development, approaches to enhancing children's social-emotional well-being and behavioral development, individualizing of instruction for children, and enhancement of parents' ability to teach their children at home.

The plan for the QRC Consortium project has called for the researchers to implement the interventions on a pilot basis during the first year, and to collect outcome data on children during the fall and spring of that program year. Subsequently, using insights gained from the first year of implementation, the researchers are to replicate the interventions in several other program sites, and, using a random-assignment design, are to evaluate the effectiveness of the interventions. Grantees are funded to continue these replications for several additional years, in different programs. The pilot implementation year yielded many important lessons that were used to modify the design of the interventions and the strategies for implementing the interventions for the replication year. The QRC Consortium's experience supports the need for pilot implementation of an intervention to learn practical lessons about the feasibility of implementing the intervention on a larger scale in the context of an evaluation.

ACF established the Data Coordinating Center (DCC) to systematically collect cross-site data for the QRC Consortium, in addition to the data collection carried out by the QRC researchers. The DCC provides pre- and post-intervention data collection and analysis on a core set of cross-site measures, such as program quality, child outcomes, and parents' involvement. A consistent set of measures has been used over several replication years to maximize the analysis sample. The data collected by the DCC enable researchers to compare findings from the QRC studies with those from the national FACES sample. They also provide a mechanism for comparing site-specific instruments with the FACES instruments.

The QRC studies are an important forerunner to the current effort to design a framework for Head Start research on quality enhancements and program variations. Their fundamental approach—initially implementing an intervention to a high degree of fidelity on a pilot basis, and then implementing it with a rigorous experimental design and measuring children's outcomes—is a strong one for studying the impacts of promising practices in Head Start. Because many of the QRC evaluations have included relatively small samples, their most important contribution may be to strengthen development of the enhancement idea and providing important lessons on the value of implementation partnerships. The QRC projects could thus offer a set of ideas and implementation strategies for Stage 2 evaluations.

4. Other Studies of Promising Practices that Include Head Start Programs

Several other studies funded by the U.S. Department of Education (ED) and a consortium of agencies including the National Institute for Child Health and Human Development (NICHD), ACF, the Office of the Assistant Secretary for Planning and Evaluation (DHHS), and the Office of Special Education and Rehabilitative Services (ED) are examining the impacts of promising practices in preschool programs on children's school readiness. All of these projects include Head Start programs, although many other types of preschool programs also are included in the samples. The studies include program evaluations that are similar in design to the Head Start impact study, as well as evaluations of promising practices (primarily curricula) that are similar in design to the QRC Consortium projects.

The NICHD consortium study, called the Effectiveness of Early Childhood Programs, Curricula, and Interventions in Promoting School Readiness study, includes eight grantees, some focusing on or including Head Start programs. The grants are given to study the effectiveness of early childhood interventions and programs across a variety of early childhood settings in promoting school readiness for children, from birth through age five, who are at risk of later school difficulties. The interventions include several teacher training and curriculum studies (focusing on such areas as children's language development, emergent literacy, mathematics skills, and social-emotional well-being) and a parent-focused intervention to enhance parent-child interactions. Researchers will use rigorous designs when implementing the interventions and will measure children's development subsequent to the children's exposure to the intervention (or control) early childhood environments. However, there is not a common, cross-site core of measures being administered by a data coordination center.

Head Start programs also are among those included in ED's evaluations of alternative preschool curricula (the Preschool Curriculum Evaluation Research [PCER] project). This project includes 12 grantees selected during two rounds of grant competitions that are charged with implementing and rigorously testing alternative preschool curricula. Two national coordinating contractors, one for each PCER cohort, are collecting cross-site data on classroom quality and children's outcomes and will compare the treatment curricula with prevailing practice, estimate overall impacts of the treatment curricula on a common set of child outcomes, and analyze subgroup impacts. Most PCER grantees have implemented the treatment curricula and are in their first year of data collection on children in preschools.

The Early Reading First evaluation will estimate the impacts of a new funding stream from ED designed to enhance the language and literacy environment of preschool classrooms on children's language development and emergent literacy. Grantees can use the funds for a variety of purposes, including teacher education and training and the purchase of books and materials. Sixty grantees (each with approximately five preschool programs) were selected during two rounds of grant competition. All of them were expected to implement their Early Reading First programs by January 2004. Head Start programs comprise about 20 percent of the preschool programs receiving Early Reading First funds. A national evaluation contractor will select programs receiving Early Reading First funds and will compare the language development and emergent literacy skills of children enrolled in Early

Reading First with those of children who are not enrolled in Early Reading First programs. Preschool programs receiving Early Reading First funds and those without such funds will be compared on the basis of observational assessments of the language and emergent literacy environments of the preschool classrooms.

ED's Even Start program is designed to enhance the literacy skills of two generations—the parent and child—through an intervention focusing on adult literacy instruction; the enhancement of parent-child interactions; and center-based, child-focused literacy activities. Families are eligible if they have a child between three and eight years of age. Three previous evaluations did not find any impacts of the Even Start program on child outcomes (St. Pierre et al. 1995; Tao et al. 1998; St. Pierre et al. 2003). The current evaluation has been designed to implement and test alternative program enhancements. Two classroom curricula and two parent education curricula (focusing on parent-child interactions around language and literacy activities) are being implemented and evaluated using a rigorous design. The Classroom Literacy Interventions and Outcomes evaluation will estimate the impacts of the alternative classroom curricula (relative to current Even Start program practices) and of the additional impact of each of the parent education curricula on children's language development and emergent literacy. Several Head Start programs offer Even Start services and are included in this study. Programs that agree to participate in the study will be randomly assigned to implement one of the classroom curricula or a classroom curriculum and a parent education curriculum, or to implement none of these curricula, but continue operating under their current designs.

F. LESSONS FROM PREVIOUS HEAD START PLANNED VARIATION STUDIES

One historic set of Head Start research initiatives that can inform the present design of research on Head Start quality enhancements is the Head Start Follow Through and Planned Variation studies of the 1960s. Like the current efforts to learn what works best in Head Start programs, these initiatives sought to address the question of how best to enhance the educational attainment of disadvantaged children participating in Head Start. Follow Through was conceived as an ongoing program that would provide additional education from kindergarten through grade three to children coming out of Head Start programs in an effort to preserve the gains children made during the Head Start year. When funding for Follow Through as an ongoing program was not forthcoming, policymakers decided to establish the program on a smaller scale, and to use it to test promising educational practices. A group of communities was invited to participate in the Follow Through study and was offered a choice of program models to implement. A year after Follow Through began, some communities already participating in that program were invited to participate in the Head Start Planned Variation study. If they agreed to participate, they were asked to implement in their Head Start program the Follow Through program models used in their communities.

More than 20 program models were tested in one or both of these studies. The models were extremely diverse. Some focused on promoting skills through drills (such as the Engelmann-Becker model), whereas others focused on fostering children's development

through experimentation, exploration, and verbalization. Still others focused on bilingual educational approaches. Some tested models that focused on parents' involvement in the schools.

Policymakers and researchers alike concluded that these planned variation studies provided little new information; methodological weaknesses in the evaluations shed doubt on the findings (Rivlin and Timpane 1975). Yet, the lessons that emerged about the best approaches to designing research to accompany the development of new ideas for serving Head Start children remain useful today. These lessons provide critical insights for current efforts to design a research approach that will enable the Head Start community to identify and create better approaches to supporting the development of economically disadvantaged children participating in Head Start. Those lessons include:

- ***Clarify the Objectives of the Study for All Stakeholders.*** Perhaps because Follow Through originally was viewed as an ongoing program, many stakeholders were unclear about the differences between a service program and an evaluation. For example, many communities instituted large-scale adaptations to the program model, so that it was difficult to interpret the findings.

- ***Use an Experimental Design.*** Communities were not randomly assigned to models in either study. Many of the schools and Head Start centers used as comparison groups were chosen from different communities and frequently served a less disadvantaged population. Consequently, it was unclear whether the findings from the studies were due to the intervention or to the differences in the characteristics of families in the treatment and control groups.

- ***Fully Develop and Define the Interventions Before Testing.*** Time constraints precluded full development of some of the models tested in Follow Through and Head Start Planned Variation at the time they were implemented. Much of the translation from theory into practice occurred during the first year of the evaluation. Because the key elements of the models were not well defined, interpretation of the findings was difficult.

- ***Provide Considerable Training and Technical Assistance.*** Teachers participated in a short workshop and then were asked to implement the model in the classroom. Many became overwhelmed and frustrated. In addition to the more theoretical underpinnings of the model, teachers need practical advice about the arrangement of the physical class space and the structure of the day. Curriculum assistants who observed the classes and provided feedback to the teachers were viewed as an essential component of implementation.

- ***Assess the Success of Implementation.*** The Head Start Planned Variation studies were criticized as having inadequate measures of how well the program model was implemented. As a result, it was not clear whether the absence of any effects was due to the program model or to problems with the implementation of the model.

- ***Use Outcome Measures that Can Discriminate Among Models.*** The outcome measures used in the early Head Start Planned Variation studies were criticized for being insufficiently tied to the objectives of the program models. The outcomes focused excessively on cognitive measures and ignored measures of other aspects of development that were the focus of some of the models. The meeting participants criticized the standardized measures used, such as the Metropolitan Achievement Test, because they were designed to give comparable measures of children's performance irrespective of the curriculum, and therefore were not good discriminators among the effects of different curricula. Outcome measures should represent key outcomes for all interventions, so that the relative success of different interventions can be compared.

This report builds on the lessons of both the early Head Start Planned Variations study and the more recent research efforts. The three stages of research are intended in part to support the development of quality enhancement ideas so that implementation is well-understood and fully established before any evaluation begins. Rigorous, experimental designs are proposed for the evaluations. Measurement plans include outcome measures that are relevant to the quality enhancements being tested. Partnerships with Head Start programs, staff, and families are critical elements for the overall success of the quality enhancement development and research process.

CHAPTER III

STAGE 1: DEVELOPING QUALITY ENHANCEMENT IDEAS

Head Start programs are a natural laboratory for innovations in early childhood education. At any given time, numerous initiatives are under way to improve the quality of the Head Start experience for children. Quality enhancements to the Head Start program may be generated by programs themselves in order to address their own specific needs, or they may be initiated by the Head Start Bureau or regions in response to more general needs. The enhancement may consist of a "pre-packaged" curriculum or teacher training module that is purchased from a third-party vendor, or it may be a program management practice or community outreach program created by a Head Start program. Although there may be a wealth of ideas in practice, little information is available about which enhancements work, or, at a more basic level, how they work.

Given the plethora of ideas and current practices, it can be a challenge to the Head Start community to sift through these activities to identify program enhancements that are ready for and worthy of rigorous evaluation. An initial development stage as we describe in this chapter offers a systematic framework for an enhancement's development, ensuring that good ideas have an opportunity to progress to promising practices while ideas that are less well-developed or are not easily replicable are filtered out. This stage is a period of enhancement definition, documentation refinement, and early experimentation with implementation and measurement. If an enhancement is unable to meet the goals of Stage 1 by achieving clarity and replicability, then it should not be considered for evaluation in Stage 2. As such, an initial development stage can contribute significantly to Head Start by ensuring that only the most promising quality enhancements undergo rigorous evaluation.

We begin this chapter by discussing the rationale for a development stage. We then turn to a discussion of the specific goals of this stage, including the tools to be developed and the conditions to be met before a quality enhancement is ready for evaluation. We also discuss the duration and activities of Stage 1, including the products that should result and the players who should contribute to the work of this development stage. We conclude the chapter by presenting three examples of specific quality enhancements to illustrate the process and goals of Stage 1 activities.

A. THE DEVELOPMENT STAGE: RATIONALE AND OVERVIEW

It is not possible to evaluate rigorously every enhancement idea or strategy. Even though a planned variation evaluation design provides the opportunity to test multiple conditions at once, important decisions and choices about what to evaluate still must be made. A development stage can help to identify the ideas that are backed by well-defined theories of change; feasible measurement frameworks to assess changes; and clear, thorough documentation for implementation in broad and diverse settings. In addition, descriptive outcome studies in Stage 1 should be suggestive that the enhancement has the potential to produce positive improvements in selected child outcomes.

The development stage ensures that the techniques for full and successful implementation have been refined, and that researchers can accurately measure the quality of implementation and the fidelity to the enhancement's intended goals. An intervention can fail for two main reasons: (1) the theory underlying the intervention is flawed, or (2) the intervention has been implemented poorly. Understanding the challenges of implementing an enhancement in the real world is important, but a test of a poorly implemented enhancement may provide misleading results. During the development phase, researchers and program developers will test approaches to implementation to ensure that programs can both implement the quality enhancement well and maintain fidelity to the model. Especially if broader implementation is anticipated, a careful study of implementation and fidelity can provide useful information for future training and technical assistance (T/TA) efforts. It can also provide information about the kinds of variation in implementation to be expected across different types of programs.

Stage 1 can also serve as a continuous feedback cycle for positive program improvement. While programs are defining implementation procedures, they also are continuously refining the enhancement model by adjusting implementation in the face of unexpected challenges or circumstances. In this same way, Stage 1 also may demonstrate that an enhancement cannot be replicated in all Head Start programs, but that it has value that could be relevant for specific population or risk subgroups.

The challenge, and the ultimate value, of the development stage is to take the quality enhancement ideas bubbling up from programs, curriculum developers, researchers, and others, and to move them toward greater clarity and replicability. Thoroughly testing, documenting, and measuring implementation during Stage 1 will improve the overall quality of a rigorous evaluation in Stages 2 and 3 by limiting the likelihood of poor implementation, and by increasing the confidence in the results.

B. GOALS OF THE DEVELOPMENT STAGE

The goals of the development stage are to (1) define the enhancement by documenting the theory of change and by detailing the key elements that must be visible at full implementation, (2) refine and document the implementation process, and (3) establish measures that can be used to assess the quality of implementation and fidelity to the enhancement model. At the end of this stage, clear, thorough documentation of the enhancement must have been developed for use by other programs for replication, and by

researchers to assess implementation as part of an evaluation. The extent of work required to accomplish these goals by any particular enhancement strategy will vary. Some enhancements may be fairly well documented and may already have been implemented in some Head Start programs. Nevertheless, a development period—the length of which may vary by enhancement—provides an opportunity to refine documentation, and to determine whether the enhancement can be implemented successfully by other Head Start programs.

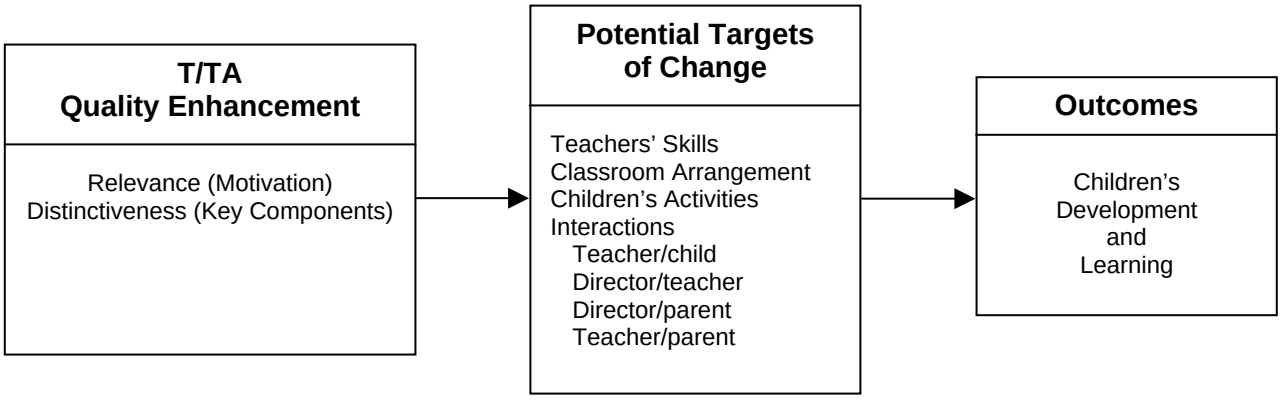
1. Defining the Enhancement

At the outset, enhancements in Stage 1 should meet the initial criteria of relevance and distinctiveness. As such, each enhancement should be supported by a strong, relevant theory of change and should have key elements that set it apart from current Head Start practice. The first goal of Stage 1 is to ensure that an enhancement has a clear, comprehensive definition through a documented theory of change. A firm understanding of the theory of change will motivate staff, guide implementation, and focus eventual measurement selection and development. A clear and focused theory of change also can serve as a good "sales" pitch to other Head Start programs at the point of replication.

At this stage of enhancement definition, the theory of change should communicate three key components, as shown in Figure III.1:

1. What is the quality enhancement? What are the essential components that set this enhancement apart from current practice? What should this enhancement look like when fully implemented?
2. Who or what is the target of change? What is expected to change in order to bring about the improvement in program and service quality?
3. What aspect or domain of children's development and learning is expected to improve as a result of the quality enhancement? Through what avenues?

Figure III.1. Conceptual Model: Theory of Change



T/TA = Training and Technical Assistance.

Note that Figure III.1 is not an actual example of a logic model, but is a broad representation of the key components of a theory of change and their connections. At this stage, a highly stylized logic model may not be necessary. What is necessary is a fundamental understanding of the underlying theory of the enhancement that can be communicated in clear, focused, descriptive documentation. Documenting the theory of change in Stage 1 is equivalent to telling the "story" of the enhancement—what it is, what changes it should generate in the program, and what improvement it should produce for Head Start children.

2. Documenting Implementation

After the enhancement has been clearly defined, the next step is to understand the implementation process in detail, refine it, and document it. Implementation documentation refers to implementation manuals, plans for classroom activities, parent-training materials, teacher-training protocols, and other materials that specify the steps necessary to implement the quality enhancement with high fidelity to the enhancement model in a large number of Head Start programs.

Many innovative, exciting enhancement strategies already are under way throughout the Head Start community, but programs may be implementing these new ideas on the basis of their intuition, rather than on documentation. For the field to truly benefit from these innovations through evaluation, greater clarity in documentation is required. Without it, other programs will have difficulty replicating the techniques, which, in turn, may limit the potential for success. In essence, then, Stage 1 documentation activities are an opportunity for programs to share what they believe to be "best practices" while laying a firm foundation for Stages 2 and 3 that follow. The implementation process of quality enhancement strategies generally consists of three key components:

(1) the initial training and/or technical assistance to launch the enhancement

(2) the ongoing training and/or technical assistance to support implementation

(3) the ongoing activities to enhance services (that is, day-to-day implementation)

In Stage 1, it is the task of enhancement developers and/or program staff to specify each of these steps, and to then document them clearly and consistently for replication. Specification and refinement of the enhancement may be a somewhat experimental process in a development stage in which identifying the most effective and most efficient approaches to implementation is encouraged. In Stage 1, programs can test different approaches to implementation to identify the lowest-cost strategy that is effective in implementing the enhancement initiative to a high level of fidelity. Specification of the enhancement does not mean that, ultimately, the final model will be a one-size-fits-all model. Rather, the documentation will specify the dimensions that can be varied to accommodate differences in program resources, child and family diversity, or staff qualifications and/or composition, as well as the dimensions that are critical and that must be uniform in order to maintain fidelity to the enhancement model. Furthermore, even though Head Start programs share core requirements, each program is unique. As a result there may be a range of starting points for an enhancement, possibly determined through an initial program needs assessment before training begins.

a. Training and Technical Assistance: Initial and Ongoing

In general, implementation begins with some form of *initial T/TA* to Head Start program staff who are expected to deliver an enhanced level of services to children and families. For example, delivering teaching services according to a new curriculum will require that teachers receive training in that curriculum. Delivering teaching services with appropriate responses to children's behavioral challenges will require that teachers receive training in techniques for managing and responding to behavioral issues in the classroom. Improving links between the program and community-based health services could require the training of directors or key administrative staff in techniques for identifying community resources, initiating new relationships with health care providers, and developing partnership agreements to support new services.

A variety of strategies have been used to train Head Start staff to implement enhancement ideas. They include different options for training format, training intensity, and the skill level of the person conducting the training. Various approaches to providing resource materials for participant training have been used as well (for example, manuals, Internet-based resources, and distance-learning programs). These approaches differ in terms of cost (because of differences in the skill level of the trainers, the length of training, and other factors); they also are likely to differ in their effectiveness in conveying essential information about the enhancement to teachers or other staff who will implement the model.

The next step is to observe how well the training worked, and to continue to support effective implementation of the enhancement through *ongoing T/TA*. Some enhancement components or skills may take time to practice and incorporate into daily routines, but after the most important changes have been made, the program should be delivering an enhanced level of services consistent with the quality enhancement. Accordingly, the purpose of ongoing T/TA is to work with the teachers and program staff as they institute changes in the classroom or program to ensure that the enhancement is implemented to a high degree of fidelity. Technical assistance staff begin this process by assessing program services to determine how well they conform to the model. If inconsistencies are identified, technical assistance staff work with the teacher to further modify the classroom environment, activities, or behavior so that services adhere more closely to the model.

Documentation of both initial and ongoing T/TA produced in Stage 1 should specify the four key dimensions of the T/TA strategy:

1. **Content.** Details a training curriculum and provides training materials and activities that are specific to the quality enhancement to be implemented
2. **Intensity.** Specifies the amount of training to provide within a particular time period (such as a week), and the trainer-trainee ratio. For example, is training provided during a workshop with 50 participants or one-on-one in the classroom?
3. **Duration.** Specifies the length of training at a given level of intensity. For example, does an initial training workshop last three days, or is it provided during

a series of daylong workshops over a longer period? How many months of follow-up support are provided?

4. **Quality.** Specifies the educational and skill level of the T/TA provider and the quality of the training curriculum and other resource materials. For example, how much experience in the practical implementation of the enhancement or how skilled at using adult learning strategies should the trainer have?

The fundamental questions underlying the choice of T/TA strategies involve achieving a balance between the cost of the approach and the approach's effectiveness in ensuring that the quality enhancement initiative is implemented with a high degree of fidelity to the model. Understanding the range of potential implementation strategies and the relative costs and likely effectiveness of each one is essential for developing research designs that produce useful information for the Head Start community. The development stage provides an opportunity to change the intensity, duration, or quality of training in strategic ways to tailor the implementation. Some answers about implementation choices may be suggested in the development stage (for example, because a program that chooses a shorter initial training workshop than another program fails to meet fidelity thresholds). In contrast, other questions could be candidates for experimentation in the evaluation (for example, because two programs vary in the duration and/or intensity of the initial training while continuing to meet fidelity thresholds).

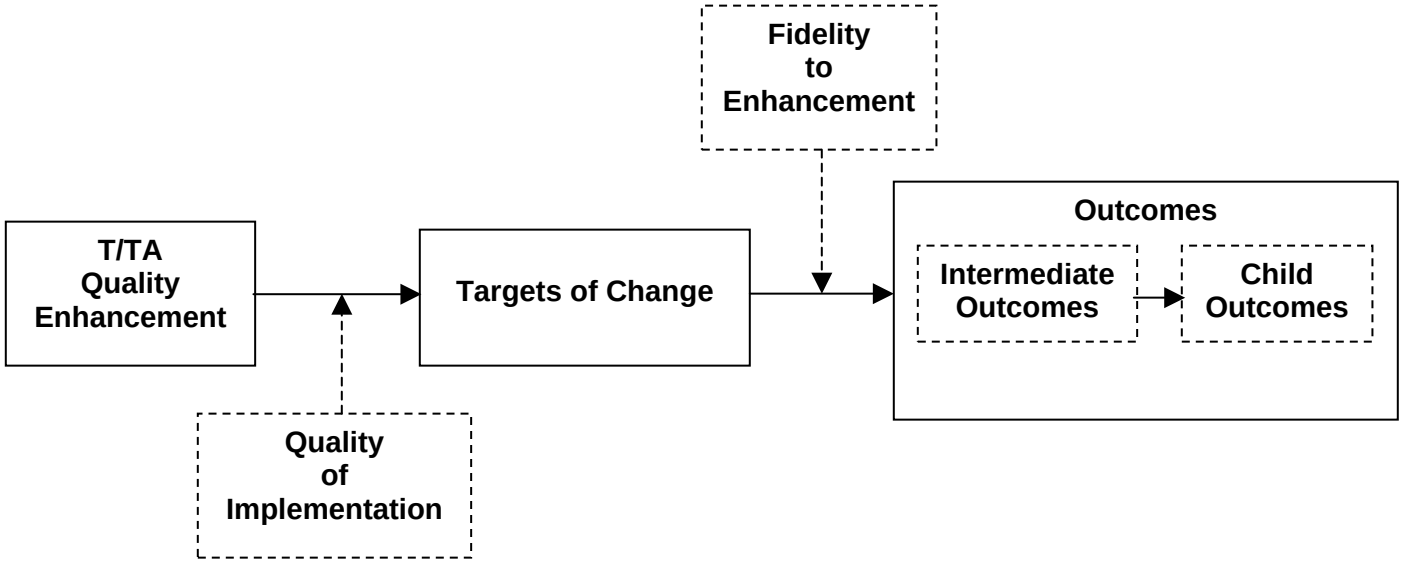
b. Day-to-Day Implementation Activities

The last key component to the implementation process is the *day-to-day implementation*. Day-to-day implementation is the actual content of all T/TA as well as of the teaching, procedural, and resource guides for ongoing program activities. The content is the heart of the enhancement, and it will take considerable thought and consideration to document each aspect of it. The content for a curriculum will include class plans, teaching aids, classroom management techniques, the details of the timing and flow of activities, and details about lessons for a particular day and over time. For some enhancements, implementation may extend beyond the literal walls of the classroom or even the center. For example, some enhancement strategies, such as those designed to increase children's access to health services, may involve the formation of partnerships with other local service providers.

3. Developing Measures

After implementation has been thoroughly documented, the next goal of Stage 1 is to create a measurement framework to gauge the success of implementation, the effects of the enhancement on children's environments, and ultimately, the effects on children's development. This framework will include measures and methods to assess the quality of implementation, the degree of fidelity to the enhancement model, and intermediate and child outcomes (Figure III.2).

Figure III.2. Conceptual Model: Theory of Change with Measurement



T/TA =Training and Technical Assistance.

The implementation process, as discussed in the preceding section, refers to the inputs and steps necessary for putting the enhancement idea in place. *Quality of implementation* refers to the extent to which programs are able to bring together all the resources necessary (such as staff with sufficient qualifications and classroom materials) and to carry out all the steps necessary (for example, initial training, the formation of partnerships, and group supervision activities) to effectively reach the targets of change, and to implement the enhanced services as planned. *Fidelity to enhancement* is the degree to which the enhancement model delivers the enhanced services as intended. In other words, the targets of change look or function as should be expected after the enhancement has been fully implemented. For example, measures of implementation quality for an enhanced classroom intervention might include an assessment of the teachers' qualifications, the teachers' training on the enhancement, supervisors' support, technical assistance from an outside trainer, and materials available in the classroom. To measure fidelity, teachers would be observed in the classroom to determine whether they are implementing the enhanced activities at the levels of quality and frequency expected. *Intermediate outcomes* differ from implementation and fidelity measures in that they provide more-global measures of change that are known to be related to child outcomes. If an enhancement affects the intermediate outcome measures, then the potential exists to affect child outcomes. Finally, *child outcomes* capture the changes expected to occur in children's development and learning as a result of the enhancement.

a. Measuring Quality of Implementation

To assess implementation quality, enhancement developers or evaluators can develop criteria that gauge how well each step is being implemented, based on plans, procedures, and other documentation. Criteria can be developed for each phase of implementation, from initial training to the intensity and duration of the enhanced service provision. Data on the

processes of implementation would then be used to assess the extent to which resources have been brought to bear and implementation steps carried out according to plans and procedures. For example, to assess the quality of initial training, evaluators might compare the actual qualifications of trainers, content of training, intensity and duration of sessions, and teacher participation rates with those projected in the enhancement training plan. Some programs already may use available data to conduct their own self-assessments; these data could readily be adapted to assess implementation quality. Alternatively, if new measures are developed to assess implementation quality for an evaluation, they also may be useful for continuous program improvement efforts, and to identify programs' technical assistance and training needs.

Depending on the resources available for the task, enhancement developers or evaluators could develop measures that are sensitive to fine-grained differences in implementation across programs, and that rely on multiple data sources, or they could develop measures that focus on assessing central features of the enhancement, and that require less-detailed information. Sets of Likert-type scales to rate each implementation step along various dimensions could provide a more in-depth analysis. Such an analysis might show, for example, that qualified trainers covered all the specified training topics, but that the training was shorter than expected, and that teacher participation was somewhat low. Alternative measures might consist of sets of "yes/no" indicators for whether specific steps were completed.

The National Evaluation of Early Head Start provides a useful example of how implementation quality measures can be developed and used in an evaluation. The implementation study conducted as part of the Early Head Start evaluation sought to examine all aspects of the comprehensive services provided through the program; thus, it probably was a larger, more complex effort than may be necessary for most quality enhancements. Nevertheless, the design and methodology of the implementation study could be adapted for more modest efforts. For the Early Head Start implementation study, researchers developed a set of 25 rating scales to assess implementation quality, referred to as "full implementation" in the Early Head Start evaluation (Paulsell et al. 2002). The scales were based on key requirements of the Head Start Program Performance Standards (U.S. Department of Health and Human Services 1996). Each scale contained five levels, ranging from minimal implementation (Level 1) to enhanced implementation (Level 5). Using a similar methodology to assess implementation quality of a Head Start quality enhancement, program developers or evaluators might use training plans and operations manuals to develop rating scales for key aspects of implementation (for example, initial training, support provided to teachers implementing the enhancement, and the frequency and intensity of enhanced services provision).

The rating process for the Early Head Start evaluation drew on multiple data sources (for example, semistructured interviews and focus groups conducted with staff, parents, and community partners; staff surveys; reviews of program records; and service use data) and summarized a large amount of detailed information about program implementation into a concise set of ratings. Evaluators aggregated the 25 ratings into a summary rating for each program area and an overall implementation rating for the program. Similarly, measures

developed to assess the quality of implementation of Head Start enhancements in Stage 1 might draw on multiple data sources.

b. Measuring Fidelity

When a quality enhancement is fully implemented, it should be possible to observe changes to usual practice. Aspects of what children experience in Head Start classrooms or home environments should have changed due to the specific behavior, home, or classroom characteristic targeted for intervention. Measures of fidelity to the enhancement model quantify how closely those changes adhere to the "ideal" vision of how the enhancement should be implemented in the classroom or other setting. Because fidelity measures aim to quantify both aspects of the classroom or home environment and the behavior of teachers or others, the fidelity data usually will be collected through observation. Data will be collected to answer the following questions, among others: Did a target behavior occur? How often? How well was the target behavior carried out by the teacher or by an assistant teacher?

Because fidelity measures are designed to document the occurrence of specific behaviors or features of the environment targeted by the enhancement model, they must be tailored to the specific enhancement under study. The range of choices in existing measures will be restricted by the type of enhancement chosen; for example, if the enhancement is a classroom intervention, then evaluators must choose from among existing classroom observation measures. Evaluators will have to determine during Stage 1 whether they will adapt existing measures, or whether they will develop new measures that better tap the important features of the enhancement. The fidelity measurement status at the start of a study may fit one of four main variations: (1) a measure exists that can be used as it is; (2) a measure exists that has to be adapted slightly; (3) no suitable measure exists, but the structure and content of an existing measure can be adapted; or (4) no suitable measure of fidelity exists. Here, we describe these scenarios and their implications at each stage of evaluation.

1. ***A measure exists that can be used as it is.*** Some enhancements may have measures that are suitable for capturing fidelity. For example, a quality enhancement for programs that do not use a specific classroom curriculum might be to adopt such a curriculum. If the evaluation of this enhancement strategy involves comparing adoption of the High/Scope curriculum with adoption of the Creative Curriculum, then fidelity measures that have been developed by the respective curriculum publishers can be used at Stage 1.[1] Evaluators can determine whether the existing cutoffs provided by the publishers are meaningful in distinguishing what is happening in classrooms; they also could develop additional cutoffs based on Stage 2 analyses of how the fidelity measures related to intermediate and child outcomes. In addition, because these measures

[1] In this example, we assume that the two fidelity measures and their thresholds for determining fidelity are equally rigorous. If they are not, it will be difficult to compare fidelity across the two curricula.

of curriculum fidelity are long, evaluators might determine at Stage 2 whether they can be streamlined for the Stage 3 field test.

2. ***A measure exists that has to be adapted slightly.*** In some cases, an existing fidelity measure may require only minor changes in the way that it is conducted or in the content of the items. For example, the PCER team studying *Ready, Set, Leap!* adapted an observational measure from a longitudinal study (Vellutino and Scanlon 2001) into a very detailed sampling procedure called Classroom Language Arts Systematic Sampling and Instructional Coding (CLASSIC; Scanlon et al. 2003; and Scanlon and Vellutino 1996 and 1997). When an existing measure is adapted, Stage 1 work could determine whether the measure taps the key dimensions of the activities that teachers should be doing; cutoffs for different levels of fidelity can be developed in Stage 2.

3. ***No suitable measure exists, but the structure or content of an existing measure can be adapted.*** The Early Language & Literacy Classroom Observation Toolkit (ELLCO; Smith and Dickinson 2002) is an observational measure of the classroom literacy environment with a structure that could easily be tailored to specific literacy enhancements. For example, the ELLCO, as it currently exists, would not be sufficiently tailored to an enhancement that focuses on dialogic reading, but the items in it could be altered to address the target behaviors while preserving its basic observation structure.[2] Alterations would be made in Stage 1, and cutoffs would be determined in Stage 2 based on the relationship of the new fidelity measure to intermediate and child outcomes.

4. ***No suitable measure exists.*** During Stage 1, evaluators will work with program developers and Head Start staff to document exactly what they expect to happen in classrooms if a particular enhancement is implemented with a high degree of fidelity. One way to discuss this topic with program developers and front-line staff is to facilitate discussions about what would look different in a classroom if the classroom were to implement the enhancement with fidelity. What would an observer expect to see? What observation period would be necessary in order to determine whether what is happening is faithful to the model?

The development or adaptation of a fidelity measure for a given enhancement, as well as its use in an evaluation, must meet a number of important criteria. The implementation framework described for T/TA is adapted here to apply to the fidelity measures. In addition to determining whether a target behavior occurred during a classroom or a home observation, measures of fidelity must meet the following criteria in order to document whether enhancements are implemented with fidelity at the classroom level:

[2] Dialogic reading refers to adults involving children in reading by prompting children to speak about the book, asking questions about the content, evaluating the children's responses for the level of understanding, and extending the information presented during daily activities.

1. **Intensity.** How often do classroom staff perform the target behaviors with children? Are they done with the children in a group or in one-to-one interactions?
2. **Duration.** How long do target behaviors at the group or child level last? Are the behaviors sustained throughout the program year?
3. **Quality.** How well do classroom staff members implement the target behaviors? Is the quality of the implementation consistent across all staff?

The QRCs and the PCER grantees provide excellent examples of preschool experiments that have developed fidelity measures. The fidelity measures that each project team is using are tailored to the specific intervention under study and range from existing measures tailored to the study to measures developed specifically for the study. For example, in addition to adapting the CLASSIC measure to study *Ready, Set, Leap!*, the PCER team studying *Building Language for Literacy* developed a new fidelity checklist that observers use to code whether a list of specific activities occurred during the observation period. The experiences of the QRC and PCER grantees with these fidelity measures will provide ACF with important information about the challenges to measuring fidelity and the successes on which future Head Start enhancement studies can build. Evaluators studying future enhancements may be able to draw on or to adapt these measures to new research projects.

As we have discussed, evaluators must set a minimum threshold for determining whether children's experiences are faithful to the enhancement model. An enhancement that has been designed to change a variety of behaviors requires many fidelity measures and either a minimum threshold for each measure or an overall threshold. A narrowly focused enhancement may require only one fidelity measure and one cutoff for determining whether the implementation is faithful to the model.

Cost Considerations. Clearly, the cost of measuring fidelity will be determined by the specific approach taken, whether a preexisting measure of fidelity can be used, the training requirements for the measure, and the frequency of the observations. The following cost-related questions are the main ones:

- How much measurement development work is required?
- What are the training requirements for observers/coders?
- How long is each observation, and how often does each occur?
- How can the work conducted during the development stage inform reduction of the resources required to measure fidelity during Stages 2 and 3?

Measures of fidelity to the enhancement model, based on observations of classrooms and teachers or of parents and home environments, are expensive to develop and implement. Observational measures may be supplemented with teacher or parent logs that record the frequency with which these individuals perform target behaviors every day. Depending on the type of enhancement, reviews of teachers' lesson plans may provide additional information about fidelity. However, the validity of the self-report measures must

be tested against an observational measure. For most enhancements, any existing measures considered for adoption will likely be too general to capture the essence of the enhancement; as we discuss in the following section, they may be more suitable as intermediate outcome measures. For example, the fidelity of a literacy intervention that focused on teachers' regular use of dialogic reading techniques (Whitehurst et al. 1994) with individual children could not be adequately or efficiently measured by using broad-based measures of quality and the literacy environment, such as the ECERS-R (Harms et al. 1998) or the ELLCO. The PCER research team resolved this problem by working with CIRCLE to develop its measure of language and literacy activities in the classroom for use in a multi-site evaluation. Where possible, adapting existing measures is one way to reduce the measurement development costs of the research phase. Of course, for many types of enhancements, no measures are available, so they will have to be developed on the basis of Stage 1 collaboration among the enhancement developers, Head Start program staff, and evaluators. This process would build on hypotheses about what must take place in classrooms in order to provide evidence that teachers have incorporated the enhancement into their daily activities.

Implementing observational or other qualitative measures of fidelity well can be costly. To ensure that all observers/coders rate fidelity in a similar manner, the observers must establish inter-rater reliability with either the developers of the measure or with someone who knows the measure well enough to be considered a "gold standard" coder. As the fidelity measures are refined through input from Stages 1 and 2, evaluators will determine the most efficient ways to establish inter-rater reliability on these measures. They also will design observer training to address common problem areas. For Stages 2 and 3, the cost of establishing inter-rater reliability can be incorporated into the cost of conducting a centralized training of observers that includes reliability visits to community child care and Head Start facilities during training. For example, High/Scope (2003) reports that training to acceptable levels of inter-rater reliability on its Preschool Program Quality Assessment (PQA), a comprehensive rating instrument "designed to evaluate the quality of early childhood programs and identify staff training needs," takes three days.[3] High/Scope researchers recommend that the first two days include review and time to practice the items using videotapes of early childhood settings and actual visits to preschool programs. The third day is to be used for a full observation (with half the day spent observing a classroom to complete the classroom items, and half spent conducting interviews to inform the agency items).[4]

Most measures of classroom quality are conducted in two to four hours. Given that Head Start's center-based services must be offered for at least three and one-half hours per day, observation for two to four hours may be sufficient for measuring fidelity. Some large-

[3] The authors note that the PQA may be used in any center-based setting, rather than only in those using the High/Scope curriculum.

[4] This level of intensity of training and reliability testing for classroom observations is similar to the training efforts included as part of large-scale evaluations, such as the Early Head Start National Research and Evaluation Project, the Fragile Families and Child Well Being Study, and FACES.

scale studies require two half-day observations to ensure reliability, but the majority of studies require only one. For enhancements that focus on what parents do with children, the fidelity measure requires either an observation in the home or the use of a structured interview. When the enhancement is parent-focused, the evaluators must determine how to sample the children in the classroom, as home observations are very expensive.

Reliability and Validity. To choose among existing fidelity measures, evaluators must determine whether the measures have sound psychometric properties. In the absence of existing measure or after a measure has been adapted, evaluators must gather information about the measure's reliability and validity. The most important characteristics are:

- **Inter-Rater Reliability.** On each fidelity measure, can observers be trained to meet high standards of inter-rater reliability? (In other words, are all of them able to demonstrate that they are counting or rating what happens in the classroom in the same way?) Most researchers require that exact agreement or agreement within one rating point between observers correlate with each other at the .90 level to meet reliability standards.
- **Internal Consistency Reliability.** Do the fidelity measures "hang together" in meaningful scales that have good statistical properties? Most researchers require that the intercorrelation among scale items reach at least .70.
- **Concurrent and Predictive Validity.** Do the fidelity measures correlate with other measures of classroom quality (for example, the ECERS-R or the Arnett Caregiver Interaction Scale)? To provide evidence that the fidelity measures capture something important about classroom quality, they should correlate with other measures. Do the fidelity measures tap dimensions important for children's outcomes? If so, they should correlate with or be predictive of child outcomes.

c. Intermediate Outcomes

Intermediate outcomes are outcomes affected by the enhancement prior to its influencing child outcomes. Intermediate outcomes are global measures reflecting characteristics or conditions of the center, classroom, or family environment that are likely to change (intentionally or unintentionally) due to the enhancement. Intermediate outcomes also are theorized to predict child outcomes. Thus, it is through changes in intermediate outcomes resulting from implementing a Head Start enhancement that children are presumed to be affected by the enhancement. For example, intermediate outcomes germane to teacher- and classroom-focused enhancements include changes in the classroom environment and teaching practices that result from the teacher- or classroom-focused enhancement. Intermediate outcomes germane to a center-level intervention (such as management training of Head Start directors) include subsequent changes in center operations and management practices that result from the center-focused enhancement. If teacher- or classroom-level outcomes also change as a result of the center-focused enhancement, then an evaluation of a center-focused enhancement should plan to measure the classroom-level intermediate outcomes as well. Intermediate outcomes relevant to a

parent- or family-level enhancement (for example, educating parents about activities and parent-child interactions that support children's language development and literacy) include subsequent changes in parenting behavior and other aspects of the child's home environment that result from the family-focused enhancement.

Intermediate outcome measures are designed to (1) capture broader aspects of change that will not be captured by the implementation and fidelity measures, (2) provide data using a measure with a proven link to targeted outcomes in children and adults, and (3) allow for comparison with other studies that have used the measure. In an enhancement that focuses on dialogic reading, evaluators might complement an observational fidelity measure of the number of times and duration that teachers used dialogic reading with a proven measure of classroom quality, such as the ECERS-R (Harms et al. 1998) or the Assessment Profile (Abbott-Shim and Sibley 2001). Including intermediate outcome measures enables evaluators to study why an enhancement might be effective in one classroom but not in another. For example, a classroom that implemented the dialogic reading enhancement with high fidelity according to a narrow measure of fidelity might not experience any other changes in overall classroom quality. The effect of the enhancement in that classroom might be less than the effect of the enhancement in a classroom that had both high fidelity and an overall increase in observed quality. Implementation study data could be used in this case to determine what else may have changed in the setting that increased classroom quality relative to the one that did not increase it. Choosing an intermediate outcome measure that has been widely used in research also provides a way to benchmark Head Start findings against other studies. For example, it would be possible to determine whether a classroom launching an intervention had a higher score on the ECERS-R than did the typical Head Start classroom measured in the Head Start FACES study, and whether its score was higher than the typical preschool classroom studied in a state pre-kindergarten study.

The specific methods and measures for assessing intermediate outcomes will be tailored to the enhancement under study. We expect that the four main categories of intermediate outcomes are those that assess changes in (1) the knowledge and skills of adults (directors, education coordinators, teachers, and parents); (2) classroom quality and what teachers do in the classroom with the children in their care; (3) the home environment and what parents do with their children; and (4) program partnerships related to providing additional services to children and families. Many potential intermediate outcome measures exist, but a goal of the development stage is to identify or create measures that meet several criteria:

- ***Relevance and Sensitivity to Enhancement Goals and Potential Spillover.*** The measures should focus on aspects of the environment and adult behavior targeted by the specific enhancement, but they also should be broad enough to capture other changes that might occur. An enhancement that focuses resources and staff attention on building early mathematics skills may unintentionally reduce the frequency or quality of teacher-child interactions related to literacy activities. Including a global measure of classroom quality enables evaluators to determine whether implementing the enhancement results in any additional positive or negative effects on Head Start classrooms other than the ones directly

targeted by the enhancement. The selected measures should have demonstrated sensitivity to these changes in inputs as staff training, education, and experience.

- ***Adequate Psychometric Properties.*** All measures should have adequate reliability and validity for use in classrooms designed for children from low-income families and with program staff and parents who care for these children. In general, measures should have a demonstrated internal consistency reliability of .70 or higher. (This level is generally accepted as an adequate demonstration of reliability.) In some cases, however, reliabilities as low as .65 might be tolerable if better measures of the same construct are not available. In addition, measures collected through observation must demonstrate good inter-rater reliability. The general standard is an exact agreement or agreement within one rating point, or a kappa correlation between observers of .90 or higher. In some cases, lower values (.85, for example) may be tolerable if no other measure exists. It is important to include measures with demonstrated predictive validity (that is, significant correlations with child outcomes or other outcomes targeted by the enhancement) because they provide another gauge of the likelihood that the enhancement will be effective.

- ***Previous Use in Large-Scale Surveys and Intervention Evaluations.*** To increase the comparability with other national studies and intervention evaluations, measures used in other national studies of similar populations (for example, FACES, the Head Start Impact Study, and the PCER project) should be selected. If a measure taps an important intermediate outcome but has not been used in a large study, evaluators should determine whether it has ever been used in settings similar to Head Start.

- ***Reasonable Cost and Burden.*** The measures must be able to be administered reliably by trained field staff, rather than by highly experienced graduate students or evaluators. Most of the measures will be observational, and therefore potentially costly. Consequently, efforts to streamline the observation protocols and to reduce the length and complexity of observer training are critical. In addition, the intermediate outcome measures should impose minimal burden on Head Start children, parents, and classroom staff. At most, a few clarifying questions can be asked of staff or parents as part of the observational protocols, but minimal disruption of the setting is the usual standard for observational measures.

As part of Stage 1 activities, evaluators, program staff, and enhancement developers will discuss their theories about the avenues through which the enhancement will affect children, and the ways in which the enhancement might spill over beyond its more narrow targets to affect other aspects of classroom quality, adults' skills and knowledge, and other important aspects of program quality. After review of existing measures for their coverage of the identified areas, a consensus on which measure or measures to use must be reached. In the absence of an existing measure, evaluators will have to either adapt an existing measure or develop a new one.

d. Child Outcomes

A critical first step in selecting outcome measures is to clearly articulate hypotheses about how the Head Start enhancement is expected to affect children. Which child outcomes in which domains (whether targeted or not) are expected to change as a result of a child's one-year exposure to the Head Start enhancement? The same criteria described for the selection of intermediate outcomes also apply to child outcomes. Specifically, measures of child outcomes must be relevant to school readiness goals; have demonstrated sensitivity to enhancement goals; be appropriate for use with a culturally diverse, low-income population; have adequate psychometric properties; be able to be administered with reasonable cost and limited burden to Head Start children, parents, and program staff; and, preferably, have been used in previously conducted large-scale surveys and intervention evaluations. The child outcome measures also must be valid and reliable for the intended mode of administration.

In addition to using sound, high-quality measures of important child outcomes, evaluations of Head Start enhancements must have a sound overall measurement strategy delineating the types and characteristics of outcomes that are important to measure. Several additional measurement issues should be considered as researchers and Head Start program partners begin identifying or creating a set of outcome measures to gauge the effects of a particular quality enhancement:

- **Measuring Nontargeted and Targeted Child Outcomes.** At the least, selected child outcome measures should reflect the particular aspect of school readiness that the Head Start enhancement is targeting. However, a particular Head Start enhancement may affect, for better or worse, outcomes that the enhancement is not targeting directly. On the one hand, school readiness is a multidimensional concept (see, for example, Love 2003; McWayne et al. 2004; Raver and Knitzer 2002), and children's progress in one area of readiness has been shown to predict readiness in other areas (see, for example, Hampton 1999; Ladd and Coleman 1997; Tramontana et al. 1988). Progress in a few targeted areas may therefore produce positive spillover into other, nontargeted areas of development (Bronfenbrenner 1979 and 1986; Zaslow et al. 1995). On the other hand, as a direct consequence of focusing an enhancement on one area of children's development, the enhancement may focus less on other areas of development (relative to the regular Head Start control group), leading to negative spillover (or "displacement effects") into nontargeted areas of development. Thus, an evaluation strategy should consider including measures of school readiness that may not be targeted directly by a particular Head Start enhancement, but that theory or practice suggests may nevertheless be affected by the enhancement.

- **Measuring Positive and Negative Child Outcomes.** Outcome measures that cover all aspects of a given child development construct must be included. If they are not, it will not be possible to detect both improvements in positive outcomes *and* decreases in problem outcomes (favorable impacts) and decreases in positive outcomes *and* increases in problem outcomes (unfavorable impacts).

- **Measuring Narrow and Broad Constructs/Outcomes.** Child outcomes targeted by Head Start enhancements can be narrow, broad, or very broad. The Head Start Child Outcomes Framework delineates narrow constructs ("indicators," such as confidence), broad constructs ("domain elements," such as self-concept), and very broad constructs ("domains," such as social and emotional development). Evaluators must clearly articulate not only which domain or domains of child development are likely to be distinctly affected by the enhancement over and above "regular" Head Start, but which constructs within a developmental domain (for example, at the domain element or indicator level) are likely to show impacts. The more narrow the construct expected to be affected by the enhancement, the more fine-grained or detailed the measure must be. For example, an enhancement that explicitly targets improving children's concentration will need a measure of "concentration" that is sensitive to the Head Start enhancement. A more global measure of "approaches to learning"—of which concentration is but one key ingredient—may or may not be sufficiently sensitive to, or distinctly affected by, the Head Start enhancement (relative to a regular Head Start control group). Of course, constraints relating to cost and respondent burden play a key role in selecting a measurement strategy. The fewer the resources, the more the evaluation may have to focus on measuring only those child outcomes—positive or negative, narrow or broad—most directly targeted by the Head Start enhancement.

Evaluators also face considerations regarding the mode of measurement as they select child outcome measures. Specifically, they must determine the best data collection mode for a given outcome, the cost of collecting the data using the preferred mode, and whether measurement trade-offs exist such that a less preferred mode might be chosen instead of a more-costly preferred one. Some measurement modes are not interchangeable. For example, because some outcomes lack valid, reliable parent report measures, another mode must be used. In addition, the selected mode may constrain the types of outcomes that can be measured.

The timing of measurement must be considered when selecting child outcome measures. The frequency of measurement will affect costs, and measurement duration (with one or more follow-up periods after Head Start completion) will influence the use of particular measures. Some child outcome measures (for example, the PPVT-III, the Woodcock-Johnson Tests of Achievement, and the Child Behavior Checklist) are suitable for use with both preschool children and with elementary school-age children. In other cases, it may be necessary to use measures tapping newly relevant developmental or school performance constructs (for example, *actual* reading ability, rather than measures of pre-reading skills and knowledge) in order to collect data on relevant child outcomes during the elementary school years. Deciding when to measure child outcomes will depend largely on (1) the particular enhancement being implemented, (2) the theory explaining which aspects of child functioning the enhancement is likely to affect (in the short and the long run), and (3) the value placed on including measures of known predictors of later school success or failure (regardless of whether the enhancement explicitly targets them).

During the development stage, evaluators will have to integrate decisions about the type of child outcome measures (targeted/nontargeted, positive/negative, and narrow/broad), the timing of measurement, and the measurement mode to develop the outcome measurement framework. If suitable outcome measures that are sensitive to the effects of the enhancement do not exist, then evaluators will have to conduct their measurement development work during this stage, before beginning the evaluation.

C. ACTIVITIES AND DURATION OF STAGE 1

Members of the Head Start community currently are practicing numerous enhancement strategies, and additional ideas undoubtedly are percolating each day. Because these ideas and strategies are at different points along the continuum of achieving the goals of a development stage, the specific activities and duration of a Stage 1 will vary with the enhancement strategy under consideration. In general, we expect that potential enhancement strategies may be at one of four levels in achieving Stage 1 goals. The levels are not necessarily mutually exclusive, and we expect that many enhancements would be placed in one or more of them.

- **Level 1: Defining the Enhancement.** Some enhancements may be in the very early development stages, with little more than a good idea based on a potentially strong intervention theory. These enhancements still are detailing key components, defining how the enhancement should function at full implementation, and articulating the theory of change.
- **Level 2: Documenting Implementation.** These enhancement strategies will lack clarity in that they have not completely documented the theory of change and/or the implementation process. This category will include enhancements that are little more than good ideas that have yet to be implemented by any program, or that are practiced by one or a handful of programs managing mostly by intuition. It also will include enhancements that have some documentation, but not at a level necessary for effective replication.
- **Level 3: Developing Measures of Fidelity and Outcomes.** Some enhancement strategies may have strong documentation and may have been adopted by multiple programs, but the measurement work to select or develop measures to assess the quality of implementation and the fidelity to enhancement has not occurred or is incomplete. Programs implementing these enhancements may be using available data to assess progress but have not developed standardized measurement for use over time or across programs to assess the success of implementation and/or replication.
- **Level 4: Replicating the Enhancement in a Variety of Settings.** Documentation has been developed and measures for quality of implementation and fidelity to enhancement exist, but neither are being used to assess implementation of the enhancement in diverse sites. Enhancements implemented in fairly homogeneous settings may need expanded, measured

replication that will provide critical feedback in refining documentation and measurement work to support future replication in a variety of settings.

We expect enhancement strategies that are at Level 1 (defining the enhancement) to require between one and one-half to three years to achieve the goals of Stage 1. These enhancements will benefit from a planning phase devoted to defining the details of the key components, and to documenting the theory of change. This early planning phase may last six to nine months. Another six to nine months may be necessary to refine documentation, and to develop and test measures through implementation in a number of early sites. Additional replication might be necessary for testing implementation and measurement in a variety of settings. Depending on the documentation's level of thoroughness and clarity and the diversity of early implementation settings, enhancements at Level 2 (documenting implementation) may take one to two years to achieve the goals of Stage 1. Enhancements at Level 3 (developing measures) may require one to one and one-half years to complete measurement work and to test replication. Those at Level 4 (replication) may take as little as six months or as long as one year to assess replication on a small scale before achieving Stage 1 goals.

Enhancements may be considered for a small-scale evaluation (Stage 2) when they have a well-defined theory of change; clear, thorough documentation of implementation; and a sound measurement framework that has been tested on a small, diverse scale. Implementation and outcome studies conducted during Stage 1 will determine the extent to which these criteria are met.

1. Implementation Studies

During the development stage, implementation studies will help to refine documentation, and to assess replication. The focus of an implementation study will depend on where the enhancement strategy lies along the spectrum of the four levels of achieving Stage 1 goals.

Levels 1, 2, and 3. The goals of an implementation study for enhancements at the first three levels will be explorative. Their purpose will be to determine the details of the implementation process and the lessons that will help to produce clear, thorough documentation, and to formulate quality and fidelity measures. The key research questions for these initial implementation studies are the following: (1) What are the processes of implementation, including initial training and ongoing technical assistance? and (2) What lessons can be drawn from the program's implementation experiences?

Collecting information about lessons from a program's early implementation experiences is especially important during this early stage of evaluation. Conducting interviews and focus groups with staff and technical assistance providers can identify implementation challenges and strategies that have the potential to resolve those challenges. Similarly, staff can provide important information about the usefulness of their training and support, as well as about aspects of implementation in which they need additional support. This information can be used to refine and strengthen implementation plans and strategies

for use by other programs, and to determine the types and intensity of T/TA that staff must receive if they are to do a good job of implementing the enhancement strategy.

Level 4. Level 4 enhancements will require an implementation study that is more evaluative in assessing replication and the effectiveness of quality of implementation and fidelity measures. This type of study may look very similar to an initial explorative study, and it will contribute to the feedback cycle for continuous program improvement and refinement of documentation. However, it also will include the use and testing of the specific quality and fidelity measures under development. The following two research questions will be added to the initial implementation studies' key research questions: (1) What is the quality of implementation? and (2) What is the degree of fidelity to the enhancement model?

Measurement experimentation. The Stage 1 development phase for new fidelity measures can allow for exploration of alternative measurement methods that can cost-effectively provide accurate measurement for the subsequent stages. For example, in the development stage, a fidelity measure that requires direct observation of the classroom at multiple points in time may be used in tandem with a staff survey about attitudes and beliefs. If the two measures demonstrate high correlations with each other, it may be possible to use only the staff survey during the later evaluation stages, when cost parameters may be more restrictive. Alternatively, experimentation in Stage 1 may determine that a short version of a detailed observation tool may function equally or nearly as well as the longer version. Information collected through an evaluative implementation study can inform these decisions.

Data Collection Methods. Table III.1 provides a detailed set of implementation study topics that could be included in both explorative and evaluative implementation studies. The table indicates the implementation study questions (about process, quality, and lessons learned) that each topic addresses, along with their possible data collection methods. Depending on available resources, one or more data collection methods could be used. The use of multiple methods during the development stage would enable evaluators to collect more-detailed information about key topics from a variety of respondents, and to triangulate findings across several data sources. For example, teachers may report in a staff survey that, after initial training, they still did not feel adequately prepared to implement the enhanced services. During a focus group with the teachers, evaluators could explore the aspects of training that appeared to be inadequate and could seek ideas about ways to improve the training. Interviews with technical assistance providers could yield additional information about teachers' training needs, perhaps based on perceptions of the teachers' educational levels and learning styles.

Table III.1: Potential Implementation Study Topics for a Study of Head Start Quality Enhancements

	Implementation Study Questions			Data Collection Methods					
Implementation Study Topics	Implementation Processes	Implementation Quality	Lessons Learned	Direct Observation of Service Delivery	Staff Survey	Parent Survey	Program Records	Semistructured Interviews with Staff or T/TA Providers	Focus Groups with Staff or Parents
Program Characteristics									
Where is the program located?	X				X		X	X	
What is the program's size?	X				X		X	X	
What type of agency operates the program?	X				X		X	X	
How many years has the agency participated in Head Start?	X				X		X	X	
Target Population									
How does the program decide which families receive enhanced services?	X	X			X		X	X	X
What are the characteristics of the families and children served?	X				X	X	X	X	X
Are services provided to children, parents, or both?	X	X			X	X	X	X	X
Staff Characteristics and Roles									
Who provides the enhanced services?	X				X			X	
What are their qualifications?	X	X			X			X	
Who supervises and supports these staff?	X	X			X			X	
Which staff participated in training?	X	X			X			X	
What is the rate of staff turnover?		X			X		X	X	
How well do staff understand the enhancement strategy?		X			X			X	X
Initial Staff Training									
What was the content of training?	X	X		X	X			X	X
What were the qualifications of trainers?	X	X			X			X	
What was the intensity and duration of training?	X	X		X	X			X	
How well did training prepare staff to provide the enhanced services?			X		X			X	X
How could training be improved?			X					X	X

X

Table III.1 *(continued)*

Implementation Study Topics	Implementation Study Questions			Data Collection Methods					
	Implementation Processes	Implementation Quality	Lessons Learned	Direct Observation of Service Delivery	Staff Survey	Parent Survey	Program Records	Semistructured Interviews with Staff or T/TA Providers	Focus Groups with Staff or Parents
Providing Enhanced Services									
What enhanced services are provided to parents and children?	X	X		X	X	X	X	X	X
How often are the services provided?	X	X			X	X	X	X	
What is the intensity of service delivery?	X	X		X	X	X		X	
How long are the services provided?	X	X			X	X	X	X	
Are staff able to provide enhanced services at the intended intensity and duration?		X	X			X	X	X	X
If not, why not?			X					X	X
Technical Assistance and Support									
What support do staff receive in providing enhanced services?	X	X			X			X	X
Do supervisors observe service delivery and provide feedback to staff? How often?	X	X			X		X	X	X
Does the program receive regular technical assistance? How often?	X	X					X	X	X
What topics have been covered?	X	X						X	X
How helpful is the technical assistance?			X					X	X
What other types of support do staff need?			X					X	X
Lessons Learned									
Which aspects of implementation have worked well?			X		X	X		X	X
What factors facilitate implementation of the enhanced services?			X		X			X	X
What challenges have staff experienced in providing enhanced services?			X		X			X	X
What strategies have staff used to overcome the challenges?			X					X	X
What are the staffs' views on the enhancement strategy?			X		X			X	X
What are families' views on the enhancement strategy?			X			X			X
How could the enhanced services be improved?			X			X		X	X

2. Outcomes Studies

An initial outcomes study in the development stage will begin to lay the framework for measuring intermediate and child outcomes in a rigorous evaluation. These studies will examine selected intermediate and child outcomes that are expected to change based on implementation of the enhancement. They are likely to use a standard pre/post methodology to measure the outcomes early in implementation (for example, during the fall semester or at program entry), and again at one or more points after implementation (for example, during the spring semester or at the end of the academic year). Although it will not be possible to attribute changes in the intermediate and child outcomes to the presence of the enhancement (or to its absence, if a comparison site design is used), the findings may be suggestive of an enhancement's potential for success. If numerous sites can be included in the outcomes study component, the analysis also should examine the correlations among the levels of the quality of implementation, fidelity to the enhancement model, and changes in intermediate and child outcomes. In addition, the study should examine how intermediate outcomes are correlated with child outcomes.

Measurement experimentation. During this stage, when there are few implementation sites and costs can be contained, evaluators should consider using more-intensive methods and several measures to examine, test, and select appropriate outcome measures (for both intermediate and child outcomes). Their goal should be to identify the soundest, most efficient, and, ideally, least costly measures for use in the subsequent, larger stages of evaluation. For example, to measure child outcomes, evaluators may choose to use both a teacher-child interaction measure and a teacher-report measure to tap the children's attention, engagement in classroom activities, and positive teacher interactions. If the two types of measures demonstrate high correlations with each other and with implementation of the enhancement, the less burdensome, less expensive teacher-report measure may be a good alternative for use in Stages 2 and 3. Furthermore, if new measures have been developed or if existing measures have been substantially revised for the enhancement, Stage 1 provides the opportunity to test the sensitivity and pyschometric properties of "created" measures against existing normed measures.

3. Products from Stage 1

At the end of Stage 1, a number of products should exist that fall into the categories of (1) implementation products and, (2) evaluation products. Implementation products are the documentation that will prove useful to other programs that decide to implement the enhancement strategy. Evaluation products should be designed to contribute to the knowledge of enhancements throughout the Head Start community, so that program administrators can make informed decisions about enhancement choices, and evaluators can make informed decisions about which enhancements merit more-rigorous study. Evaluation products should be disseminated more broadly than the detailed implementation ones.

Implementation Products

- **Implementation Manual(s).** The implementation manual should present the theory of the enhancement (by describing what it is and what it has been designed to accomplish), and a step-by-step guide to implementation from startup through ongoing activities. The implementation manual also should contain a measurement section that details the data sources and measures selected for assessing the quality of implementation, assessing the fidelity to enhancement, and examining intermediate outcomes. This section should include background information about each measure (how and why it was selected and/or developed), specifics about reliability and validity (psychometric properties, inter-rater reliability, and internal consistency), and the actual measurement instruments and data sources.
- **T/TA Guide.** The T/TA guide may be either a separate document or part of the implementation manual. Regardless of the method of presentation, the guide must describe the details of the content, intensity, duration, and quality of both the initial and ongoing T/TA.

Evaluation Products. The evaluation products will report findings from the implementation and outcomes studies, either separately or combined.

- **Implementation study findings.** Findings from the implementation studies will summarize the implementation process and will discuss the challenges to implementation and lessons for replication. Implementation study findings should include a discussion of the overall quality of implementation and the fidelity to the enhancement model in both the original site and in new sites.
- **Outcome study findings.** Findings from the outcomes study should report on intermediate outcomes, the outcomes' association with the child outcomes of interest, and correlation of the outcomes to the level of quality of implementation and fidelity to enhancement measured in the study sites.

The evaluation product (or products) should also make recommendations for continued research involving the specific enhancement strategy, such as the potential for rigorous testing based on changes in intermediate and child outcomes, or on potential changes in the enhancement strategy that should be explored before beginning more-rigorous testing.

4. Entities Involved in Stage 1 Activities

The activities in the initial development stage would be carried out by a combination of Head Start program staff and/or enhancement developers and outside researchers, possibly connected with a university partner or research firm.

- **Enhancement developers.** In partnership with Head Start programs, enhancement developers are responsible for defining the key elements of an enhancement, the details of each element, and the methods of integration into

the Head Start program. Enhancement developers also should formulate a clear motivation and theory of change. Some developers, such as curriculum design experts, may contribute to the selection or development of measures to test fidelity to the enhancement model.

- **Head Start program staff.** Program staff may function as enhancement developers and, as such, would perform their roles either on their own or with the assistance of researchers. Ideally, Head Start program staff will be involved in the feedback cycles that contribute to implementation and documentation refinement based on their experiences with early rounds of implementation.
- **Research partners.** Research partners will play a substantial role in measurement development by explicitly defining the theory of change, and by selecting and/or developing measures for each measurement domain. Researchers will also develop and conduct implementation studies in Stage 1, the findings of which would contribute to enhancement refinement and measurement testing. In addition, research partners will formulate and conduct outcomes studies in Stage 1 designed to measure the magnitude of any changes in intermediate and child outcomes that could be suggestive of effects of the enhancement. All of these activities will be stronger with strategic input from and discussion with Head Start program staff.

D. EXAMPLES OF STAGE 1 ACTIVITIES FOR THREE ENHANCEMENTS

Many enhancements to Head Start programs begin with a planning phase that might resemble the development stage discussed in this chapter. In practice, however, planning phases can look quite different from one initiative to another. The staged approach to evaluation detailed in this report adds guidance and prescription to enhancement development by viewing numerous potential enhancements as part of a holistic research agenda, rather than as many independent initiatives, each following its own course. Although enhancements may develop through different methods and along different timeframes, the development stage adds a degree of uniformity by specifying the types of activities that must occur and the products that must be produced during this early stage. By the end of Stage 1, each enhancement will have the elements in place that are critical for replication and for progression to the next evaluation stage.

The types of activities for Stage 1 that we have detailed are being encouraged and supported through Head Start Innovation and Improvement (I&I) grants. These grants, totaling $2.9 million, have been awarded to a range of national, state, and local organizations to support one-year planning phases for strategies that have the potential to strengthen Head Start programs and services. In the remainder of this chapter, we use three enhancement ideas developed by I&I grantees to discuss the type and flow of activities and the development of measurement frameworks that could be expected to occur in a development stage as outlined in this chapter. Note that these ideas are the basis for our examples and have been elaborated upon to illustrate the steps necessary for Stage I, but the activities we discuss may not actually be in process in the exact manner they are presented here.

1. Family Foundations Project, University of Arkansas for Medical Sciences (Little Rock, AR)

The University of Arkansas for Medical Sciences is conducting pilot-phase work under an I&I program grant to fully define the key components of the Family Foundations Project (FFP) for eventual implementation in Head Start programs throughout Pulaski County, Arkansas. The FFP has been designed to build family resources and supports by connecting targeted Head Start parents with job opportunities and community support services through a Work Program, by improving parental skills and practices through Brief Parenting Interventions, and by increasing fathers' involvement through a Fatherhood Initiative. In this section, we describe the steps to be taken to define the enhancement, which is in the very early stages of development; to refine and document implementation; and to develop measures of implementation, fidelity, parent, and child outcomes.

a. Level 1: Defining the Enhancement

Motivation and goals. The University of Arkansas for Medical Sciences initiated the FFP in response to the specific vulnerabilities of Early Head Start and Head Start families in Pulaski County. The county has higher-than-average teenage birth and child poverty rates (38 births per 1,000 teens and 21.3 percent poverty, respectively). Eighty percent of the children in the targeted Early Head Start and Head Start programs live in single-parent families. Many county residents have poor access to health care; families are at risk for financial instability and experience high unemployment. In addition, 18 percent of pregnant women tested positive for illicit drug use across 16 public health clinics and 2 private clinics in the county. Combined, these factors explain the limited resources of Early Head Start/Head Start families and the low level of the parents' involvement in the educational lives of their children. FFP developers believe that increasing the resources available to Head Start families through the Work Program and increasing the engagement of families in Head Start will increase the parents' availability to, engagement with, and responsibility for their children, thereby minimizing the potential effect of multiple risk factors on child outcomes.

The goals of the FFP are to increase parents' capacity to provide financial support and resources for their children, improve the quality of parent-child interactions and relationships, and increase fathers' involvement with their children and with the Head Start program. In FFP sites at full implementation, we would expect to observe parents of Head Start children, particularly those in specific subgroups (fathers, mothers in substance abuse treatment programs, and teenage parents), connecting with community support services and jobs through the Work Program's job development, case management, and job coaching services, and participating in Head Start's Brief Parenting Interventions.

Enhancement components. At the time of the University of Arkansas for Medical Sciences I&I proposal, the FFP was little more than a conceptual framework. The components of the FFP—the Work Program, the Brief Parenting Interventions, and the Fatherhood Initiative intended to integrate noncustodial fathers into FFP activities—had been outlined, but the proposal for the planning grant focused on the core development work necessary to completely define this enhancement. For example, the details of the

Work Program still must be finalized. These details relate to where the Work Program will be housed, how it will be staffed (the composition and qualifications of program staff), which services will be provided and how, and how the program's funding and resources will be sustained. A Steering Committee will accomplish part of this work by establishing workgroups to target specific development areas (for example, adult education, corporate/marketing, and relationships with employers). As with the Work Program, the Brief Parenting Interventions and the Fatherhood Initiative also need specification. To help them to develop the Brief Parenting Interventions' content and the Fatherhood Initiative's methods of parent engagement, FFP developers have proposed conducting focus groups during the planning period with the targeted at-risk parent populations (fathers, mothers in substance abuse treatment programs, and teenage parents); the goal of the focus groups will be to identify parenting beliefs, norms, values, and areas of concern. In addition, the Fatherhood Initiative still has to identify and define father-centric activities that can be incorporated into the Work Program and Brief Parenting Interventions, determine whether father-only activities also may be provided, and identify those activities.

Theory of change. In addition to defining the components of the enhancement, enhancement developers must formulate an initial framework for the theory of change as part of the overall definition of the enhancement. The measurement work takes place later in the development stage and will involve researchers who will thoroughly define and specify a logic model for an evaluation; during this early phase, however, FFP enhancement developers should define their theory's basic elements about how the FFP could improve child outcomes for Head Start children. For example, FFP developers have suggested that Head Start staff and the highest-risk Head Start families are the primary direct beneficiaries of FFP. According to the theory of change, these groups may therefore be the *targets of change*. Specifically, the FFP is intended to enhance staff skills in parenting education and staff resources for use in connecting families with jobs and community services. In addition, FFP is to target high-risk Head Start parents in order to increase the parents' engagement in parenting workshops, improve their links with community and job resources, and increase their involvement in the educational lives of their children in Head Start.

After the targets of change have been identified, FFP developers should consider both the *intermediate and child outcomes* that they believe FFP will affect. To identify intermediate outcomes, the developers will find it helpful to examine the specific targets of change. FFP's intermediate outcomes should encompass the program's goals for Head Start parents who participate in FFP, such as increased job entry and retention rates, income gains, increases in social support networks, increases in parenting skills and knowledge, and improved parent-child interactions. Child outcomes are not specifically targeted by FFP but are likely to result from the intermediate changes in families. For example, increased father involvement can lead to improved child cognitive and social outcomes (Fagan 2000; Nord and West 2001; Nord, Brimhall, and West 1997; Parke et al. 1992; Tamis-LeMonda et al. 2004), and higher family income is associated with larger child vocabulary and better school achievement (Duncan and Brooks-Gunn 1997; Hart and Risley 1995; Future of Children 2005). The FFP intervention seeks to influence the child's home environment, but not to achieve specific changes in that environment. Because the intervention does not include

components that directly focus on the child, changes in child outcomes depend on the extent to which the FFP intervention changes the child's home environment.

b. Level 2: Documenting Implementation

The process of defining the FFP enhancement more fully will provide a framework for the comprehensive documentation necessary to effectively replicate and evaluate the FFP. However, the necessary additions and refinements are likely to become evident only with the implementation of FFP in the Head Start programs throughout Pulaski County.

A development stage for FFP will have to produce a number of "implementation" products to fully describe both the components of FFP and the choices and decisions associated with implementation of each component. These products may or may not be stand-alone documents; we discuss them separately for the purpose of establishing the importance of each one. Ultimately, it will be necessary to integrate these pieces in order to present the holistic approach of FFP as a Head Start program enhancement.

Implementation Manual: FFP developers will have to produce an implementation manual explaining how to establish and administer the FFP by detailing implementation of the three key components—the Work Program, Brief Parenting Interventions, and the Fatherhood Initiative. The Work Program component of FFP may pose specific challenges when replicated among Head Start programs more broadly and will therefore need thorough documentation of the different avenues that programs may pursue to ensure full implementation. An implementation manual for the FFP should include the following elements:

- **Community resource assessment.** Conducting a community resource assessment may be the first step necessary for new Head Start programs to determine whether the Work Program can be implemented using existing community resources, or whether this component needs full development driven by the Head Start program itself. The Work Program calls for case managers and job developers who can help Head Start parents to connect with jobs, support services, and other social services. Hiring, training, and housing these staff is likely to be beyond the resource capacity of many individual centers and/or Head Start grantees. The implementation manual will have to address critical start-up and sustainability issues, such as identifying and securing funding from state and local government, foundation, or private sector sources, and identifying case management and job development resources within the local area that can be linked with and, possibly, adapted to meet the criteria of the Work Program.
- **Staffing and services of the Work Program.** The manual should address such topics as the Work Program staff's qualifications, critical staff training elements particular to serving Head Start parents, a suggested ratio of case managers/job developers to the number of Head Start parents served, guidance on identifying and linking with potential employers and service agencies, and details about the

types of job development resources that the Work Program should provide (for example, job listings, job skills workshops, and assistance with resumes).

- **Staffing and services of the Brief Parenting Interventions.** This section should provide detailed information about such items as the qualifications of staff who will facilitate Brief Parenting Interventions, methods to increase participation in Brief Parenting Interventions, specific lesson and facilitation plans for specific Brief Parenting Interventions sessions, decisions about the timing and frequency of the Brief Parenting Intervention sessions, the suggested numbers of participants in the sessions and/or ways to adapt sessions to various group sizes, and identification and linkage of parents with parenting education resources in the community,
- **Staffing and services of the Fatherhood Initiative.** This guidance should include details about identifying the appropriate Head Start staff to lead and implement the Fatherhood Initiative, training elements for staff, methods of engaging fathers in the Work Program and Brief Parenting Interventions, and developing and implementing father-centric activities in FFP components and/or throughout the Head Start program.
- **Potential for replication.** FFP has been developed to serve a specific county in Arkansas, but it probably has broader applicability. Enhancement developers, with assistance from evaluators, should consider and document how the program parameters may vary by community or by particular Head Start program in order to provide thoughtful, comprehensive guidance that can support replication across a diverse set of programs.
- **Accessing T/TA:** Implementation guidance should detail any on-going technical assistance that will be made available from the enhancement developers or, possibly, from Head Start regions and/or evaluators for the different components of the FFP. For example, FFP developers plan to have web-based training materials available for the Brief Parenting Interventions.

T/TA Guide for Brief Parenting Interventions. In addition to an implementation manual, a T/TA guide will have to be developed for the FFP's Brief Parenting Interventions. This guide will provide detailed information for the trainers who will train Head Start program staff. Guidance for trainers will have to include details about the necessary qualifications of the trainers themselves, the Brief Parenting Interventions parenting learning strands (training content), which Head Start staff are to be trained, culturally appropriate methods of Brief Parenting Intervention facilitation that may vary with the predominant language/culture of Head Start parents and children, and the intensity and duration of Brief Parenting Interventions training for staff. As with any training, different options should be presented so that trainers may accommodate the needs of different Head Start programs.

c. Level 3: Developing Measures

After the enhancement has been clearly defined, FFP developers and evaluators can begin creating the measurement framework. This work can be accomplished simultaneously

with documentation creation and refinement. After this step has been completed, the measures may undergo early testing in the FFP implementation sites in Pulaski County.

Quality of implementation. Measures of the quality of implementation for FFP should focus on how well the sites meet the parameters of implementation as established by the developers. Although the routes to FFP implementation may differ among the sites, there are likely to be critical standards for the qualifications of staff, the breadth of the network of potential employers and service providers for referrals, the quality of staff training sessions, and the quality of the activities of the Fatherhood Initiative (Table III.2). Measures of the quality of implementation typically can be obtained and observed by conducting a qualitative study of site activities. In the case of FFP, it would be important to observe initial staff training sessions so as to gauge content coverage, the level of staff participation, and the intensity of that participation. Other measures could be obtained through observation and semistructured interviews with program staff, administrators, and community partners. Thresholds for the quality of implementation measures must be set with care, particularly in the case of those that measure the breadth of the potential employer/community support network and methods of engaging targeted Head Start parents. Because the FFP intervention reaches children only through the work of engaging and supporting parents, the performance bar in these areas will have to be set rather high if the intervention is to produce its intended effects on child outcomes.

Fidelity to enhancement. Fidelity to enhancement measures should capture whether the FFP in each site functions as intended, and whether the FFP's targets of change accomplish what they are intended to do. For example, the staff-parent interactions should have the content, frequency, and duration expected with full FFP implementation; Work Program staff should make the link to jobs and community resources for Head Start families; the appropriate program staff should have a high level of competence in parenting education; Brief Parenting Interventions sessions should address parents' needs for content and frequency; and the targeted Head Start families should participate in the Work Program and the Brief Parenting Interventions (Table III.2).

In many cases, fidelity measures will have to be created for the specific enhancement, and they will have to rely on direct observation. It may be possible to adapt existing measures for some of the fidelity measures, or to use existing principles when creating them. For example, when developing measures of the quality of interactions between parents and Work Program case managers or job developers, FFP developers and evaluators might refer to studies of employment programs that serve disadvantaged populations for guidance on which key elements to assess, and how to conduct those assessments. If systems already have been developed to track information specific to the Work Program and to engage parents in Brief Parenting Interventions sessions, it may even be possible to capture some FFP fidelity measures through administrative data. For example, evaluators should be able to use administrative data to determine the number and quality of links to jobs and community resources for targeted Head Start parents and to document the degree to which targeted parents participate in FFP components.

Table III.2 Potential Measurement Options for the Family Foundations Project

	Concept Domain	Potential Measurement Method Existing Tool
Quality of Implementation	Work Program and Brief Parenting Interventions staff meet qualifications for positions.	Resume review; interviews with staff during site visits
	Resources to support the Work Program and/or connections built with necessary community partners are in place to effectively implement and sustain the Work Program.	Interviews during site visits
	The work program has established effective links with the intended number/variety of potential employers and community service providers.	File/data review; interviews with staff, employers, and community service providers during site visits
	The ratio of Work Program staff to Head Start parents meets parameters.	Observation during site visits
	Brief Parenting Interventions staff training sessions meet quality parameters for qualifications of trainer/facilitator, frequency, and duration.	Observation and interviews during site visits
	Multiple methods of engaging targeted parent populations are in place.	Observation and interviews with staff during site visits
	Fatherhood Initiative activities meet quality parameters for staff qualifications, content, frequency, and duration.	Observation and interviews during site visits
Fidelity to Enhancement	Interactions between Work Program staff (case managers and for job developers) and Head Start parents meet quality parameters for content, frequency, and duration.	Observation during site visits
	The intended number and quality of job links for Head Start families have been made (staff skills).	File/data review; focus groups with Head Start parents and employers
	The intended number and quality of links between Head Start families and community services/resources have been made (staff skills).	File/data review; focus groups with Head Start parents and service providers
	Head Start program staff has competency in parenting education.	Post-test of staff trainees tailored to specific Brief Parenting Interventions concepts (must have knowledge of a specified percentage of training content)
	Brief Parenting Interventions sessions for parents meet quality parameters for content, frequency, and duration.	Observation during site visits
	The overall level of Head Start parent participation in the Work Program and Brief Parenting Interventions is as intended.	Administrative data observation
	The level of participation in the Work Program and Brief Parenting Interventions by targeted parent groups (fathers, mothers in treatment programs, and teenage parents) is as intended.	Administrative data observation

Table III.2 *(continued)*

	Concept Domain	Potential Measurement Method Existing Tool
Intermediate Outcomes	Employment entry, retention, and income of Head Start parents.	Program administrative data UI data
	Level of involvement in Head Start program.	Parent Involvement and Satisfaction with Head Start (FACES Spring 2003)
	Breadth of social support network.	Carolina Parent Social Support Scale
	Parenting knowledge and skills of Brief Parenting Interventions participants.	Post-test of parent participants tailored to specific Brief Parenting Interventions concepts (must have knowledge of a specified percentage of content)
	Increased frequency and duration of parent-child reading time.	FACES Parent Interview "Activities with Your Child"
	Parenting stress.	Third Edition
	Household conflict.	FES Conflict Items
	Father presence and involvement with child.	Father Activities with Child—Early Head Start Father Study Measures
	Increased parenting support and warmth; decreased detachment and intrusiveness.	Parent-Child Interaction Task (NICHD) Parent-Child Relationship Scale (NICHD)
	Child activity level (minutes in brisk activity and/or television viewing, other media).	Parent report; activity diaries
	Number of meals from fast food restaurants.	Parent report
Child Outcomes	Child vocabulary.	PPVT-III
	Book knowledge (interest in books and reading-related activities, such as listening to and retelling stories and pretending to read).	- Story and Print Concepts (FACES) - COR Language and Literacy Scales
	Print knowledge (recognition of words as a unit of print, increased ability to associate spoken words with written words, and increased awareness of the mechanics of reading).	- Story and Print Concepts (FACES) - Pre-CTOPPP Print Awareness Subtest - Woodcock-Johnson III Tests of Achievement Letter-Word Identification Test - TERA-3

Table III.2 *(continued)*

Concept Domain	Potential Measurement Method Existing Tool
Alphabet knowledge, phonological processing, early reading.	- Letter Naming Task (NRS) - Pre-CTOPPP - TERA-3 - COR Language and Literacy Scales - Woodcock-Johnson III Tests of Achievement Letter-Word Identification Test
Behavioral problems.	- Child Behavior Problem Index (NICHD) - Behavior Problems Scale (or Classroom Conduct Problems) (NICHD) - Child Behavior Checklist (NICHD)
Prosocial behavior.	- Howes Peer Play Observation Scale (FACES) - Parent-Child Interaction Task (FACES) - Social Skills Scale of the Social Skills Rating System - Social Competence Subscale of the Social Competence and Behavior Evaluation
Child weight for height.	Height, weight, BMI
Overall child physical health.	Consult large-scale health studies of low-income children

BMI = Body mass index; COR = Child Observation Record; FACES = Family as Child Experiences Survey; FES = Family Environment Scale; FFP = Family Foundations Project; NICHD = National Institute of Child Health and Human Development; NRS = National Reporting System; PPVT-III; Peabody Picture Vocabulary Test-Third Edition; Pre-CTOPPP = Preschool Comprehensive Test of Phonological and Print Processing; PSI = Parenting Stress Index; TERA-3 = Test of Early Reading Ability-Third Edition

As before, FFP developers will have to set the thresholds rather high, particularly for participation and for frequency of intervention activities. Meeting the parental participation thresholds in FFP components may be a challenge, but it will be critical when testing the intervention theory. Even if the implementation of FFP is considered "full" and of high quality, it is possible that participation will be low given that many government and social service programs have difficulty engaging the types of individuals targeted by FFP. Factors exogenous to FFP that influence participation may have a stronger, detrimental effect. If FFP sites cannot engage parents to a sufficiently high degree, the intervention will have only a limited potential to produce the size of intended effects on child outcomes indicative of positive impacts. Similarly, if the frequency (or dosage) of exposure to Brief Parenting Interventions is too low, it is unlikely that child impacts could be detected.

Intermediate and child outcomes. We presented possible intermediate and child outcomes in a previous section. It should be possible to collect some of these intermediate outcomes, specifically the ones for the Work Program, with relative ease and at limited cost by using administrative data. For example, employment and income outcomes of targeted parents could be measured using employment placement data collected by Work Program staff and data from the state's unemployment insurance records. Other intermediate outcome measures for the Work Program could use or adapt existing tools. For example, the Carolina Parent Social Support Scale may be a measurement option for assessing the breadth and depth of a parent's social support network prior to and following participation in FFP.

d. Examples of Brief Parenting Intervention Models and Measures

Without specific definition to the components of FFP, it is difficult to specify additional intermediate and child outcomes. For the sake of example, we propose three Brief Parenting Interventions modules that might be universal among individual programs that are targeting noncustodial fathers, mothers with substance abuse problems, and teenage parents: reading with children; family relationships; and improving nutrition and health. Using these modules, we can propose intermediate and child outcomes and their corresponding considerations for use in an evaluation of FFP. Measurement tools exist to measure many of these concepts. In Table III.2, we make specific suggestions based on these examples, but other measures may be considered as well.

Module 1: Reading with Children. The first Brief Parenting Intervention module would focus on reading with children. The module's goals would be to establish reading as a daily routine, introduce dialogic reading (in which a parent asks questions to further engage the child with the book), and demonstrate effective read-aloud techniques to parents. The intermediate outcome of interest would be increased frequency and duration of parent-child reading time. To measure this outcome, the FFP evaluation could adapt the "Activities with Your Child" section of the FACES parent interview, in which parents are asked six distinct questions pertaining to their support of and involvement with literacy activities. The child outcomes could include increased vocabulary, increased book and print knowledge, and increased letter identification (or alphabet knowledge).

To measure these child outcomes, FFP evaluators probably would be able to rely on existing measures. Measurement of these child outcomes typically is conducted through direct assessment. Although a number of measures on vocabulary are available and have been used with Head Start children, the PPVT-III is one of the most commonly used measures of children's vocabulary, and it has strong psychometric properties. The PPVT-III has been used with Head Start populations; for FACES, trained paraprofessionals administered and scored a short version of the PPVT-III in about 10 minutes. The PPVT-III also can be used with elementary school-age (and even older) children, which could prove useful in an evaluation of FFP, given that child impacts may occur after the Head Start year.

Measures for assessing book knowledge, print knowledge, and letter identification are available; many of them capture elements of all three of those concepts. Some of these measures contain only a single relevant item, some have subscales, and still others are more comprehensive. For example, the Child Observation Record (COR) contains two items on children's knowledge and appreciation of books, and one item on children's knowledge of letters and numbers. The most in-depth measure of book knowledge and appreciation is the Story and Print Concepts measure, which also is available in Spanish. However, information from FACES on this measure's psychometric properties suggests less-than-optimal reliability and mixed evidence of its validity with a Head Start population. For print awareness and concepts, the Preschool Comprehensive Test of Phonological and Print Processing (Pre-CTOPPP) includes a "print awareness task" that also contains a few items that tap this domain element. The Conventions Subtest of the Test of Early Reading Ability, Third Edition (TERA-3) is another in-depth assessment that has good reliability, and that can be used with elementary school-age children. (However, it is not available in Spanish.) The Letter-Word Identification Test of the Woodcock-Johnson III has good psychometric properties; it has been used with diverse populations, can be administered in eight minutes, and can be used with elementary school-age children.

Many of the same tools that measure print awareness also can be used to measure alphabet knowledge. The "print awareness task" of the Pre-CTOPPP taps the child's ability to identify letters. The Alphabet Subtest of the TERA-3 has good psychometric properties and takes about 10 minutes to administer and score. The English and Spanish versions of the Letter-Word Identification Test of the Woodcock-Johnson III can be used with children of all ages to measure alphabet knowledge. A measure specific to alphabet knowledge is the Letter Naming Task (and its Spanish counterpart, Nombrando Las Letras) that was developed by the Head Start Quality Research Centers and is being used in the Head Start National Reporting System. It is a psychometrically sound measure that takes only about five minutes to administer and score.

Module 2: Family relationships. The second module would focus on supporting both the mother-father relationship, and the parent-child relationship (both mother-child and father-child) with the goals of providing parents with constructive ways to engage each other in support of their children's development, reducing parent conflict, reducing parenting stress, and increasing positive communication between parents and their children. The expected focus on communication and positive social interactions would lead to the

intermediate outcomes of reduced parenting stress, reduced household conflict, and increased father presence and involvement with the child. Existing measurement tools may be adapted for an FFP evaluation or may guide development of more refined measures specific to FFP.

Intermediate outcome: reduced parenting stress. The Parenting Stress Index (PSI) covers four major domains (total stress, child domain, parent domain, and life stress) and can be used with parents of children between the ages of one month and 12 years. The parent domain would be most germane to capturing the intermediate outcome of reduced parenting stress that could result from a Brief Parenting Intervention focused on family relationships. This domain consists of seven subscales that measure competence, isolation, attachment, health, feeling of role restriction, depression, and spousal support. FFP evaluators may find that the PSI can be informative in developing a measure of parenting stress. The short form of the PSI could also be explored as an alternative. In addition, although the PSI can be administered and scored by staff who do not have formal training in psychology or social work, interpretation of PSI scores does require such training.

Intermediate outcome: reduced household conflict. The Family Environment Scale (FES) is a tool that might be used in the second module to measure household conflict. The FES measures the family's social environment and family functioning through 90 true-false items on a paper form that a parent completes. The FES's instructions are self-explanatory, no training is required for administration, and the tool is easy to score. It is available in a number of languages and was used in the impact study of Early Head Start.

Intermediate outcome: father involvement. The Early Head Start impact evaluation also can be informative as a source of measures for father presence and involvement with children. The study developed four factor scores that measure the frequency with which fathers, or father figures, engage in four types of activities: (1) caregiving activities, such as helping their children to brush their teeth or take baths; (2) social activities, such as taking children to visit friends or relatives and going to restaurants; (3) cognitive activities, such as reading or telling stories and singing to the children; and (4) physical play, ranging from calm activities, such as rolling a ball, to rough and tumble activities, such as chasing games. The Early Head Start father activity scores have good psychometric properties (internal consistency reliability ranging from .72 to .84) and can be administered and scored with relative ease.

Intermediate outcome: parenting support. Increased parenting support and warmth and decreased detachment and intrusiveness are other possible intermediate outcomes of supporting the parent-child relationship. Even though a few measures of parent-child interactions exist, some measurement development work in this area probably would be necessary. The Parent-Child Interaction Task, used by the National Institute of Child Health and Human Development (NICHD), involves observation of a parent and child interacting during a semistructured, 15-minute play interaction. Ratings scales are used to assess the quality of the interaction, expressions of affect, and the child's emotional regulation with the mother in a potentially exciting or frustrating activity. This measure has good reliability, but validity specific to Head Start children has not been determined. A less intensive, but potentially less reliable measure, is the parent report (the Parent-Child Relationship Scale)

used by NICHD. This 15-item questionnaire assesses how warmly parents view their relationship with their children by asking parents to rate items on a five-point Likert-type scale.

Child outcomes for Module 2. The expected child outcomes would be reduced behavioral problems and increased prosocial behavior. ("Prosocial behavior" refers to a child's interest in and ability to develop friendships and positive relationships.) Because many existing measures capture some aspect of children's behavior and social relationships, FFP evaluators will have to make an informed assessment about what aspect (or aspects) of behavior are likely to be influenced by the Brief Parenting Interventions (or other components of FFP). Two potential broad measures of problem behaviors, both of which have been used in the FACES study, are the Behavior Problems Scale (BPS) and the Child Behavior Problem Index (CBPI). The BPS relies on teachers' ratings of children's behavior; the CBPI is a parent report based on 12 items of children's negative behaviors that are relatively common among preschool children. Both measures are applicable only to preschool children. For the FFP evaluation, however, it may be necessary to identify or develop a measure that can be used with elementary school-age children, as impacts on child outcomes may be expected to occur beyond the Head Start year. The Child Behavior Checklist contains a parent and a teacher rating component of social competence and problem behavior for children aged 4 to 18 years. Measuring prosocial behavior may overlap with measuring problem behavior, as positive social relationships require such traits as good self-control and cooperation. However, prosocial behavior generally is a broader construct than is social competence that includes such behaviors as children's willingness to talk with and accept guidance and directions from teachers, their ability to develop friendships, and their ability to express empathy and to care for others. Many measures of prosocial behavior require observational coding of interactions with peers (such as with the Howes Peer Play Observation Scale used in FACES), observational coding of interactions with parents (using the Parent-Child Interaction Task), or some other method to tap the dyadic nature of children's social relationships (such as with a teacher). Other measures constitute validated scales or subscales of children's social competence more broadly conceptualized on the basis of teacher or parent reports (for example, the 10-item social competence subscale of the Social Competence and Behavior Evaluation and the Social Skills Scale of the Social Skills Rating System).

Module 3: Improving nutrition and health. The third module would focus on nutrition and health with the goals of educating parents about how to establish food shopping and food preparation goals, sharing tips on stimulating preschoolers' interest in eating nutritious meals and snacks, and engaging parents in reading their children's satiety cues and increasing their children's activity levels. The intermediate outcomes for this module could be measurement of activity level, such as minutes per day in outdoor activities or hours per day viewing television, or, possibly, the number of meals per week from fast food restaurants. These intermediate outcomes could be collected through parents' reports, using questionnaires, interviews, or, possibly, activity diaries. Possible child outcomes of interest include child weight for height (or body mass index for children) and overall child physical health. Our review of potential child outcome measures conducted as part of a previous task in this study highlighted one study among recent large-scale studies and

program evaluations involving low-income families with preschool-age children that measured children's health status and practices. The NICHD Study of Early Child Care assesses children's height and weight and also asks parents about any hospitalizations, diagnosed health conditions, and severity and impact of any illnesses that their children have experienced. In addition, the Early Childhood Longitudinal Study of a Birth Cohort measured height, weight, and other growth measures, as well as obtaining parent report about health status and medical services. In addition, both the Early Head Start Evaluation and FACES obtained parent report of health status and medical care. Measurement work in this area should focus on identifying or adapting existing measures of the specific aspects of child health related to nutrition that can be informed by large-scale health studies of low-income children.

e. Activities and Length of Stage 1 for the FFP

The I&I planning phase for the FFP, intended to last nine months, ends with the completion of the enhancement definition through the specification of the FFP's various components. However, to fully complete the activities of Stage 1, FFP will require additional time to refine documentation, and to develop and conduct preliminary testing of measures. At the end of the initial nine-month planning phase, FFP will be implemented in Head Start programs throughout Pulaski County. Over the course of an additional six to nine months, FFP developers can use the experience of the sites in Pulaski County to learn about the resources required for successful implementation, the different methods that programs use to develop employer and community resource networks and to engage parents, and the challenges to implementation. This information will feed into the enhancement refinement process to assist FFP in developing the comprehensive documentation necessary for expanded replication. During the six- to nine-month period, measurement-development work should be under way as well.

It is feasible and recommended that, as part of Stage 1, FFP be implemented in additional sites outside of Pulaski County. This expansion will enable developers to test replication in sites that may have characteristics and resources that differ from those of Pulaski County. Implementation should occur after an initial period of implementation in the original sites, when documentation can be considered comprehensive, and when measures have been developed or selected. The quality of the implementation and fidelity measures should be tested through implementation studies in both the original sites and the second-phase sites in order to assess threshold levels, and to test multiple measurement approaches, where applicable. In addition, developers should conduct pre-post outcome studies of the intermediate outcomes in all the sites. It is unlikely that any descriptive child outcomes study of FFP would be informative in Stage 1. Because this enhancement would affect child outcomes through the parent, changes in child outcomes are unlikely to occur in the short term.

For FFP to progress to a Stage 2 evaluation, the components of FFP must have been developed fully and documented clearly, and the implementation studies must suggest that replication is feasible, particularly in diverse sites. Two issues will be of particular importance for determining the replicability and feasibility of FFP necessary for progression

to Stage 2. First, FFP developers will have to give substantial thought to the options for establishing the Work Program component in order to ensure that Head Start programs serving diverse populations and having varying levels of resources within Head Start and within the community can successfully achieve implementation. Second, the early implementation sites in Pulaski County and elsewhere will have to achieve the thresholds for participation so as to provide important lessons about methods of engaging parents and building participation to the levels necessary for evaluation. The studies of intermediate outcomes for FFP parents should reflect positive improvements of a magnitude that has the potential to create significant changes in child outcomes. As an enhancement in the very early stages of development, it would likely take two to three years for FFP to be ready for a Stage 2 evaluation.

2. English Language Learners Project, Community Development Institute (Denver, Colorado)

The goal of the English Language Learners Project (ELLP), a project of the Community Development Institute (CDI), is to integrate ELL assessment and instruction into Head Start programs that serve large numbers of children who are learning English as a second language in order to improve the children's English language acquisition and literacy skills. The ELLP will provide initial intensive training on ELL assessment and instruction for Head Start teachers, home visitors, and administrators, supported by ongoing training and technical assistance for three years. The Community Development Institute, with I&I support, is refining the ELLP through planning and implementation activities with 22 Head Start grantees (2 per region). This enhancement is well-defined, and its *Guide to Working with English Language Learners* has established a strong framework for documentation. Additional documentation, particularly on training and technical assistance, still must be developed, and measurement work will be necessary. Pilot implementation in the 22 sites will provide important information about the variations and challenges that can arise across different sites. This information will help to refine the enhancement and the supporting documentation, as well as assist in developing and refining measures of implementation, fidelity, and outcomes.

a. Level 1: Defining the Enhancement

Motivation and goals. More than one-quarter of Head Start children speak a language other than English at home (see Table II.7; and U.S. Department of Health and Human Services 2002). Spanish is the most common non-English language, spoken by 23 percent of Head Start children. Relative to the national average, Migrant and Seasonal Head Start programs and programs in the West and Northeast have substantially higher percentages of Spanish-speaking children.[5] Based on demographic projections, the proportion of ELL children in Head Start is projected to increase during the next decade and beyond (Espinosa 2004; Iglesias 2004). According to CDI, recent studies indicate that more ELL children drop

[5] Data are based on analyses of 2002-2003 Head Start Program Information Reports (PIR) data.

out of high school than graduate. Studies also suggest that most U.S. schools do not adequately address the needs of ELL children, and that practical guidance that can be applied to the classroom is lacking. CDI initiated the ELLP with the belief that ELL instruction in Head Start classrooms will improve English language acquisition and will enhance the emerging literacy skills of English-language learners in the short run, thereby setting a course for potential improvement in educational outcomes in the long run.

Therefore, the short-term goals of the ELLP are to improve English language acquisition and literacy skills among Head Start children who are English language learners, and to increase the social and emotional competence of these children. To achieve these goals, Head Start programs in which the ELLP enhancement has been fully implemented should (1) conduct assessments using methods that accurately gauge the skills and abilities of ELL children, (2) provide effective ELL instruction in Head Start classrooms, and (3) use methods to engage families of ELL children in their children's education. To fully define the enhancement, ELLP developers will have to detail those three key components of the project.

Enhancement components. To improve the accuracy of assessments for ELL children, the ELLP, presumably, will provide a toolkit of resources that can be used to tailor Head Start's individual child assessments to the needs of ELL children. This toolkit is likely to include nonverbal methods of gauging a child's abilities. Nonverbal methods generally are viable options only for measuring abilities in domains other than language development and literacy (for example, tasks that can be modeled for the child, such as identifying the "odd" item in a set of items sorting objects by size, and showing numbers on fingers). However, some basic literacy cues can be gleaned from nonverbal practices, such as whether a child is drawn to looking at books on his or her own, and whether the child holds the book upright and turns the pages from left to right. Other methods that typically are used in an assessment portfolio to gauge early language development and literacy include parent questionnaires/interviews, alphabet knowledge, and word associations with pictures presented on flashcards. The use of these methods will help teachers to gauge the level of English language acquisition; however, unless they are conducted in the child's primary language, they will provide little information on the level of early literacy. In theory, each Head Start program has a staff member who is able to communicate with children and parents in their primary language. Methods of involving these individuals in the process of assessment may be another aspect of the ELLP's approach to individual child assessment.

ELLP instructional activities for Head Start teachers, the second component to the ELLP, are not part of a formal curriculum. They consist of a set of lesson plans and activities for classroom use that are designed to integrate into such approaches as the Creative Curriculum or High Scope. The ELLP methods and principles will have to be broad enough and flexible enough to incorporate into any curriculum in use in Head Start classrooms. ELLP instructional activities may be shaped around the four stages of second language acquisition: (1) home language use (children continue to speak in their home language even while they are surrounded by people who are speaking in English); (2) the nonverbal period (children realize that their primary language is not understood, become quiet, and become observant of the uses of English); (3) the use of telegraphic and formulaic

speech (children use individual English words or use these words in short, often incomplete, and ungrammatical sequences); and (4) the use of productive language (children begin to speak English relatively well, using phrases and sentences) (Gonzales 2005). Each stage may have specific teaching tips and goals, but ELLP methods are likely to integrate an array of practices into each lesson and/or activity for two reasons. First, children in Head Start classrooms are likely to be at different stages of English language acquisition. Second, children do not necessarily proceed through the stages in a linear fashion but exhibit traits of each stage at various points in time. For these reasons, teachers may be able to "customize" their practices when working with children independently but will need more inclusive ELL approaches for group lessons.

The third component—employing methods of engaging families of ELL children in their child's education—may be the one that is least sharply defined by ELLP developers at this time. Because the ELLP's goal is to better integrate families into their children's educational lives and to promote literacy activities in the home, ELLP developers will need to develop an array of activities designed to reach out to and communicate with families, including home visits by staff members who speak the families' primary languages, translation of program materials and child progress/assessment information into the primary languages spoken in the home, parent/teacher meetings in which teachers and a translator are present, and informational sessions and parent activities specifically for non-English speaking parents to increase their comfort level.

Theory of change. Through its three key components, the ELLP expects to increase English language acquisition and literacy by having impacts on both the Head Start classroom and the home environment. The enhancement's *targets of change* are general staff's level of knowledge about ELL principles (such as the four stages of second-language acquisition and the use of inclusive teaching practices), teachers' skills in the areas of ELL assessment and instruction, and the level of skills of all Head Start staff (home visitors, teachers, and administrators) in interacting with the families of ELL children. *Intermediate outcomes* will measure the extent of change observable in the classroom and in parents' behaviors that could result from increasing Head Start staffs' ELL skills. Intermediate classroom-focused outcomes will examine changes in the quality of the language and literacy environment. They also will include a more global measure of classroom quality across a broader range of topics (to assess any "spillover" effects). Intermediate parent-focused outcomes might measure the parents' level of involvement with the Head Start program and the frequency and duration of literacy activities in the home. *Child outcomes* of interest—the ELLP enhancement's final outcomes—will focus on the domains of language development and literacy. In addition, through an improved ability to communicate with their teachers and their peers, ELL children could become more interested and engaged in learning as well as develop better social relationships.

b. Level 2: Documenting Implementation

Implementation products for the ELLP will fall into one of two categories: (1) guidance on specific T/TA modules for use by trainers and consultants who will prepare Head Start staff for the ELLP enhancement, and who will then support these staff

throughout ELLP implementation; and (2) implementation guidance for use by the Head Start program staff that would describe in detail the assessment methods for ELL children, the ELL instructional methods for the classroom, and the methods of engaging the families of ELL children.

Program needs assessment. Because of the variation in Head Start programs, T/TA necessary to achieve full implementation is likely to vary across programs as well. To determine the content and intensity of T/TA needed to launch the ELLP, ELLP consultants will have to work with Head Start program staff to conduct a program needs assessment. Guidance on conducting this type of assessment should therefore be a primary element of any T/TA guidance. The goal of a program needs assessment is to identify prominent gaps in a program's ELL knowledge, assessment, and instruction. T/TA guidance should detail the specific areas of interest for a program needs assessment, the sources of data and information, and the steps of the assessment. Three of the more obvious factors for consideration in such an assessment are:

(1) What percentage of the Head Start student population is non-English speaking? What primary language do these children speak at home?

(2) What does the current child assessment system look like? Does it include an assessment of a child's status and progress in his/her home language? If it is not coordinated and comprehensive, it may be necessary to provide T/TA to help programs enhance their current assessment system for all children in order to achieve the goal that ELL children are assessed accurately. In the absence of a solid framework for child assessment, programs probably will not be able to achieve the thresholds for quality of ELLP implementation and fidelity that will be necessary to progress to the next stage of evaluation.

(3) What is the Head Start staff's level of knowledge of ELL assessment and instruction? To design initial T/TA plans that will raise the staff's ELL knowledge to a standard level, T/TA providers will have to be aware of the range of staff ELL skills and abilities in ELL instruction. T/TA providers may be able to integrate the assistance of staff with strong ELL competency into training efforts.

T/TA guide. The ELLP T/TA guidance should specify the training options and set parameters for the critical and required training elements. For example, training on the four stages of language acquisition may be a standardized training module that should have little to no modification because the principles apply across all ELL students. However, training on specific instructional approaches may vary in content to adapt to cultural differences associated with children's primary language. The T/TA guide could be organized around different topical modules (for example, the stages of language acquisition, assessment, classroom instruction and activities, and interactions with families), with each module detailing the core elements, such as trainers' qualifications, core content, timing, and flow. Options for each module could offer choices about the intensity and method of training (which would depend on program resources, location, or program staff's level of existing

ELL knowledge), and that provide adaptations to content based on the predominant language of ELL children in the Head Start program.

Implementation manual. The ELLP's current implementation manual for participating Head Start program staff (the *Guide to Working with English Language Learners*) gives this enhancement a strong starting framework for the second type of implementation product—program implementation guidance. The implementation manual should essentially be a "user's guide" to the ELLP enhancement that provides detailed guidance on each of the key components—assessment, ELL instruction, and family engagement. Although all of the information in the implementation manual should complement and reinforce what staff learn through training, the manual differs from T/TA guidance in that it is targeted to direct "users," rather than to trainers.

The implementation manual for ELLP will need to include detailed guidance for each of the three components. All three components to the ELLP are likely to look like a menu of options from which different Head Start programs can choose the approaches that would work best given their current child assessment practices, primary language of non-English-speaking children and families, size of the non-English-speaking student population, and staffs' language abilities. The basic elements to the implementation manual should include:

- **General implementation guidance.** The implementation manual might begin with a section on general implementation information that would outline such items as the center staff who must participate in ELLP training for the center to reap the full benefits of the enhancement, recommended distribution of ELL children across the classrooms in participating Head Start centers, and whether and how the ELLP approach can be adopted for use by programs with smaller numbers of ELL children.
- **ELL assessment.** The assessment section should include examples of assessment tools and methods, accompanied by detailed descriptions of how to use them, when, and by whom.
- **ELL instruction.** The instructional practices section would include specific lesson plans and/or classroom activities, with guidance on the flow and timing of each. This section must provide suggestions on how to integrate ELL methods with the classroom's existing methods. Head Start teachers have multiple goals, so the ELLP should provide specific examples on how ELL methods can be integrated into existing activities and lessons, rather than displacing them, or how ELL can be targeted to individuals or small groups of children.
- **Engaging families.** The section on engaging families should present the critical topics for communication with parents (for example, child assessment information and activities to promote family literacy activities) and methods for encouraging parental involvement in the Head Start program.

c. Level 3: Developing Measures

Development and selection of evaluation measures are critical to the ELLP's development stage. As a well-defined and focused enhancement, there are some clear directions that measurement of the ELLP could take (Table III.3).

Quality of implementation. Measures of the quality of implementation assess the extent to which initial and ongoing ELLP T/TA is delivered as prescribed by the program developers at CDI. In addition, an objective of the ELLP enhancement is to develop agreements with local community colleges to offer Head Start staff credit for their ELLP T/TA. Credit may be indicative of the quality of implementation by reflecting comprehensive T/TA coursework and a commitment on the part of the program.

Other measures of implementation should examine the extent to which the components of the ELLP have a solid framework for success. For example, implementation measures should assess the extent to which a comprehensive, coordinated child assessment system is in place by examining the existence of a common set of procedural guidelines for collecting assessment data on children. These guidelines can be an indication of the level of planning and consideration that has been given to a child assessment system. Similarly, a measure of the depth and breadth of methods to engage the parents of ELL children with the Head Start program can reflect the level of implementation of the family component of ELLP. Finally, plans for ongoing T/TA to support ELLP should be examined for an indication of a lasting commitment to the continued professional development of staff on ELL assessment and instruction and to the project as a whole. Implementation information will largely be gathered through direct observation or through interviews during site visits that should be conducted during the period of initial intensive training and during full-scale implementation.

Fidelity to enhancement. Fidelity measures will capture how well participating Head Start programs meet the ELLP's functional objectives of conducting assessments, delivering ELL instruction to children, and engaging the children's families. This measurement work will be intensive and will require careful consideration because measures will have to be developed for specificity to the ELLP model in measuring how well the targets of change (teachers' skills, teacher/child interactions, and staff/parent interactions) react and adhere to the ELLP enhancement. Before fidelity measures can be developed, ELLP developers will have to define broader concepts relating to fidelity. For example, they will have to define the critical elements of ELL competency and accuracy in ELL assessments in order to develop measures to capture these concepts. Fidelity measures that focus on the classroom will have to measure the specific actions, practices, methods, and languages that teachers use in relation to the ELLP guidance on instructional practices.

Intermediate outcomes. Intermediate outcome measures could be drawn from existing tools. The Early Language and Literacy Classroom Observation (ELLCO) Toolkit measures the quality of the language and literacy environment in early childhood classrooms, has good psychometric properties, and has been widely used in large-scale studies. The Early Childhood Environment Rating Scale-Extension (ECERS-E), another widely used

Table III.3 Potential Measurement Options for the ELL Project

Area of Measurement	Concept/Domain	Potential Measurement Method/Existing Tool
Quality of Implementation	Duration and intensity of ELL teacher training, staff training, and TA meet intended parameters.	Observation and interviews during site visits
	Training covers intended content and activities.	Observation and interviews during site visits
	Trainer(s) meet standards for qualifications.	Resume review; observation during site visits
	Availability of college credit for ELL training	Review of course curriculum and agreement with local community college
	Level of participation in T/TA by "required" Head Start staff	Observation and interviews during site visits
	Common set of procedural guidelines for collecting individual child assessment information is in place.	Paper reviews and interviews during site visits
	Multiple methods of engaging parents of ELL children are in place.	Observation and interviews with staff during site visits
	Plans for continued T/TA to support the ELL enhancement are in place.	Paper reviews and interviews during site visits
Fidelity to Enhancement	ELL competency of trained Head Start staff	Post-test of staff trainees tailored to specific ELL concepts (must have knowledge of a specified percentage of training content)
	Breadth, frequency, accuracy, and consistency of assessments completed among ELL students meet ELL parameters.	Observation and interviews during site visits
	Use of assessment information to inform instructional activities in the classroom	Observation and interviews during site visits
	Teacher practice of ELL instructional skills (teacher skills, classroom activities, teacher/child interactions)	Classroom observation
	Interactions between Head Start program staff (home visitors, teachers, administrators) and parents of ELL children meet quality parameters in terms of content, frequency, duration.	Observation during site visits
Intermediate Outcomes	Language and literacy environment in the classroom	ELLCO
	Global classroom quality	ECERS-E
	Level of parental involvement in Head Start program	Parent Involvement and Satisfaction with Head Start (FACES Spring 2003)
	Increased frequency and duration of parent-child reading time	FACES Parent Interview "Activities with Your Child"

Table III.3 *(continued)*

Area of Measurement	Concept/Domain	Potential Measurement Method/Existing Tool
Child Outcomes	Listening and understanding	PPVT-III TVIP - Pre-LAS 2000 Oral Language Component—Simon Says (English and Spanish versions)
	Speaking and communicating	Pre-LAS 2000 Oral Language Component—Simon Says or Art Show (English and Spanish versions)
	Phonological awareness	Pre-CTOPPP—Elision Task (English and Spanish versions)
	Book knowledge and appreciation	- Parent Report of Child's Emerging Literacy (Parent Emergent Literacy Scale) (English and Spanish versions) - Story and Print Concepts (FACES) (English and Spanish versions)
	Print awareness and concepts	- Bateria Woodcock-Munoz Pruebas de Aprovechamiento, Identificacion de Letra y Palabras - Woodcock-Johnson III Tests of Achievement—Letter-Word Identification Test - Parent Report of Child's Emerging Literacy (Parent Emergent Literacy Scale) (English and Spanish versions) - Pre-CTOPPP—Print Awareness Subtest (English and Spanish versions) - Story and Print Concepts (FACES) (English and Spanish versions)
	Early writing	- Bateria Woodcock-Munoz Pruebas de Aprovechamiento—Revisada, Dictado - Woodcock-Johnson Revised Tests of Achievement—Dictation Test - Parent Report of Child's Emerging Literacy (Parent Emergent Literacy Scale) (English and Spanish versions)

Table III.3 *(continued)*

Area of Measurement	Concept/Domain	Potential Measurement Method/Existing Tool
	Alphabet knowledge	- Bateria Woodcock-Munoz Pruebas de Aprovechamiento, Identificacion de Letra y Palabras - Woodcock-Johnson III Tests of Achievement—Letter-Word Identification Test - Letter Naming Task (NRS) English & Spanish versions - Parent Report of Child's Emerging Literacy (Parent Emergent Literacy Scale) English and Spanish versions - Pre-CTOPPP—Print Awareness Subtest (English and Spanish versions)
	Initiative and curiosity	- COR Initiative Scale - PBLS - Social Skills Scale of the Social Skills Rating System—Teacher
	Social relationships	- COR Social Relations Scale - Social Skills Scale of the Social Skills Rating System—Teacher - Social Competence Subscale of the Social Competence and Behavior Evaluation

COR = Child Observation Record; ECERS-S = Social-Emotional Learning Checklist; ELL = English language learners; ELLCO = Early Language & Literacy Classroom Observation Toolkit; FACES = Family and Child Experiences Survey; NRS = Head Start National Reporting System; PBLS = Preschool Learning Behavior Scale; PPVT-III Peabody Picture Vocabulary Test, Third Edition—Revised; Pre-CTOPPP = Preschool Comprehensive Test of Phonological and Print Processing; TA = technical assistance; T/TA = training and technical assistance.

measure of global classroom quality, examines the four curricular areas of literacy, mathematics, science and environment, and diversity. The ECERS-E will capture indirect effects of the ELLP enhancement that may occur in the classroom environment beyond the direct effects on language and literacy. Other intermediate outcome measures involve changes in parental behavior. Existing tools use parent interviews to gauge the parents' level of involvement in the Head Start program and the extent of literacy activities in the home. It may be too much to expect changes in the literacy activities that parents engage in with their children. However, these changes depend on the nature and intensity of the outreach activities to parents, and the extent to which the ELLP emphasizes family literacy. Many parents of ELL children may be illiterate not just in English, but in their primary language as well.

Child outcomes: language development and literacy. Measuring child outcomes for an enhancement focused on ELL children is particularly challenging. Gauging English language acquisition over the course of the intervention may be somewhat straightforward, but measuring the broader concepts of language development and literacy skills that can occur in the primary, secondary, or both languages is more difficult. Recent reviews of the available child outcome measures available in languages other than English reveal that there are fewer than 20 norm-referenced standardized tests in Spanish, and almost none in other languages (Kochanoff 2004; Kochanoff et al. 2003). In addition, some of the existing measures have problematic flaws, such as being normed using data from monolingual speakers in other countries (for example, Mexico), rather than having norms that reflect the experience of English language learners in the United States. Another criticism of the existing tests is that the Spanish in the tests may reflect only one dialect, which makes the test more difficult, and possibly invalid, for speakers of other dialects (for example, Spanish spoken by children in Puerto Rico can be quite different from Spanish spoken by children who are second-generation Mexican immigrants to the United States).

Because few choices are available for standardized assessments in Spanish, many researchers create Spanish versions of assessments and interview questions without completely understanding the importance of equating the level of difficulty across the test versions. Simply translating an English assessment into another language will not solve the problem of the absence of valid, reliable tests in other languages.

In Table III.3 we list the domain elements in the language development and literacy domains, along with existing measurement tools that are available in both English and Spanish versions. However, even if a test has both English and Spanish versions, the two versions may yield data that are not comparable; the two tests may differ completely from one another or the norming samples may not have been comparable. The PPVT-III and its Spanish version, the Test de Vocabulario en Imagenes Peabody, are one of the best examples of this problem. The Spanish version of this widely used test is much older than the English version and yields data that cannot be combined or compared validly with the English version.

Researchers have attempted to avoid these potential problems by testing children who demonstrate competency in English and Spanish in both languages. This approach builds

on empirical findings about the course of bilingual children's language development (August et al. 2002; Iglesias 2004; Tabors et al. 2004). In summary, relative to their monolingual peers, children learning two languages perform lower on tests of their abilities in either language. When researchers combine bilingual children's pass rates on comparable questions (for example, expressive or receptive vocabulary) across both languages, children perform at the same level as their peers. That is, they know as many words across two languages as their peers know in one language. This finding has important implications for studies of Head Start enhancements that focus on supporting language development for English-language learners. It is not clear that results from two different tests that are not calibrated at the same level of difficulty can be combined validly. Furthermore, it is not clear how such data would be used in impact analyses.

These new findings about how vocabulary development progresses among children learning two languages also have practical implications for evaluation design. Assessments cannot be too long if some children in a study must complete twice as many assessments as others (one in each language). Burden on the children also must be taken into account as measures are selected. Approaches such as matrix sampling, where not all children receive every test item, may be particularly useful for reducing burden on children who must take two versions of the assessments.

Child outcomes: approaches to learning and social relationships. Measuring child outcomes in other domain elements, such as initiative and curiosity and social relationships, typically is accomplished through parent or teacher reports, rather than through observational assessments. In Table III.3, we list only existing measures that rely on teacher reports, as the use of teacher reports reduces problems associated with questioning a group of parents who speak a variety of languages. "Initiative and curiosity" refers to a child's eagerness to learn. It includes an increased ability to make independent choices, and to choose to participate in a variety of tasks and activities. The COR's four-item Initiative Scale taps a child's ability to express choices, solve problems, engage in complex play, and cooperate with program routines, but it does not contain items relating to a child's curiosity. The Preschool Learning Behavior Scale, used in FACES 2003, assesses the child's approaches to learning, including his or her motivation to learn and behaviors that enhance the child's learning. "Social relationships" refers to a child's growing interest in and ability to develop friendships and positive relationships with adults and peers. As shown in Table III.3, several existing scales measuring social relationships through teacher reports have been used in large-scale studies of low-income children.

d. Activities and Length of Stage 1 for the ELLP

The ELLP expects to spend about three months refining definitions of the enhancement's key components. Early implementation experimentation will then begin in the 22 test sites. If the ELLP enhancement is intended for broad replication, it will be particularly useful to include in the 22 sites Head Start programs serving children who speak a range of languages, as well as programs with varying-sized ELL enrollments. The initial intensive T/TA for ELLP should begin well in advance of the trial implementation program year (for example, in January, in preparation for full implementation in the academic year

starting the following September). CDI will use the experience of the 22 sites during the initial T/TA and the first year of implementation to further refine the definition and the documentation for the ELLP. Measurement development work also will take place during this time. With early implementation in 22 sites, CDI enhancement developers and selected evaluators may be able to experiment with different approaches and methods of measurement. This type of experimentation could be particularly useful given the range of issues, as discussed earlier, that are unique to the measurement of language development and literacy among preschool children for whom English is a second language.

An evaluative on-site implementation study should be conducted during the initial T/TA period to gather information about the quality of implementation. ELLP developers should conduct a second on-site study during the program year, to gauge fidelity. An outcomes study should measure intermediate classroom outcomes and child outcomes at the beginning and near the end of the program year. Some improvements in language development and literacy would be expected during the course of a normal program year in any Head Start program. Thus, at this early stage, changes in the intermediate outcomes may be more meaningful than any measured changes in child outcomes. The intermediate outcomes have a more direct relation to the activities of the enhancement and therefore may be more suggestive of the potential for change through the ELLP.

The ELLP might be ready for a Stage 2 study at the end of about one and one-half years, a period allowing for initial T/TA and one program year of implementation in the original test sites. However, because the plan for the full enhancement is to provide three years of ongoing T/TA, evaluation plans will have to consider this full implementation period. An important consideration for rigorous study of this enhancement (or of any other ELL-focused enhancement) is the breadth of its replicability and feasibility. First, a decision should be made about the applicability of the ELLP enhancement within the broader Head Start community, including programs with smaller numbers of ELL children. It is possible that the ELLP enhancement could be adapted to meet the needs of these programs, but that the programs will not be able to achieve the thresholds of fidelity if only minimal attention to ELL practices (or more focused attention to individual children) is needed. The decision could be made to test the ELLP enhancement only in Head Start programs that have a predetermined percentage of ELL children. Second, even with testing only in programs with relatively large populations of ELL children, evaluators will have to assess the feasibility of measurement across a range of languages if children will have to be tested in both their primary language and in English.

3. Violence, Intervention and Prevention Program, Circles of Care, Melbourne, Florida

Circles of Care has developed the Violence, Intervention and Prevention (VIP) enhancement to integrate mental health problem prevention and intervention services into Head Start programs. Circles of Care is a behavioral health care organization in Brevard County, Florida, that provides mental health, alcohol, drug abuse, and related services to county residents through its hospital-based settings and state and county contracted programs. The VIP enhancement has four key components: (1) the Second Step preschool curriculum, which teaches social and emotional skills for violence prevention; (2) the

Building Strong Families program for parents, which focuses on intensive child and family social skill building; (3) individual, group, and family counseling services to children and families with specific mental health needs; and (4) case management services for fathers. The VIP program will be a centralized effort, managed and staffed by Circles of Care for implementation in five Head Start programs in Brevard County. The I&I grant to Circles of Care supports an initial planning period to refine the development of the VIP program, and to build the infrastructure for implementation and evaluation.

a. Level 1: Defining the Enhancement

Motivation and goals. The prevalence of social and emotional problems among preschool children is an issue of growing concern, as these problems can affect academic performance as early as first grade (Raver and Knitzer 2002) and persistent problems can have longer-term effects, including substance abuse, depression, violent behavior, and school failure. The Early Childhood Longitudinal Survey found that approximately 10 percent of children in an average kindergarten classroom easily become angry or engage in arguments or fights (West et al. 2000). Teachers in one study reported that about 40 percent of preschoolers in Head Start exhibited at least one disruptive or unsafe behavior each day (Kupersmidt et al. 2000). In response to increasing prevalence rates of problem behaviors in classrooms, clinical disorders, and children in families with risk factors, Circles of Care initiated this early intervention program to address the emotional and mental health needs of Head Start children and their parents.

When fully implemented, the VIP enhancement intends to increase the social and emotional competence of Head Start children by providing: (1) one or two 30-minute sessions for preschoolers each week, using the Second Step curriculum, which teaches prosocial behavior and improving teachers' skills in helping children to incorporate the practice of emotion management and problem solving throughout the Head Start day, (2) instruction in social skills-building and parenting skills to the parents of Head Start children through Circles of Care's Building Strong Families program, (3) on-site individual, group, and family counseling services, facilitated by Circle of Care's mental health specialists, to an estimated 20 percent of Head Start children and families across the five pilot implementation sites, and (4) case management services to fathers of Head Start children through a Circles of Care Outreach Specialist.

Enhancement components. During a development phase, VIP developers will have to specify the targeted population, as well as the content, dosage, and duration for each of the four key components. The first component, the Second Step curriculum, is a pre-packaged curriculum developed by the Committee for Children that teaches empathy, impulse-control, anger-management, and problem-solving skills. It is recommended for use throughout an entire school (or, in this case, throughout an entire Head Start center), rather than for use in a single classroom in a school or center. The preschool lessons last approximately 30 minutes, and children are motivated by poster-sized photo-lesson cards and puppets and soft toys (Impulsive Puppy, Slow-Down Snail, and Be-Calm Bunny) that relate to the cards. The cards depict children expressing emotions in real-life situations. Instructions for the teacher on the back of each card present an overview of the concepts to

be covered, list the lesson's objectives and required materials, and provide a story about the photo that is accompanied by discussion topics and specific questions for the children to answer. Each lesson also contains guidance about supplemental activities, such as role-plays, and offers teachers suggestions about how to model the skills taught in the lesson throughout the week. The Second Step curriculum also provides ideas for extension activities that can be incorporated into classroom activities throughout the week, with minimal preparation or materials.

VIP's second component, the Building Strong Families program for parents, is designed to build stronger relationships within the family, and to prevent mental health and substance abuse problems. The enhancement will have to define whether this program component will target all parents of Head Start children in the implementation sites, or whether families with specific risk factors (such as single-parent households, low-educated parents, or low proficiency in English) will be the primary targeted population. The method of delivery and dosage for this parental component will need definition such as whether Building Strong Families will provide parental instruction and guidance through small group workshops, one-on-one sessions, or a combination of the two methods, and the frequency and length of the sessions. Presumably, the content of Building Strong Families sessions will include such topics as stress management, handling of conflict, parenting skills, and relationship skills.

For the third component, on-site mental health counseling will be provided at each implementation site one day each week. Circles of Care expects that 20 percent of the Head Start families within the VIP program sites will need individual, group, or family counseling. The screening procedures to identify these families should be detailed, as should the specific counseling plan and approach (for example, the method, frequency, and duration of counseling).

Finally, the case management support for fathers to maximize fathers' participation in the VIP program specifically, and in the Head Start program more broadly, needs definition. Again, any criteria that Circles of Care will use to target fathers with specific characteristics or circumstances for services should be made explicit. In addition, VIP developers will have to describe the specific outreach activities to engage fathers, as well as the breadth and depth of the case management support (for example, employment referrals, social service referrals, and assistance with supportive services, including transportation and child care).

Theory of change. Circles of Care also will have to define explicitly the theory of change for the VIP program in order to guide early evaluation efforts, and to identify the intermediate and child outcomes of interest. As a comprehensive approach, the goal of the VIP enhancement is to increase children's social and emotional competence through both the classroom and home environment. In this section, we describe the targets of change and the intermediate outcomes for each of the four components of VIP separately. Although each component of the VIP enhancement may have its own targets of change and intermediate outcomes, all four components are intended to work together to positively affect child outcomes in the area of social and emotional development.

The Second Step curriculum is designed to improve teachers' skills and increase the number of classroom activities that focus on prosocial behavior. The intermediate outcomes

that measure progress in this component would focus on the classroom and, possibly center, environment. For example, the overall quality of the classroom environment could be expected to improve, specific lessons or tasks may take less time if children remain "on task," and fewer classroom disruptions due to arguments, fights, or other problem behaviors may occur. Other changes may be visible in the center more broadly and could be captured by an intermediate outcome that gauges the climate on the playground, in the lunchroom (if applicable to full-day programs), or during other group activities that integrate children across classrooms (such as an indoor play area or group music program).

For the sake of this example, we assume that the Building Strong Families component will focus on some of the same issues that we presented in our discussion of the Brief Parenting Intervention component of the FFP example (discussed in Section D.1 of this chapter). Specifically, the topics of this intervention would focus on supporting the mother-father relationship and positive parenting skills. We also assume that the Building Strong Families component directly addresses mental health and substance abuse issues by discussing the risk factors and indications of a problem, and by offering routes to screening, counseling, and treatment, if necessary. The targets of change are individual coping skills, parent-to-parent communication skills, and parent-child positive interactions. Intermediate outcomes to measure these changes might include a reduction in household conflict, an increase in father presence and involvement with the child, a reduction in parenting stress, and an increase in parenting support and warmth. Other intermediate outcomes of the Building Strong Families component might measure the percentage of parents screened for mental health and substance abuse problems, as well as a follow-up measure for those who are diagnosed with a problem and subsequently have entered counseling or treatment.

The counseling component of VIP will have different routes to the child depending on the method of delivery. In some cases, a child with significant behavioral problems or a diagnosed problem will receive direct individual counseling from a Circles of Care mental health specialist. In other cases, the entire family may receive counseling; in still others, parents may receive group counseling. The end goal is improved mental health of the individual receiving the counseling. When the parents or the whole family unit are the recipients of counseling, the intermediate outcomes affecting the child might include reductions in household conflict and parenting stress. In addition, an intermediate measure should capture the psychological well-being of the parent.

We assume that the outreach and case management support provided specifically to fathers would have overlapping intermediate outcomes with the Building Strong Families component. Specifically, the intermediate outcomes from the case management support to fathers would include increased father engagement with the child and increased parenting support and warmth specifically between the father and the child. Circles of Care also expects that increased father engagement will lead to increased financial support of the child.

b. Level 2: Documenting Implementation

As the curriculum developer, the Committee for Children has developed complete documentation for the Second Step curriculum. The Committee for Children provides the

training for teachers and all classroom materials, lesson plans, and extension activities necessary to implement Second Step. Circles of Care will have to produce the detailed documentation necessary for replication of the VIP model as a whole by clearly documenting the three remaining components during the development stage.

Community resource assessment. Implementation guidance should first help Head Start programs to assess the mental health and social service agencies in their communities in order to identify partners for the VIP program. Circles of Care is a unique organization with the ability to provide all the distinct pieces that comprise the integrated VIP program. Aside from the Second Step curriculum, the other program components may or may not be housed within the Head Start program itself. Counseling services (either on- or off-site) must be delivered by mental health specialists wherever the VIP program is implemented. Head Start programs are required by the Program Performance Standards to have relationships with mental health professionals for screening and referrals, and may want to build on these existing partnerships for implementation of the VIP program. Individual Head Start programs can be given the choice of administering the Building Strong Families parenting and family supports and the case management services for fathers themselves or seeking community partners for these services. Cost and resource constraints may prevent a program from adopting the whole model for in-house administration.

Implementation guidance for program components. Circles of Care will have to consider the finer points of implementation that a Head Start program or partnering organization will have to take into account when implementing the components of the VIP program. In fact, each component could have its own stand-alone documentation.

- *Counseling services.* Circles of Care should define the framework for counseling services in terms of the qualifications of counselors, the types of counseling that should be made available, and the frequency and duration of services.
- *Case management services for fathers.* Guidance should specify the qualifications of case managers, the ideal ratio of case managers to fathers, and the range of services and supports for fathers. Although services may vary from community to community, documentation should include some protocols for the quality, frequency, and duration of interactions between case managers and fathers. It will also be critical to describe outreach techniques for engaging fathers and sustaining contact with them over time.
- *Building Strong Families.* Documentation for the Building Strong Families component may be the most comprehensive to produce. This component will have to include a training section that details the qualifications of trainers, the content of training, and the intensity and duration of training to prepare staff members who will administer this parent instruction component of VIP. Documentation also will have to include instructional guides for staff who will facilitate Building Strong Families workshops or sessions with parents that detail the topics, learning objectives, and activities to engage parents in discussion.

Communication and referrals between components. A key element to replication of the VIP program by other Head Start programs will be achieving the appropriate level of service integration and communication among the program components. If separate entities administer separate pieces of VIP, it will be critical to establish proper protocols for communication to create an integrated system of referrals and supports. For example, a trainer's experiences with the parents of a Head Start child during the Building Strong Families component may lead the trainer to recommend individual or family counseling with the mental health specialist. Implementation documentation should therefore specify methods to ensure that the communication and referrals among the various components of VIP are created and sustained.

c. Level 3: Developing Measures

Given the VIP program's integrated approach to improving the social and emotional competency of Head Start children and families, VIP developers will have to perform a substantial amount of measurement work during the development stage. The developers will have to develop or identify measures of the quality of implementation and fidelity to the enhancement for each program component as well as to capture the level of integration among services, where appropriate. This measurement work will benefit to some degree from measures that already have been developed for the Second Step program; as with any enhancement, however, measures in these areas typically need development for specificity to the enhancement.

Quality of implementation. Quality of implementation measures should capture the extent to which programs are able to garner resources, build partnerships, establish administrative procedures, identify qualified staff, and deliver the T/TA necessary to successfully launch each of the VIP program components (Table III.4). The Committee for Children has developed a Second Step Implementation Checklist that measures whether the planning steps and resources are in place to launch and sustain Second Step, the percentage of teaching and non-teaching staff trained, whether the appropriate level of resources and instructional materials have been obtained for each classroom, and the level of center-wide support for the curriculum. In addition, trainers complete a Second Step Training Evaluation immediately after training. This form, which takes about five minutes to complete, includes nine questions about the trainer's skill level, the training content, and the value of the session, all rated on a scale of poor, fair, good, and excellent, as well as five open-ended questions about ways in which the training could be improved. This type of participant feedback can be helpful, but VIP developers also may have to develop a tool for use by an objective, nonparticipant evaluator.

VIP developers and/or evaluators will also need to develop tools for measuring the quality of implementation for the additional three components of the VIP program. Among these should be an observational tool to measure the quality of staff training for the Building

Table III.4 Potential Measurement Options for the VIP Project

Area of Measurement	Concept/Domain	Potential Measurement Method/Existing Tool
Quality of Implementation	Second Step staff training sessions meet quality parameters in terms of qualifications of trainer/facilitator, frequency, duration.	- Observation and interviews during site visits - Second Step Training Evaluation
	School-wide implementation of Second Step meets parameters advised by developers.	Second Step Implementation Checklist
	Level of participation in Second Step T/TA by "required" Head Start staff	Second Step Implementation Checklist
	Partnership formed with appropriate entity to provide mental health counseling to individuals and/or families.	Interviews during site visits
	Mental health counselors meet qualifications.	Resume review; interviews with staff during site visits
	Criteria and methods for engaging individuals and/or families in mental health counseling are in place.	Interviews during site visits
	Resources to support Building Strong Families and/or connections built with necessary community partners are in place to effectively implement and sustain this component.	Interviews during site visits
	Building Strong Families staff training sessions meet quality parameters in terms of qualifications of trainer/facilitator, frequency, duration.	Observation and interviews during site visits
	Case managers/outreach specialist to work with fathers meet qualifications for position.	Resume review; interviews with staff during site visits
	Ratio of case managers to Head Start fathers meets parameters.	Observation and interviews with staff during site visits
	Multiple methods of engaging targeted parent populations in the various components of VIP are in place.	Observation and interviews with staff during site visits
	Communication/referral protocols among program components are in place.	Observation and interviews with staff during site visits
Fidelity to Enhancement	Competency of teachers and other trained Head Start staff on the concepts of the Second Step curriculum	Post-test of staff trainees tailored to specific Second Step concepts (must have knowledge of a specified percentage of training content)
	Teacher practice of Second Step instructional skills (delivery of specific lessons, teacher skills, classroom activities, teacher/child interactions outside of lessons).	- Classroom observation - Lesson-Completion Record - SELC

Table III.4 *(continued)*

Area of Measurement	Concept/Domain	Potential Measurement Method/Existing Tool
	Individuals/families are effectively engaged in mental health counseling, when appropriate; frequency and consistency of participation meets expectations.	Interviews during site visits
	Competency of trained staff on the concepts of the Building Strong Families parental instruction content	Post-test of staff trainees tailored to specific Building Strong Families concepts (must have knowledge of a specified percentage of training content)
	Building Strong Families sessions for parents meet quality parameters in terms of content, frequency, duration	Observation during site visits
	Interactions between case managers and Head Start fathers meet quality parameters in terms of content, frequency, duration	Observation during site visits
	Frequency of referrals/communication among components of VIP	Observation and interviews with staff during site visits
	Overall level of Head Start parent participation in Building Strong Families component	Administrative data/observation
	Level of participation of fathers in case management services and in Building Strong Families	Administrative data/observation
Intermediate Outcomes	Global measure of the quality of the classroom environment	- ECERS-R - Preschool program quality assessment
	Time on specific classroom activities or child-directed tasks	Observation during site visits
	Disruptions in the classroom	Observation during site visits
	Playground (or other integrated setting) climate	Playground and Lunchroom Climate Questionnaire
	Parenting stress	PSI
	Household conflict	FES Conflict Items
	Father presence and involvement with child	Father Activities with Child—Early Head Start Father Study Measures
	Increased parenting support and warmth; decreased detachment and intrusiveness	- Parent-Child Interaction Task (NICHD) - Parent-Child Relationship Scale (NICHD) - Father-Child Interaction Task for the Three-Bag Task
	Financial support of child provided by the father	Interview/questionnaire completed by the mother and the father
	Percentage of parents in Building Strong Families screened for mental health and/or substance abuse problems	Program data

Table III.4 *(continued)*

Area of Measurement	Concept/Domain	Potential Measurement Method/Existing Tool
	Percentage of parents in Building Strong Families with positive screens who are diagnosed with a problem; percentage who seek counseling or treatment	Program data
	Parental mental health/psychological well-being	- CES-D full version or abbreviated - Pearlin Mastery Scale, Locus of Control
Child Outcomes	Children's knowledge about empathy, problem solving, management of strong emotions, and ways to respond to problematic situations with peers	- Second Step Knowledge Assessment for Non- or Beginning Readers - Social Problem-Solving Test Revised—Child Measure
	Empathy	Empathy Subscale of the Social Skills Rating System—Teacher Report
	Self-control, regulation	Self Control Subscale of the Social Skills Rating System—Parent or Teacher Report
	Problem solving	Social Problem-Solving Test—Revised, Teacher Report
	Anger management/behavioral problems	- Behavior Problems Scale (or Classroom Conduct Problems) (NICHD)—Teacher Report - Child Behavior Problem Index (NICHD)—Parent Report - Child Behavior Checklist (NICHD)—Parent or Teacher Report
	Prosocial behaviors	- Howes Peer Play Observation Scale (FACES) - Parent-Child Interaction Task (NICHD) - Social Competence Subscale of the Social Competence and Behavior Evaluation—Teacher Report - Social Skills Scale of the Social Skills Rating System—Teacher Report

COR = Child Observation Record; ECERS-S = Social-Emotional Learning Checklist; ELL = English language learners; ELLCO = Early Language & Literacy Classroom Observation Toolkit; FACES = Family and Child Experiences Survey; NRS = Head Start National Reporting System; PBLS = Preschool Learning Behavior Scale; PPVT-III Peabody Picture Vocabulary Test, Third Edition—Revised; Pre-CTOPPP = Preschool Comprehensive Test of Phonological and Print Processing; TA = technical assistance; T/TA = training and technical assistance.

Strong Families component. Observations and interviews during implementation site visits should be performed to gather information about the infrastructure in place to support the Building Strong Families component, the case management services to fathers, and the mental health counseling services, whether it is housed in the Head Start agency or with partnering agencies. Finally, developers should measure other components that are critical to the success of implementation, including whether staff meet the necessary qualifications, whether the ratio of staff to parents/fathers is appropriate, the extent to which multiple methods of engaging parents in the VIP components have been developed, and the level of quality and comprehensiveness of communication and referral protocols to support integration of services.

Fidelity to enhancement. Fidelity measures will capture the extent to which the VIP components look and function as intended by the developers at Circles of Care. Again, each component should be considered separately in measuring fidelity, and at least one measure should assess the frequency and quality of referrals and communication among components (Table III.4).

Second Step developers at the Committee for Children have two tools that may help to measure fidelity for the Second Step curriculum. Both are based on either teachers' or staffs' reports and may therefore have to be combined with an independent classroom observational tool specific to the Second Step curriculum. The first tool, the Lesson-Completion Record, is used by teachers to track when specific lessons were conducted, which students participated in the lessons' role-plays, and how many students participated. The Social-Emotional Learning Checklist (SELC) is a method of assessing teachers' (or other staffs') support of students' skill use outside of Second Step lesson instruction. Using a scale of "never, once, 2-3 times, 4+ times," teachers are asked to rate how often a series of nine events occurred outside of specific lesson instruction. Events include such items as asking students for solutions to a problem in the classroom, modeling problem-solving or anger-management strategies, and intervening in a student conflict by prompting students to use problem-solving or anger-management strategies. Second Step developers recommend completion of the SELC once a month throughout the program year.

Other fidelity measures should assess the competency level of trained teachers or other staff on the concepts of Second Step and Building Strong Families training. These measures are likely to be post-tests tailored to the specific concepts of each training that can be conducted both immediately after training and, ideally, at a later point during implementation to assess the degree of retention.

VIP developers also should conduct observations during implementation site visits to assess the quality and frequency of interactions between case managers and fathers and the quality of Building Strong Families sessions. For reasons of confidentiality, evaluators will be unable to observe counseling sessions, but a fidelity measure should capture the percentage of clients engaged in counseling services and the frequency and consistency of the clients' participation.

Finally, developers will have to develop measures of the level of parents' participation in the Building Strong Families sessions and the level of fathers' engagement in case

management services. As the FFP example indicates, programs will have to achieve a sufficiently high threshold for participation in order to generate effects of the magnitude that can significantly change child outcomes.

Intermediate outcomes. The range of activities in the VIP program present program evaluators with a number of potential intermediate outcomes that could be examined (Table III.4). Part of the measurement work during the development stage may be to prioritize the usefulness and value of different intermediate outcomes. In the development stage, evaluators may be interested in collecting a broad array of outcomes to experiment with different measures and methods of data collection. However, they should consider the cost and resource implications that a large data collection effort could have on a larger-scale evaluation.

Center and classroom environment. One tool that has been used in previous process evaluations of Second Step and that could inform development of an intermediate outcome measure is a Playground and Lunchroom Climate Questionnaire. This 23-item, 5- to 10-minute survey is designed to collect information about the structure of, monitoring of, and staff collaboration in unstructured settings in which problems among students may occur. The concept of measuring the climate of unstructured settings is an important one; however, this tool was developed to identify areas of concern and to suggest methods of improving the facilitation of unstructured activities. It may be a useful starting point, but it probably will not achieve the full objective of an intermediate measure to capture the social-emotional climate among and between children in unstructured settings. Other intermediate measures related to changes in the classroom that could result from Second Step will require some development including an observational tool that measures time spent performing specific classroom activities and/or child-directed tasks and the number of disruptions in the classroom over a specified period. Evaluators also should consider including a global measure of the quality of the classroom environment; they have the option of choosing from a number of existing tools with good psychometric properties, such as the Early Childhood Environment Rating Scale, Revised Edition and the Preschool Program Quality Assessment.

Home environment. A number of the intermediate outcomes that would assess changes in the home environment were presented in our discussion of FFP. Details about measures of parenting stress, household conflict, father presence, father involvement with the child, and increased parenting support and warmth are also presented there. In addition to those outcomes, evaluators should collect information about the level of financial support that the father provides to the child, as the VIP program developers believe that this level could increase as a result of the case management services provided to fathers.

Mental health and substance abuse problems. Additional intermediate outcomes are process measures that assess the extent to which discussion of mental health and substance abuse issues in Building Strong Families sessions may lead to screening and treatment. These measures can be collected from program data (paper flow or electronic). The last intermediate outcome shown in Table III.4 assesses the change in parents' psychological well-being that could result from the receipt of mental health counseling. Two existing and tested measurement tools are suggested for this purpose.

Child outcomes. Particularly when selecting child outcome measures in the area of social and emotional development, evaluators must carefully consider the theory of change underlying the enhancement and must select the measure or measures that best fit with that theory. For example, is change expected on a particular social-emotional outcome, such as empathy or problem solving, or is change expected on social competence more broadly, such as by decreasing problem behaviors and increasing positive social relationships? The answer for an evaluation of VIP may be mixed. The Second Step curriculum focuses on specific skills, such as empathy, impulse control, anger management, and problem solving, so measures that can capture practice of these skills could be important. However, the three other components of VIP are intended to affect the child indirectly through improvements in the home environment that contribute to social and emotional competence more broadly. Certainly, children who do acquire the skills taught through the Second Step curriculum might be expected to improve in broader measures of social and emotional competence as well.

Second Step developers and previous evaluators currently are using two direct child measures that might be useful tools for measuring child outcomes for the VIP program. The Second Step Knowledge Assessment for Non- or Beginning Readers uses black-and-white pictures to depict social situations, and to assess children's social-emotional knowledge and skill. The assessment tool uses a story-and-question format similar to the format of the Second Step lessons. It also has been used in evaluation research with preschool and kindergarten students. The tool could provide useful information as a measure of children's knowledge about the concepts of empathy, problem solving, management of strong emotions, and ways to respond to problematic situations with peers that are taught in Second Step lessons—children who show a learned competency in these social-emotional skills may be more likely to practice these skills on their own. Another option for measurement of this outcome is the Social Problem-Solving Test-Revised, which is another direct child measure designed to assess the quantitative and qualitative dimensions of social problem solving. Although it may be useful to examine this existing tool, the assessment tool that is specific to the Second Step curriculum may be the most productive route.

Many other good options exist for measuring aspects of preschool-age children's social and emotional development. In Table III.4, we present a few measurement options for the more specific skills of empathy, self-control, and problem solving, as well as broader measures of behavioral problems and prosocial behavior. These outcomes are typically measured through parent and teacher reports. However, when measuring social relationships, observational tools may be used to directly observe a child in play tasks with either another child or a parent. The Empathy Subscale of the Social Skills Rating System is a validated subscale based on teachers' reports. Measures of self-control often can be thought of as measures of social relationships more broadly; however, the Self-Control Subscale of the Social Skills Rating System measures self-control more directly and is a valid and reliable multi-item measure of children's self-control at home (the parent report) or in the classroom (the teacher report). Problem-solving skills could be measured using the teacher report of the Social Problem-Solving Test-revised. This tool is a complementary one to the direct child assessment discussed above. However, rather than directly testing a child's knowledge, the teacher report assesses a child's actual practice of problem-solving

skills. Our example of the FFP included a discussion of the broader measures of behavior problems and prosocial behavior that are presented in Table III.4. We refer you to that discussion for details on these measures.

d. Activities and Length of Stage 1 for the VIP Project

Circles of Care has developed what they consider to be a comprehensive approach to promote the social and emotional development of Head Start children combined with the integration of mental health problem prevention and intervention services for both children and their families. At the end of the six-month planning period that is supported by I&I funding, Circles of Care will have further refined and promoted its model and intends to implement VIP in five Head Start programs in Brevard County, Florida. During implementation, Circles of Care should use the experience of the five sites to develop and refine documentation so that the enhancement is portable to other Head Start programs. Because of the breadth of this enhancement, substantial measurement work must be undertaken. This work can begin at any time and measures can be tested in the existing sites, as appropriate.

The VIP program currently is administered by one entity—Circles of Care—that is ostensibly capable of providing the range of services comprising the full enhancement. As in FFP, the VIP program will have to be implemented in sites beyond its core service area to ensure that Head Start programs with diverse resources and in diverse communities can identify the appropriate partners to support and sustain the full VIP model. This expanded implementation should occur only after documentation has achieved a sufficient level of depth and clarity to assist other programs in replicating VIP, and only after measurement tools have been identified or developed. The second round of implementation sites should inform replication efforts as well as serve as testing grounds for the measurement framework.

If it can meet the criteria for progression to Stage 2, the VIP enhancement could be an interesting candidate for a planned variation study. A number of evaluation studies already have been completed on the Second Step curriculum that suggest promising results. One pre-post outcomes study of 109 predominantly African-American and Latino three- to seven-year-old children from low-income urban families found that children demonstrated an increased conceptual knowledge of social skills, a decrease in observed levels of physical and verbal aggression, and a decrease in disruptive behavior (McMahon et al. 2000). A rigorous evaluation of the VIP program could include both sites that implement only the Second Step curriculum and sites that implement all the VIP program's components. This type of approach could test the added value of a full family strategy versus a direct classroom approach.

CHAPTER IV

SMALL-SCALE EVALUATION: MATHEMATICS CURRICULUM

Mathematics skills are critical for success in today's world. Yet, despite the importance of these skills, studies suggest that low-income preschoolers' mathematics skills lag behind those of their more-advantaged peers by as much as seven months (Starkey et al. 2004). Similarly, mathematics achievement scores of children in persistently poor households continue to lag behind those of children from higher-income households into the early grades of school (Smith et al. 1997). The gap in mathematics achievement is partly a function of differences in maternal education and family income. In a study of preschool-age children in 65 Los Angeles neighborhoods, children's reading and mathematics achievement scores related most closely to the mother's educational attainment and to neighborhood poverty (Lara-Cinisomo et al. 2004). The differences in children's scores related to maternal education and neighborhood poverty could thus stem in part from differences in support for early reading and mathematics skills in the children's home environments. Early intervention programs, such as Head Start, have the potential to provide a grounding in early mathematics concepts that children may not be acquiring at home.

To date, researchers and policymakers have focused more on preschools' support for early literacy skills development than on the ways in which the preschools can support the development of mathematics skills. Ideas about which mathematics skills should be developed—and how that can be accomplished in the preschool classroom, how preschool children should acquire those skills, the development of curricula and measurement tools to chart the children's progress, and the evaluation of alternative approaches—have lagged advancements in the early literacy area. However, interest in supporting children's early mathematics skills is growing, as is knowledge about approaches to developing early mathematics skills among preschoolers. For example, researchers have investigated which mathematics concepts children are developmentally ready to learn (Clements et al. 2004; Cordes and Gelman 2004). They also have developed mathematics-oriented preschool activities and curricula (Casey et al. 2004; Greenes et al. 2004; Sarama and Clements 2004; Sophian 2004; Starkey et al. 2004). Findings from these research efforts suggest that preschool children have some intuitive knowledge about many mathematical concepts that they have obtained through their own experiences; as a result, most adults underestimate the

ability of preschool children to increase knowledge by exploring materials and by interacting with adults who can amplify what the children are doing to inform about numbers, size, shape, time, and measurement (Tudge and Doucet 2004). Even so, many mathematics concepts are beyond the comprehension of preschool children, suggesting that mathematics activities and curricula must be designed carefully (Ginsburg and Golbeck 2004). Several researchers implementing mathematics curricula in preschools have noted that the teachers reported learning a great deal about the mathematics activities that children are capable of performing, as well as about methods of presenting a much broader range of mathematical concepts in interesting ways (Sophian 2004; Griffin 2004; Starkey et al. 2004). After implementation, the teachers reported an increase in the frequency and breadth of mathematics activities in their classrooms, with participation by all children, rather than by just a few.

As another example of an increasing focus on mathematics instruction in preschool, the National Research Council recently published a summary of a workshop on mathematical and scientific development in early childhood (National Research Council 2005). The workshop was intended as an initial step in exploring the research base on children's cognitive capacity in mastering mathematical and scientific ideas and skills. The workshop focused on availability of curricular and resource materials for early childhood settings in an effort to map the links between research and practice. Workshop panelists concluded that there is a marked lack of connection between research and practice in this field. The research base was found to be weaker than that in literacy with less well-developed basic and applied research, as well as a dearth of longitudinal studies. The NRC concluded that more research in all these strands would be valuable, as well as a synthesis study that pulls together the existing strands of research knowledge.

Mathematical skills are likely to be emphasized more strongly in Head Start classrooms as programs consider the results of the first year of Head Start National Reporting System (NRS) assessments, which included both the literacy and mathematics domains. In particular, Head Start programs are considering ways to strengthen their mathematics content to match recent efforts to improve language and literacy activities in the classroom. Yet, Head Start programs have little information about which mathematics activities or curricula might be most useful. To date, the research on available mathematics curricula has involved small numbers of classrooms and, in most cases, designs that are not rigorous (for example, comparison groups or pre-post designs). In the Preschool Curriculum Evaluation Research (PCER) project, one group of researchers is evaluating two mathematics curricula using a rigorous design.[1] Site-level sample sizes in the PCER projects tend to be sufficient to

[1] Six other PCER sites are evaluating general curricula that researchers expect could influence mathematics skills in addition to language, literacy, and social-emotional skills. We do not discuss them in this chapter because they do not focus primarily on mathematics.

detect effect sizes of .20 to .40 for language, early literacy, and early mathematics outcomes and .35 to .45 for behavioral outcomes.[2]

This chapter describes approaches to evaluating alternative mathematics curricula in which Stage 2 evaluation designs are used. These designs are called "small-scale" in contrast to the nationally representative Stage 3 designs, but in general, the sample sizes for these evaluations will be larger than those used in PCER or the Quality Research Center (QRC) Consortium projects so that smaller impacts can be detected. We focus on two mathematics curricula developed for preschools that include written manuals, and that have been implemented multiple times. Several other preschool mathematics curricula, described briefly here, are in various stages of early development. Although we focus on mathematics curricula in this chapter, the approach that we describe can be generalized to any curriculum evaluation.

A. CURRICULUM MODELS

Several approaches to supporting preschool children's mathematics ability have been developed. Two fairly well developed ones are being evaluated in preschool settings as part of the PCER project:

- ***Building Blocks*** (Sarama and Clements 2004). Building Blocks provides materials (building blocks, puzzles, and art projects) and ideas for teacher-led activities (songs and stories) to support mathematical activities through which children can enhance their mathematics skills. Topics include numbers, operations, and geometric and spatial reasoning, all including subthemes of patterns, data, and classification/sequencing. The curriculum includes both computer-based activities and concrete activities for each lesson.
- ***Prekindergarten Mathematics Curriculum*** (Starkey et al. 2004). This curriculum includes 27 small-group activities with concrete materials that teachers conduct with preschoolers. The activities cover enumeration and number sense, arithmetic reasoning, spatial sense, geometric reasoning, pattern sense and unit, nonstandard measurement, and logical relations. In addition to teacher activities, the curriculum includes computer-based mathematics games and instructions for setting up a mathematics learning center in the classroom. A set of parent activities that coordinate with the classroom curriculum is available.

Several other preschool mathematics curricula are in earlier stages of development or are narrower in scope. They have been implemented in a few Head Start and pre-kindergarten classrooms, but they would need further replication and documentation to be ready for small-scale evaluation. With additional development, they could offer helpful approaches to providing mathematics education for preschool children:

[2] The effect sizes shown here represent the minimum impact as a percentage of the standard deviation of the outcome measure that can be detected at the site level.

- ***Weekly Mathematics Activities for Head Start Teachers*** (Sophian 2004). The curriculum focuses on the concept of unit as it applies to enumeration, measurement, and the identification of relations among geometric shapes. It includes a weekly core classroom project, supplementary classroom activities, and a weekly home activity.
- ***Big Math for Little Kids*** (Greenes et al. 2004). This set of activities and stories covers number, shape, pattern, logical reasoning, measurement, operations on numbers, and space and navigation. The approach includes a focus on mathematics vocabulary.
- ***'Round the Rug Math*** (Casey et al. 2004). This approach provides six books for preschoolers containing problem-solving adventure stories. Each book focuses on a different mathematics content area, including spatial concepts, shapes and geometry, pattern, measurement, and graphing.
- ***Number Worlds*** (Griffin 2004). This program focuses on developing children's number sense, including concepts of quantities, relative size, counting, and simple operations. It offers a sequenced set of lessons/activities that identify learning goals, suggest discussion sequences, and provide concrete materials for children to work with.

B. STUDY DESIGN

The fundamental issues to be addressed by an evaluation of a new mathematics curriculum are whether the curriculum improves children's early mathematics skills, and whether pursuing that curriculum displaces other classroom activities, such as language, literacy, and social-emotional development, that would otherwise be the dominant activities. This section discusses key elements of a research design to address those questions. We begin by discussing the quality enhancement and its counterfactual, the research questions, sampling strategy, random assignment plan, and sample sizes.

1. The Quality Enhancement and Its Contrast

Head Start programs traditionally have strongly emphasized learning through play and children's social development. During the past decade, language and literacy development received increasing levels of attention in preschool settings. We can be fairly certain that, after implementation of the Strategic Teacher Education Program (the national Head Start initiative to improve teacher's knowledge about activities to promote language development and literacy in the classroom), all Head Start centers added book reading and other language and literacy activities to their class planning. Early mathematics has not received as much

attention, so a targeted mathematics curriculum would likely introduce new topics to most Head Start programs' daily schedules.[3]

Accordingly, an evaluation that contrasts Building Blocks or the Prekindergarten Mathematics Curriculum (the two well-developed preschool mathematics curricula) with normal practice in Head Start classrooms would likely reveal clear differences between the two. Relative to the control classrooms and children, the classrooms given the mathematics curriculum would likely introduce more mathematics concepts to children, and the children would spend more time expanding their understanding of mathematical topics. Control classrooms would spend relatively more time in language and literacy activities and in play than in more targeted mathematics instruction.

A much larger evaluation, perhaps in the context of a broader field test (Stage 3), could support a comparison of one curriculum relative to another by randomly assigning specific mathematics curricula to specific centers. At the small-scale evaluation stage (Stage 2), it would be more useful to focus the evaluation on a single curriculum, and to obtain solid evidence about whether that curriculum works in the set of programs participating in the evaluation. In either design, the activities of the counterfactual would be important to document.

A less-defensible alternative evaluation design might measure the impact of implementing a broader set of mathematics curricula relative to current Head Start practice. If Head Start teachers were asked to devote more time to mathematics, they might choose a curriculum on their own. Thus, the evaluation would call for teachers in the intervention group to select a mathematics curriculum to implement. The research question in this case would be, "Does implementing a preschool mathematics curriculum selected by the teacher improve children's mathematics ability relative to current Head Start practice?" This design would not support a comparison of one curriculum relative to another, as curricula would be selected by teachers and programs, rather than be randomly assigned to them. Because the intervention in this design cannot be clearly described or defined, we do not recommend this approach.

Thus, we recommend evaluating a specific mathematics curriculum for two reasons. First, the alternative mathematics curricula vary in their intensity and focus, so it would be more useful to learn from the evaluation that a particular curriculum can or cannot work, rather than to potentially confound curriculum efficacy with program characteristics that influence curriculum choices. For example, programs that are not well prepared to implement a new mathematics curriculum might choose a less comprehensive, less demanding one, thus confounding the independent effects of weaker programs and weaker

[3] The two integrated curricula most commonly used in Head Start, High/Scope and Creative Curriculum, have been revised to include greater concentration on early mathematics topics and approaches. However, if Head Start programs have not implemented more recent versions of the curricula or if they did not receive recent technical assistance and training on the curriculum, the mathematics components of their programs likely will be weak.

curricula. Second, implementing the various curricula well could be difficult if different centers choose different curricula. In addition, implementation would be more cost-effective if the centers implementing a particular curriculum were geographically close, providing easier access for the curriculum developer to provide assistance.

2. Research Questions

The first set of research questions focuses on direct impacts of the mathematics curriculum on classroom activities and child outcomes:

- Does the use of an early mathematics curriculum increase the amount of time that the teacher spends discussing mathematical concepts in class?
- Does the use of an early mathematics curriculum increase children's number sense, knowledge of geometry, and spatial sense?

The second set of research questions focuses on whether the mathematics curriculum has displaced other activities, such that mathematics knowledge increases at the expense of language and literacy activities, and on whether the shift in focus has affected children's progress in those areas:

- Does the use of an early mathematics curriculum reduce the amount of time that the teacher spends on language and literacy activities or class time spent in play and prosocial activities?
- Do children in classes with an early mathematics curriculum make less progress in language and literacy or in social-emotional development (relative to children in classes without that curriculum)?

By Stage 2, we expect that curriculum developers have honed their techniques for implementing the curriculum in preschool (particularly Head Start) settings so that implementation is likely to be successful. Nevertheless, understanding the impacts of the curriculum on children's development and on classroom activities hinges on whether or not the training and technical assistance process led teachers to implement the curriculum with high fidelity. The larger number and diversity of centers and classrooms involved in the Stage 2 study might lead to new challenges, either for the curriculum's "fit" with the program or for the training and technical assistance procedures. The evaluation should, therefore, examine whether the curriculum was implemented to high fidelity, what steps were taken to get to that point, and what was learned from the process:

- Was the curriculum implemented fully and with a high degree of fidelity?
- What strategies were used to implement the early mathematics curriculum? How much training was provided, and over what time period? What amount and types of technical assistance were provided?

- What challenges were encountered, either for the curriculum's "fit" with the program or with training and technical assistance procedures, and how were they resolved?

Finally, although the evaluation will examine whether the mathematics curriculum is effective overall, the Head Start community also will be interested in whether the curriculum is effective across different subgroups of children and families, and across different program designs:

- What were the impacts of the early mathematics curriculum on key subgroups of children? What were the impacts for Head Start programs with different characteristics?

Some caution should be exercised in conducting subgroup analyses in Stage 2 because if many subgroups are examined, some impacts will emerge simply by chance. Moreover, since the sample of grantees and centers is not representative of the broader Head Start population, the subgroups are similarly not representative. Drawing conclusions about how well the curriculum works in specific subgroups would require that the sample frame is designed to include a representative sample from those subgroups. However, this approach would not be consistent with the Stage 2 design, which is to conduct an evaluation that yields internally consistent impact estimates for a group of program grantees and centers that volunteer to participate in the study. Impact estimates for subgroups in a Stage 2 design are valid for the subgroups attending the centers included in the sample. Obtaining a representative sample of specified subgroups would increase the sample size requirements of the study. Thus, serious investigation of subgroup impacts must wait until a Stage 3 evaluation, which will examine a sample that is representative of Head Start programs at the regional or national level.

3. Major Activities and Timetable for the Evaluation

Several activities must occur to carry out the evaluation:

- Draft study design and protocol for recruiting program grantees and centers and submit for Office of Management and Budget (OMB) review
- Identify programs and centers to participate
- Develop data collection instruments and prepare and submit review packages for the Institutional Review Board (IRB) and OMB
- Conduct and monitor random assignment
- Train teachers to implement the curriculum in randomly selected centers or classrooms; provide technical assistance and additional services and supplies needed for implementation; monitor fidelity to implementation
- Identify a sample of children who have consent to participate

- Collect data from enhancement and control groups
- Analyze the data and report findings

We recommend that the first four tasks occur during the first year of the evaluation, with implementation, sampling of children and collection of data during the second year, and further data collection, analysis, and reporting during the third year (see Figure IV.1). Because the Head Start year typically runs from August or September through May or June, the timing of activities will proceed most smoothly if the evaluation activities begin in January or February. We discuss the steps in more detail in the rest of this chapter.

The first year of evaluation activities will be dominated by OMB review and recruitment of grantees and centers. OMB review of the study design and protocols for recruiting grantees and centers must be completed before researchers and curriculum developers can obtain information about prospective centers or negotiate agreements to participate. We have estimated two months to draft and submit a package to OMB that includes the study design and recruiting protocols, and six months for OMB review. During the OMB review period, researchers can conduct activities that do not involve information-gathering from prospective centers. For example, data collection protocols can be developed and submitted for OMB review and the study design and data collection plans can be submitted for IRB review. Data systems for tracking children and managing evaluation data can be developed. ACF can inform the Head Start community about the pending evaluation. Following OMB clearance, the curriculum developer and researchers will contact interested grantees and centers and begin the recruiting process. We estimate that recruitment, executing agreements, and randomly assigning centers or classrooms will require approximately four months.

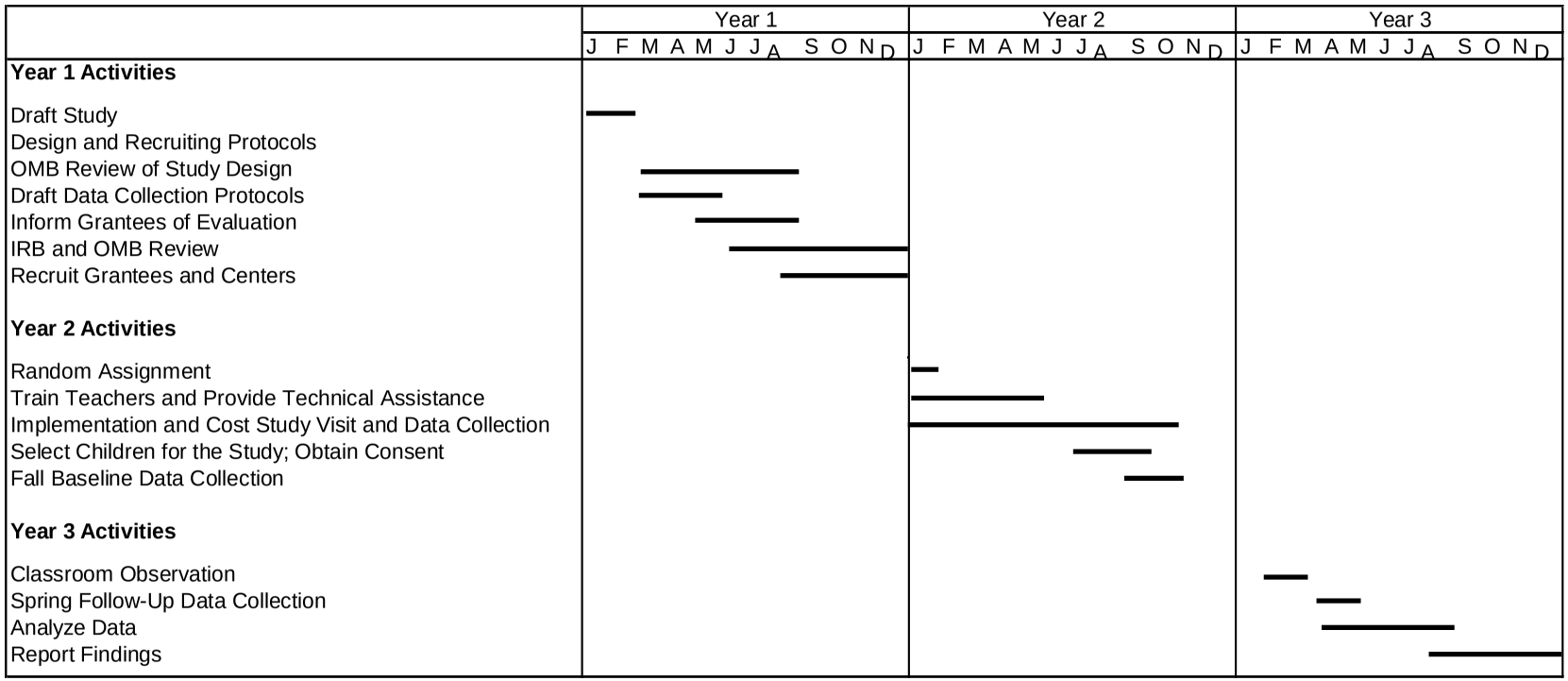
The main evaluation activities in the second year will be implementing the curriculum to high fidelity, conducting the implementation study, obtaining consent for children to participate in the study, sampling children, and collecting fall baseline data. Ideally, implementation will occur in the spring so that the Head Start classes can benefit from a fully implemented curriculum during the entire year. Third-year evaluation activities will include collecting follow-up data in the spring on classrooms and children, analyzing data, and reporting the results.

4. Sampling, Random Assignment, and Sample Sizes

Stage 2 designs generally include grantees and centers that have agreed to participate in the study and were selected because of their location or interest rather than their ability to represent Head Start centers and grantees regionally or nationally. At this stage, the evaluation focuses mainly on whether the quality enhancement *could* be effective in a set of those grantees and centers. As we discuss in more detail in Appendix B,[4] because the

[4] If impact estimates were expected to be externally valid (representative of all Head Start centers), then sample sizes would have to be larger; see Appendix B for details.

Figure IV.1. Schedule of Activities for Small-Scale Evaluation of a Mathematics Curriculum Three-Year Study Beginning in January



emphasis is on internal validity, or a rigorous test of the curriculum within the set of grantees and centers in the evaluation, the sample size requirements are lower than if grantees and centers were selected to represent all Head Start grantees and centers. For Stage 2 designs, a set of programs and centers interested in participating must be identified; random assignment must be completed, and classrooms and children must be selected to participate in data collection.

Identification of Grantees and Centers. Various methods could be used to recruit grantees. Methods of recruiting a broad set of programs include working through the Head Start Bureau to communicate with all regional offices; sending a fax to all programs; or sending an email sent to program directors. However, Stage 2 evaluations ideally should be geographically focused to support careful implementation of the enhancement, so more-targeted methods probably would be more effective than would broadly disseminated flyers or similar methods. With support from the Head Start Bureau and selected regional offices, the curriculum developer and research team also could approach directors of Head Start programs within a geographic area and ask colleagues in other areas who might help with implementation to contact program directors in their own areas. If a broader call for participation is made first, interested programs could be asked to help the curriculum developer approach other programs in their areas. The initial contact materials should briefly explain the study's focus (in this case, a new mathematics curriculum) and should briefly summarize the benefits of participation. The contact materials also might indicate that all centers (or classrooms) in the program will have an equal chance of participating in the study, but that only some of the ones in each program will be chosen to implement the curriculum. Those not chosen to implement the curriculum in the first year will be given priority to implement it, if desired, after the follow-up data have been collected. Programs will receive information about the study's findings, and that they will be partners in the research study. We expect that several programs will want to learn more about the evaluation as a result of this recruiting effort.

Ideally, program grantees and centers chosen for the study would have the following characteristics

- Centers are clustered in a small number of geographic locations to simplify implementation.
- Centers are not already implementing a targeted mathematics curriculum but are interested in doing so.
- Grantees and centers are willing to cooperate with the study's random assignment and data collection requirements.
- The centers currently follow a similar curriculum—either one of the major curricula followed by Head Start programs or a locally designed curriculum—to provide a more stable counterfactual across evaluation sites.

Recruitment of grantees and centers will be easier if their staff believe they will gain more from participation than they might lose. Accordingly, after the initial contact has been made, the benefits to sites must be explained in detail, and concerns about study burden

must be discussed. Curriculum developers and research staff should visit program directors and other relevant administrative staff to discuss the curriculum, its expected benefits to children, what will be involved in implementing it, and the research aspects of the evaluation. Researchers will explain the program's role in ensuring that the study yields useful information about the curriculum's effectiveness. For example, researchers will have to work with program staff to obtain information required to implement random assignment (for example, the number of classes in each center, teachers' names, class sizes, and children's ages). Program staff will have to maintain random assignment statuses (for example, by ensuring that teachers in the intervention group do not share information about the curriculum with control-group teachers).[5] Researchers will have to monitor the integrity of random assignment over time, a key piece of information to demonstrate the reliability of the study. They also will have to work with program staff to schedule child assessments and classroom observations. These evaluation-related requirements are balanced by the chance to implement a new curriculum that could benefit children and the chance to work with the curriculum developer to ensure that the curriculum fits the needs and interests of Head Start children and can be incorporated easily into Head Start program activities. Benefits to the control group are more challenging to identify, but an important benefit that could be offered for control-group participants in a Stage 2 evaluation is first priority to implement the curriculum once the children participating in the evaluation finish the Head Start year.

A program's agreement to participate should include a memorandum of understanding that describes the benefits to the program, and that specifies the respective responsibilities of the curriculum developer, researchers, and program in this joint research undertaking. A similar agreement should be developed with each participating center. This level of detail ensures that misunderstandings that have the potential to threaten the success of the research study do not arise after it is too late to resolve them. A memorandum of understanding also offers a useful way of informing new staff about the study.

Other Program, Center, and Child Sample Selection Criteria. Because group-randomized designs require relatively large sample sizes to detect impacts of moderate size (20 percent of the standard deviation of the outcome measure), it can be costly to expand the sample to provide sufficient power to detect similar levels of impact in subgroups. For this reason, the programs, centers, and children to be included in the sample should be defined carefully, with the most important research questions in mind when doing so.

One sample-definition strategy is to include a broad set of programs with different characteristics. This approach would address the question of the curriculum's effectiveness across the range of Head Start programs and families. In Stage 2 evaluations, however, the

[5] If teachers in the enhancement group discuss the enhancement strategies with control-group teachers, the control group could implement a version of the enhancement, and the evaluation would not provide a valid test of the impacts of the enhancement. Impact estimates would be biased downward, making it more difficult to conclude that the enhancement was effective. To avoid such contamination of the control group, we recommend implementing enhancements at the classroom level only if the potential for spillover from enhancement to control-group classrooms is small.

programs will not have been sampled randomly from all Head Start programs, so the sample will not truly reflect the diversity of programs. If the evaluation finds that the curriculum is not effective, it will be difficult to determine whether it could be effective in some program subgroups.

A more useful strategy is to target programs with a narrower set of characteristics in common, so that the evaluation measures the effects of the curriculum in a more defined set of circumstances. One possibility is to focus the evaluation on centers with full-day (six hours or more per day) programs, which would provide the greatest likelihood of finding improvements in children's mathematics ability without displacement of gains in other areas. Alternatively, given that about half the children served by Head Start attend half-day programs, the evaluation should investigate whether a mathematics curriculum can be effective in both full-day and half-day programs without displacing gains in other areas. Therefore, we recommend that the group of participating centers include an approximately even split between full-day and half-day programs.

Children's age is another characteristic that might define the evaluation sample. Including three-year-old children in the sample could be useful if the curriculum includes activities for this age; in addition, starting at a younger age could be helpful to children. Nevertheless, for a small-scale evaluation of a mathematics curriculum, we recommend limiting the sample to four-year-olds. If both three- and four-year-old children are in the sample, most analyses would have to focus separately on each age group, as mathematics performance is likely to be different in the two age groups. Focusing on age groups in this way effectively reduces the sample available for analyses or would require incurring the cost of doubling the sample size. Moreover, many three-year-old children from low-income families cannot achieve a basal score on the measures of mathematics or language typically used in these evaluations, so it might not be possible to gauge their abilities at the end of the first intervention year.

An important research question is whether the curriculum is effective across the major ethnic groups participating in Head Start (African Americans, Latinos, and whites). During the curriculum development stage (Stage 1), the curriculum developer should adapt the curriculum for each ethnic group, so that a small-scale evaluation will appropriately include all three of them. However, Stage 2 evaluation designs are unlikely to include sufficient sample to support subgroup analysis of the impacts of the curriculum in all three groups. The evaluation will reveal whether impacts are very large in any of the subgroups, but the subgroup samples will probably be too small to allow for detection of impacts of modest size. Attaining large enough sample sizes of different ethnic groups to detect modest-sized impacts might not be feasible until a Stage 3 evaluation of a curriculum.

Random Assignment. Either centers or classrooms could be randomly assigned without requiring changes in the fundamental research questions. The choice of center- or classroom-level random assignment will influence sample size requirements (and, thus, evaluation costs), as well as the potential for spillover between the intervention and control groups. Sample sizes do not have to be as large if random assignment is conducted at the classroom level than if it is conducted at the center level; randomly assigning smaller units

generally increases the power of the design (see Appendix B). In addition, classrooms within one center are more likely to be similar in population served and comparison curriculum used than are two centers. Random assignment at the classroom level might also offer a slight advantage over center-level random assignment at the recruiting stage. Center directors might prefer the 100 percent chance that at least some classrooms in the center can implement the enhancement under classroom-level random assignment over the 50 percent chance that no classrooms will implement the enhancement under center-level random assignment. Moreover, having both enhancement and control classes within their center enables directors to gain a sense (although not rigorous evidence) of how the curriculum seems to be working.

However, those are the only advantages to randomly assigning classrooms. In fact, for several practical reasons, randomly assigning classrooms can present more difficulties for an evaluation than if centers are randomly assigned. First, if the curriculum is implemented in randomly selected classrooms, care must be taken to prevent spillover from occurring. Spillover occurs if control-group teachers implement the curriculum or use techniques based on the curriculum. Thus, teachers implementing the curriculum will have to understand that the distinction between the enhancement (curriculum) and control classrooms will be eroded if they discuss curriculum-related activities and approaches with teachers in the control group. If control-group teachers implement even part of the curriculum, treatment-control differences in outcomes cannot be considered valid estimates of the curriculum's impacts. Notably, the requirement not to discuss the curriculum and methods with teachers in the control group tends to inhibit the discussions that would take place in a full-center implementation that can enhance the overall success of implementation. Second, classroom configurations and teachers change from year to year—and even from May to August—as enrollments are finalized, available space in the preschool is negotiated, and teachers' plans become settled. These changes would present fewer difficulties for the evaluation if an entire center were to either follow the enhancement or continue their usual practices than if individual teachers or classrooms were randomized to one group or the other. Finally, if, as expected, teachers receive training on the curriculum before children are assigned to classes, and if classes are the unit of random assignment, then researchers must be especially vigilant to ensure that children are placed randomly, rather than strategically, into enhancement or control classrooms. Child assignments would not be a problem if centers were randomly assigned, as decisions about children's placement in centers are based on such factors as center locations and schedules relative to the family's residence and schedule.

Random assignment should take place during the year before outcome data are collected, so that curriculum developers can fully implement the curriculum and ensure fidelity.[6] Under this schedule, if initial implementation begins in January or February of the program year, teachers will have time to learn and practice the new techniques before the next school year begins.

[6] We discuss implementation further in Section C.

If centers are to be randomly assigned, participating programs will have to give the researchers a list of participating centers, so that random assignment can begin. Within a single grantee, centers would be assigned in pairs to the enhancement group and to the control group. For example, if a grantee has eight centers, four centers would be assigned to each group. If classrooms are to be randomly assigned, participating programs and centers must provide a list of teachers' names so that teachers can be randomly assigned to enhancement and control groups. If the center has two classrooms, one classroom would be assigned to the enhancement group, and one to the control group. Similarly, four classrooms would be assigned evenly to each group. If the center has an odd number of classrooms, more classrooms should be assigned to the enhancement group than the control group. Then, if a teacher leaves the program, the replacement teacher can be randomly assigned to either enhancement or control group.[7] If the assignment is to the enhancement group, the teacher should receive training and technical assistance so as to be able to implement the enhancement as soon as possible.

Selection of Children into the Research Sample After Random Assignment. Although the evaluation design may call for random assignment of centers or classrooms, it will not be necessary to collect data from all of the children in a classroom or from every classroom and child in the center. Including 10 children per classroom (or 15 per center, if centers are randomly assigned) will provide a sample large enough to estimate average outcomes under random assignment of either classrooms or centers. Beyond that number, additional children add data collection costs at a constant rate but add very little to the power of the design to detect impacts. Therefore, Stage 2 designs typically will involve sampling of children within classrooms (if classrooms are randomly assigned) or sampling of children within centers without regard to classroom (if centers are randomly assigned).[8]

Sample Sizes. Table IV.1 shows the number of centers, classrooms, and children required for the evaluation of a mathematics curriculum under alternative designs. Randomly assigning classes is more efficient than is randomly assigning centers, as the former designs require fewer centers and fewer children than do the latter ones for each level of minimum detectable effect size and assumptions about the extent of baseline data. However, as detailed above, such designs also allow more opportunity for spillover effects that can cause an evaluation to fail.

If centers were randomly assigned, and the desired level of precision in detecting impacts was .20 (one-fifth of a standard deviation on the outcome measures), we would have to randomize 74 centers (37 to each group) and would have to randomly select 1,233 children from those centers, assuming the availability of only minimal baseline data.[9]

[7] Randomly assigning new teachers avoids the chance that new teachers will be recruited with the specific needs of the enhancement in mind, a practice that would bias the results of the study.

[8] The plan for selecting children after random assignment of centers or classrooms is discussed in Section D.

[9] We assume that the evaluation will include 16 or 17 children per center in the initial sample, with a 90 percent response rate to the spring follow-up data collection.

Table IV.1. Sample Sizes Required for a Stage 2 Evaluation of a Mathematics Curriculum Under Alternative Designs

	Total Number of Centers	Total Number of Classes	Initial Sample of Children
Randomly Assign Centers			
MDE = 0.15			
No baseline	166	Any	2,767
Minimal baseline	132	Any	2,200
MDE = 0.20			
No baseline	94	Any	1,567
Minimal baseline	74	Any	1,233
Randomly Assign Classes			
MDE = 0.15			
No baseline	115	230	2,530
Minimal baseline	92	184	2,024
MDE = 0.20			
No baseline	65	130	1,430
Minimal baseline	52	104	1,144

Note: Sample size calculations assume that the evaluation includes 10 children per class (under classroom random assignment) or 15 children per center (under center random assignment), and that 90 percent of the initial sample is available at followup. "Minimal baseline" means that demographic information is collected as part of the consent form, and that NRS fall outcome data are available.

Sample size calculations also assume a two-tailed test of statistical significance at 80 percent power and a 5 percent significance level and that centers or classrooms are divided equally among treatment and control groups. The sample size calculations do not include an adjustment for the design effect of weighting for sample nonresponse.

MDE = minimum detectable effect; NRS = National Reporting System.

Baseline demographic data allows us to control for these characteristics statistically when we estimate impacts, which reduces the variance of the impact estimate and thereby increases the power of a particular sample size to detect impacts. (Section E describes the statistical models that could be used in the analysis of impacts.) Minimal baseline data would include demographic information about the family collected as part of the consent form and children's scores from the fall NRS assessments.[10]

[10] Currently, NRS assessment scores are available only to individual programs aggregated to the program level. Making the scores available for evaluation in the way that we describe would require approval from the Administration for Children and Families and consent from the families.

If classrooms were randomly assigned and the desired level of precision in detecting impacts was .20, the initial sample would have to contain 104 classrooms (52 per group) in 52 centers, and a total of 1,144 children (11 per classroom), again assuming that minimal baseline data were available.

C. IMPLEMENTING THE MATHEMATICS CURRICULUM

The goals of implementation in Stage 2 are twofold. First, the curriculum should be implemented to high fidelity in the centers (or classrooms within centers) that were assigned to the enhancement group. High fidelity ensures that the evaluation measures the impact of the curriculum as designed, rather than measures a watered-down version of the curriculum that was implemented incorrectly. Second, procedures and manuals should be clear enough and should cover a broad enough range of Head Start program situations so that, in the future, it would be feasible to implement the curriculum on a broader scale, using Head Start training and technical assistance staff, with only minimal assistance from the curriculum developer. To achieve this goal, the procedures and manuals will have to be revised as necessary after implementation on the basis of information obtained while implementing the curriculum in the 40 to 80 centers. Issues that arise during initial training or during the technical assistance period will have to be addressed and documented, so that the training materials subsequently can be revised in a way that reflects what has occurred in the classroom during implementation of the curriculum.

At the beginning of Stage 2, the plans for implementing the mathematics curriculum and ensuring that teachers are using it properly will be available to the curriculum developer's staff, having been established during the curriculum's development period (Stage 1). This will ensure that at the start of implementation, a clear plan exists to implement the curriculum to high fidelity. Manuals will be available that describe the curriculum in detail, give teachers examples of activities, and offer trainers a plan to work with teachers during an intensive initial period and over time, while the teachers are trying the new methods in their classes. The manuals should indicate the intensity of initial training and ongoing technical assistance, the duration of the training and technical assistance, and the training and technical assistance staff's qualifications. These manuals do not guarantee that the implementation will proceed smoothly in all centers, but that a basic plan exists that has been honed in several Head Start centers. Challenges encountered and resolved during implementation for a Stage 2 evaluation will form the basis for modifying the curriculum or training and technical assistance.

For a small-scale evaluation (consisting of 40 to 80 centers, depending on whether random assignment is at the center or classroom level, and whether minimal baseline data have been collected), we expect that the curriculum developer will coordinate training for the centers and classrooms assigned to the treatment group. Depending on the design, the training and technical assistance staff will work with either half the centers or half the teachers in each center. An evaluation of this size could operate in approximately five geographic locations, each with 8 to 15 centers. The curriculum developer should have a site coordinator in each location and a staff of trainers to offer initial training and on-site technical assistance during the implementation period (January through May of the second

year) and the evaluation year (September through December of the second year and January through June of the third year). The curriculum developer should visit the sites periodically to conduct some of the large-group training, address any issues that arose during training and technical assistance, and monitor implementation.

Prior to implementation of the curriculum, the curriculum developers should communicate with the leadership and stakeholders of the participating programs at all the levels involved, including the program director, education coordinators, center directors, parent Policy Councils members, and others as appropriate. Care must be taken to ascertain the most efficient selection of staff to train. For example, previous curriculum implementation experience has shown that there is often value in training all members of a teaching team—teaching assistants and aides as well as the lead teacher—to ensure optimal implementation and continuation of the curriculum approach in the event of teacher turnover. All relevant staff should at least be informed about the goals and approaches of the curriculum.

Implementing a mathematics curriculum is likely to require three or four days of initial training in a specialized setting (such as a school or university that has facilities available for large and small groups). Training should include an overview of the curriculum and video footage of teachers engaged in activities associated with the curriculum in classrooms of preschool children, so that teachers in the enhancement group can visualize what the curriculum looks like in practice. Training should then cover a series of modules that will describe each element of the curriculum, and that will present activities and materials associated with each element. Trainers could cover the latter training component with small groups that rotate through the modules over a two-day period. Practical exercises involving role playing and hands-on experience with each module should be offered so that the teachers can practice what they have learned. Techniques for direct and observational assessments of children and links between curriculum assessments and required elements of the Head Start Child Outcomes framework can also be covered in training sessions. The training should wrap up with summaries, questions and answers, and discussion of the technical assistance plan.

After training has been completed, technical assistance staff will visit classrooms periodically to observe how teachers are implementing the curriculum, and to discuss any questions or issues. The staff should visit once per week at first, for about one to two hours, and could then taper the visit schedule to every other week and, eventually, to once per month. During the visits, the staff should discuss with the teachers the material that has been covered in previous sessions and the teachers' comfort level with the material. They should periodically complete measures of fidelity to the curriculum and should discuss with the teachers what they have observed, what aspects of curriculum implementation are going well, and what steps the teachers can take to improve their techniques.

1. Measuring Implementation

Researchers will have to monitor random assignment throughout the implementation period to ensure that the enhancement and control groups remain separate. During this

time, they also will measure the process of implementation and fidelity of curriculum implementation. During the evaluation year, they will continue to monitor implementation and fidelity and, as discussed in the next section, will collect measures of teachers, classrooms, and children's outcomes for the impact analysis.

Measures of Implementation and Fidelity. Researchers will visit a subset of the centers in the evaluation to measure implementation processes and fidelity to the curriculum. The number of centers included in the implementation study is generally determined as a compromise between the evaluation budget and the need to measure implementation experiences in several of the centers. Approximately 20 to 30 percent of the enhancement centers should be selected randomly from among those participating. Since the implementation study is intended to provide insights into what implementation strategies worked well and what did not, the implementation staff should help classify programs by their implementation experience (for example, how well and how easily implementation occurred) and centers can then be randomly selected from each of these groups to participate in the implementation study. Visits should take place during the spring of the implementation year (year 2), after teachers have received training and technical assistance. Researchers will conduct classroom observations to measure how closely the classroom practices match the curriculum and the children's level of engagement with the mathematics topics (a list of topics by data collection mode are shown in Table IV.2). Researchers will conduct semistructured interviews with training and technical assistance staff and with key program staff, including directors and education coordinators, as well as focus groups with teachers to explore the content and quality of training, the fit between the curriculum and classroom, and the content and quality of technical assistance. Questions will focus on what aspects of training and technical assistance worked well, what challenges were encountered, how well the curriculum fit with the program, and what aspects of the curriculum, training, and technical assistance might have to be changed.

While enhancements that are evaluated at Stage 2 are expected to have well-developed implementation plans and thus a high degree of success in implementing to high fidelity in the participating centers, it is not possible to anticipate every challenge to implementation. Accordingly, measures of implementation and fidelity are required both to improve the next generation of implementation materials, as described above, and to ensure that the evaluation can measure the effects of a well-implemented enhancement, as we discuss in Section E. Measures of fidelity to the curriculum and the quality of implementation obtained from the implementation study will be used to define subgroups that enable researchers to estimate impacts of the curriculum among classrooms or centers that implemented the curriculum with high fidelity. Since fidelity and the quality of implementation are not experimentally determined (relative to sites that did not implement to high fidelity), these estimates must be interpreted cautiously but they provide a measure of the effectiveness of the curriculum in "high-fidelity" sites relative to the overall impact.

Table IV.2: Potential Implementation Study Topics for an Evaluation of a Mathematics Curriculum

Implementation Study Topics	Data Collection Methods				
	Direct Observation	Program Records and Documents	Semistructured Interviews with Key Program Staff	Semistructured Interviews with T/TA Providers	Focus Groups with Teachers
Initial Staff Training					
What was the content of training?		X		X	X
What were the qualifications of trainers?				X	
What was the intensity and duration of training?				X	
How well did training prepare staff to use the curriculum?	X		X	X	X
How well did the curriculum blend with other classroom activities?	X		X	X	X
How could training be improved?			X	X	X
Curriculum					
How many times each day and each week are mathematics concepts discussed in class?				X	X
How many times each day and each week do children do hands-on mathematics activities?				X	X
Are staff presenting mathematics concepts as frequently and for as long as intended?			X	X	X
Are staff presenting all of the curriculum's mathematics concepts?	X		X	X	X
If not, why not?			X	X	X
Are children engaged by the mathematics discussions?	X		X	X	X
If not, why not?	X		X	X	X
Do all children participate in the activities? What proportion are engaged?	X				
Technical Assistance and Support					
Who provides technical assistance and support? Curriculum development staff? Education coordinators? Others?			X	X	X
What types of support do staff receive in providing enhanced services? Observation and feedback? Modeling? Conferences between teacher and technical assistance staff? Written reports?			X	X	X
How often is this assistance provided?			X	X	X
What topics have are covered?			X	X	X

Table IV.2 *(continued)*

Implementation Study Topics	Direct Observation	Data Collection Methods: Program Records and Documents	Semistructured Interviews with Key Program Staff	Semistructured Interviews with T/TA Providers	Focus Groups with Teachers
What types of questions are raised? What support is needed?			X	X	X
How helpful is the technical assistance?			X	X	X
What other types of support do staff need?			X	X	X
Lessons Learned					
Which aspects of training went well?			X	X	X
What factors associated with programs or teachers affected implementation positively?			X	X	X
What challenges were encountered in training? In technical assistance?			X	X	X
What factors associated with programs or teachers increased the difficulty of the training or technical assistance?			X	X	X
What strategies were used to overcome the challenges?			X	X	X
How well has the curriculum blended into other program activities?			X	X	X
How well has the curriculum seemed to meet the children's needs and abilities?			X	X	X
How did teachers resolve any issues regarding the fit of the curriculum with the Head Start program?			X	X	X
What do Head Start families think of the curriculum?			X	X	X
How can the curriculum be improved?			X	X	X
How can training or technical assistance be improved?			X	X	X

T/TA = Training and technical assistance.

D. OUTCOMES MEASUREMENT AND DATA COLLECTION PLANS

A comprehensive assessment of the impact of a mathematics curriculum on Head Start children would examine the impact of the intervention on children, teachers, and the overall structure of classroom activities. If the curriculum is implemented well, the teacher will be engaged in more mathematics activities with the children; children will show interest in the activities and their mathematics skills will increase over the Head Start year.

Outcome measures selected for the evaluation should have the following properties:

- ***Relevance to School Readiness Goals and the Head Start Child Outcomes Framework.*** The Head Start Child Outcomes Framework (COF) outlines eight domains of early learning and development (see Appendix A), including children's interest and early skills in the various domains of mathematics, including number and operations, geometry and spatial sense, and patterns and measurement.
- ***Sensitivity to Intervention.*** Measures should reflect outcomes malleable to intervention, as opposed to more trait-like qualities (such as temperament).
- ***Appropriateness for a Culturally Diverse, Low-Income Population.*** The selected measures should adequately assess outcomes of low-income 3- to 5-year-olds from diverse cultural populations, including those who do not speak English in the home.
- ***Adequate Psychometric Properties.*** Measures selected should be reliable and valid. Reliability means that they measure the same construct across various settings (for example, the classroom or the home), on repeated occasions (*test-retest reliability)*, when administered or rated by different interviewers/observers (*interrater reliability*), and when subsets of items are administered to identical samples (*split-half reliability*). Validity refers to the degree to which the measure taps the underlying construct it purports to measure, keeping in mind linguistic and cultural considerations.[11]
- ***Valid and Reliable for Intended Mode of Administration.*** A measure might not be valid or reliable as reported if it is used with a group for whom it was not designed or with a mode of administration for which its reliability and validity have not been tested.

[11] *Content validity* indicates that the set of items comprising a measure is a good representation of the construct being measured. *Concurrent validity* refers to sufficiently large correlations between the measure and another measure of the same construct (measured contemporaneously in the same sample). *Predictive validity* refers to a sufficiently large correlation between the child outcome measure and a subsequently measured construct that is theoretically associated with the child outcome.

- ***Prior Use in Large-Scale Surveys and Intervention Evaluations.*** Prior use is helpful because it suggests that the measure is practical and feasible to use in large-scale research and because it provides a benchmark for the scores of children participating in the evaluation.
- ***Cost and Burden.*** The cost or burden of data collection strategies, including training requirements, respondent burden, and program administration burden, should be minimized.

Table IV.3 describes the classroom and child outcomes to be measured and the recommended measures of each. Measures of teachers' knowledge and attitudes about teaching mathematics at the preschool level can be taken from the PCER evaluation and from other recent studies of mathematics curricula. Measures of classroom activities will include the fidelity measure for the selected mathematics curriculum and questions to teachers about how classroom time is allocated among mathematics, language and literacy, and play. Measures of children's development will include mathematics skills (a standardized assessment and a more detailed measure of children's mathematics abilities used in the PCER evaluation), language (the Peabody Picture Vocabulary Test, Third Edition, which could rely on the NRS assessment), early literacy (for example, subtests of the Woodcock-Johnson, such as Letter-Word Identification and Spelling), and measures of behavioral problems and approaches to learning completed by teachers.

To obtain data for the impact analysis, a baseline assessment of children's mathematics skills will be conducted during the fall, with a follow-up assessment conducted during the spring. A teacher interview, also conducted during the fall and spring, will examine fidelity to the curriculum (in both the enhancement and control classes), time spent in mathematics-related activities, knowledge and beliefs about child learning, and the background of the teachers. The classroom observation, which need be conducted only during the spring, will measure fidelity to the intervention (in both the enhancement and control classes) and time spent in mathematics-related activities. Additional observation measures of classroom quality would tap into potential improvements in early childhood practice generally; typically, these focus more on adult-child interactions and resources than on content-specific activities. However, taking stock of these important interim measures is an important step in helping to understand the pathways by which the curriculum may be changing practice and potentially affecting child learning. The child's demographic characteristics will be obtained from an information sheet completed by parents as part of the consent process.

Before any data are collected, the research team must draft instruments, obtain IRB and OMB approval, obtain parents' consent for their children's participation, and randomly select children to participate based on their eligibility status and their consent status. We discuss these steps in the remainder of this section.

Table IV.3: Measures of Intermediate and Child Outcomes Associated with a Mathematics Curriculum

Outcome	Recommended Measure	Type of Measure
Teachers Knowledge		
Expectations for children's mathematics ability	Attitudes and beliefs about children's mathematics ability	Teacher survey
Knowledge about how to present mathematics concepts	Types of mathematics activities conducted in class	Teacher survey
	Frequency of teaching mathematics activities	
Classroom Processes		
Proportion of class time spent presenting mathematics concepts	Time spent discussing mathematics concepts	Observation
	Total time in class	Teacher survey
Teachers engagement in presenting mathematics content	Fidelity measure developed for the mathematics curriculum	Observation
Children's engagement during discussion of mathematics	Number of children listening during presentation	Observation
	Level of engagement of children	
Proportion of class time spent in language and literacy activities; proportion of time spent in play	Time spent in language and literacy activities	Observation
	Time spent in free play activities	Teacher survey
Children's Development		
Mathematics ability (number sense, knowledge of geometry, and spatial sense)	Woodcock-Johnson Applied Problems (Woodcock, McGrew, and Mather 2001)	Direct assessment
	Child Math Assessment, Abbreviated	Direct assessment
Language ability	Peabody Picture Vocabulary Test III (Dunn and Dunn 1997)	Direct assessment
Early literacy	Test of Language Development—Primary—Third Edition, Phonemic Analysis (Newcomer and Hammill 1997)	Direct assessment
Behavioral problems	Behavior Problems Scale (FACES Research Team 2001)	Teacher survey
Approaches toward learning	Preschool Learning Behaviors Scale (McDermott et al. 2000)	Teacher survey

Develop Data Collection Instruments. During the first year of the study, data collection instruments will be written. These include the consent form with a form requesting demographic information; the teacher survey; the child assessments; and the classroom observation protocol. Many of these will include standardized assessments that need to be formatted to simplify administration by a trained assessor. Others (such as the teacher questionnaire) will rely on questions that have been used in prior studies. Once the data collection instruments are developed, they will be pretested to ensure that respondents understand the questions, that the flow proceeds logically and smoothly, and that the time required to complete them is reasonable.

IRB and OMB Research Review. Research on human subjects must be reviewed by an IRB, which considers the benefits of the research to society, the programs, and the participating families and weighs them against the cost of the research to the families and program staff. The IRB also reviews protection of research participants from harm by ensuring that confidentiality is maintained. In addition, if the evaluation is federally funded, data collection instruments and the research plan must be approved by OMB. The data collection instruments are reviewed to ensure that they do not overlap with other ongoing federal data collection efforts, and that burden is not excessive. These reviews will be conducted during the nine months preceding the start of the evaluation year, while the intervention is being implemented.

Consent. Consent for children to participate in the research must be obtained from parents or guardians. The parent consent form will clearly inform parents (guardians) about the duration of the study, the types of assessments that will be administered, and the voluntary nature of participation. An information sheet will be included in the consent package to collect basic demographic information about the family, such as age, race and ethnicity, and family structure.

Because many Head Start classrooms serve mixed-age groups of children, some children in a particular classroom will not meet the study's eligibility requirements. Rather than ask teachers to sort children by eligibility status and to distribute consent forms accordingly, it seems more efficient more accurate, and less burdensome for teachers to request consent from all parents of children in the classroom. The consent form could include a few questions designed to elicit the child's eligibility for the study (for example, birth date), as well as other demographic variables that will have to be collected for the study. Researchers could then review the consent forms to identify eligible children.

The consent process could proceed more smoothly if it is incorporated into the home visits that many programs make to families. During these visits, which typically occur just before children attend class for the first time, teachers bring forms that parents must complete before the start of the school year. The teacher is available to explain the forms, and to ensure that they are completed correctly. If the study's consent is part of this process, the teacher would be able to explain the nature of the study, what will happen if the child participates in the research, and the voluntary nature of the child's participation.

Child Eligibility Criteria for the Study. Children entering the research sample for the mathematics curriculum study should be four years old (or eligible to enter kindergarten in

the following year). Teachers' training on the curriculum will have begun during the year before the child assessment year, and some classes and centers include children of mixed ages. Therefore, children entering the research sample ideally should be in their first year of Head Start, so they will all have the same duration of exposure to the curriculum.

Selection of Children for the Study. Consent forms will be sent by program staff to the researchers for processing. The researchers will enter data from the forms to obtain information on the demographic characteristics of children from each center (who have been given consent to participate). They will then identify which children are eligible for the study (using age, whether new to Head Start, and any other criteria set for the evaluation). A sample of children from the pool of eligible children in the center or classroom will be randomly selected for the data collection. Thus, if the design is to randomly assign centers, children who are eligible and have consent will be listed, and 16 to 17 will be randomly selected for the research sample. If the design is to randomly assign classrooms, children who are eligible and have consent will be listed and 11 will be randomly selected for the research sample.

Parent consent rates typically vary across classrooms, as rates at which parents return the forms and consent can depend on how organized the parents are, their degree of connection with the classroom teacher, whether parents talk to each other about participating in the study, and how diligently the teacher follows up with parents to return the forms. However, the study will typically not have information about children without parental consent; in this case, the study will have to generalize the results based on those with consent to the entire class. However, if some aggregate, class-level data are available (such as demographic composition or scores from the NRS assessments), this information can be used to adjust classroom-based estimates for nonresponse by weighting the responders to reflect the true aggregate classroom-level composition.

Teacher Self-Administered Questionnaire. Teachers' attitudes about teaching mathematics to young children and their level of understanding about four-year-old children's of learning capabilities can influence how well they implements the curriculum. We recommend giving teachers a short (no more than 15-minute) self-administered questionnaire during the fall, while the child assessments are in progress. This questionnaire could be adapted from the PCER teacher interview and could include questions about the teacher's background, attitudes and beliefs about child learning, mathematics activities conducted in the classroom, and the frequency of teaching specific mathematics activities. During the spring data collection, the teacher questionnaire will omit the teacher background questions (unless the teacher was new) but will include a short behavior problems scale and a short approaches toward learning scale for each child in the research sample. The behavior problems scale will address whether the increasing academic demands in preschool have increased children's behavioral problems. The approaches toward learning scale will help identify positive factors such as curiosity and attentiveness that are associated with academic success.

Classroom Observation. Evaluating the time that teachers spend on different mathematics skills, their approaches to these learning experiences, and the proportion of

time spent in individual and group activities will help researchers to understand the pattern of child outcomes. Observational instruments that are available to measure mathematics activities in the classroom have been developed primarily as fidelity measures for specific curricula. We therefore recommend using the fidelity measure developed for the mathematics curriculum to measure the content, duration, and intensity of the curriculum's mathematics activities. In addition, more general measures of classroom quality can be used at this data collection point, in order to assess interim practice quality. We recommend that the classroom observations be conducted during the two weeks before the spring follow-up assessments begin.

Child Assessment. To allow for a more in-depth evaluation of the mathematics domains stressed in an intervention curriculum, we recommend two assessments, the Woodcock-Johnson Applied Problems subtest and the Child Math Assessment—Abbreviated for PCER (Starkey et al. 2002). The Applied Problems subtest includes simple counting, addition, and subtraction operations and requires the child to decide not only the appropriate mathematical operations to use, but which data to include in the count or calculations. It takes five to eight minutes to administer. The Child Math Assessment—Abbreviated differs from the NRS and the Applied Problems Subtest in that it uses manipulatives (three-dimensional materials that children can touch and move around) to assess the child's understanding of object counting, construction of equivalent sets, shape recognition, and pattern duplication. It takes 20 minutes to administer. To address the research question on displacement of language and literacy development, we recommend using the PPVT (from the NRS administration, if available) and the Woodcock-Johnson subtests such as Letter-Word Identification and Spelling (an early writing task which is related to small motor skills).

The importance of obtaining a true baseline child assessment, ideally before the child has any experience in the Head Start classroom, must be balanced against both the need to control data collection costs by assessing children in the Head Start center and the two- to three-week period required to stabilize enrollment in Head Start classes. We recommend a field period that starts approximately two to three weeks after classes begin at the Head Start center, with a six-week window for data collection in the sites. Similarly, to assess what children have learned during the Head Start year, the follow-up assessment should be conducted as close to the end of the year as possible. We recommend that the follow-up data be collected during a six-week window that ends two weeks before the end of the year, and that data collection be matched to the timing of the fall assessment so that classes assessed early in the fall field period also are assessed early in the spring field period.

Costs of the Enhancement. Program administrators care not only about the effectiveness of the quality enhancements, but also about the costs of these enhancements over and above current expenditures. To support analyses of the cost of the enhancement relative to its impacts or benefits, researchers will collect information on the costs of implementing and carrying out the mathematics curriculum relative to the cost of the program without that curriculum. Researchers should measure two types of costs: (1) the upfront, one-time costs of beginning implementation of the curriculum, and (2) the ongoing additional costs resulting from the enhancement. Among the first set of costs are the costs

associated with initial teacher training, including the curriculum developer's and training staff's time, the cost of any substitute teachers hired to cover classrooms while the teachers attend training, and the cost of additional staff days for teachers paid for the days they attend training. The first set of costs also includes costs associated with the curriculum training staff's technical assistance visits and any costs associated with teachers attending extra training sessions outside their normal classroom duties (for example, periodic group discussions, if any). Time spent by teachers and trainers would be valued at their hourly wage, including fringe benefits. If this training actually supplanted the usual teacher training sessions conducted during the year, those routine training costs should be subtracted. Usually, however, new curriculum training will be an add-on expense rather than a substitute. If space for training is rented, the cost of the rental will be included as well. Materials, including documents, videos, and other training materials, should be valued at their cost.

For the mathematics curriculum, the ongoing additional costs resulting from the enhancement would include the costs of materials, any computer-related costs, the cost of ongoing technical assistance, and costs to train new teachers. Teachers will have to maintain a ready supply of materials used to demonstrate mathematics concepts in the classroom, as well as materials for children to play with in a specially-designated area of the classroom at other times. Because broken or lost materials must be replaced, the cost of maintaining the supply of mathematics materials during the program year is a cost of the curriculum. Similarly, if a computer-based software program is part of the mathematics curriculum, the cost of setting up and maintaining the personal computers in the classroom with appropriate software is an additional cost. The cost of refresher training and ongoing technical assistance to teachers during a normal program year to ensure that the curriculum continues to be implemented to high fidelity is an ongoing cost as well. For example, technical assistance staff might make two or three visits to each classroom during the program year to observe, measure fidelity, and meet with the teachers and Education Coordinator to discuss classroom practices and respond to questions. Finally, overall teacher turnover in Head Start is approximately 15 percent per year (National Institute for Early Education Research 2003). Accordingly, an ongoing cost of the mathematics curriculum is the cost of training and technical assistance to 15 percent of the teachers in the enhancement group each year. Assuming that the evaluation includes 50 to 100 classrooms in the enhancement group, 7 to 15 teachers would have to be fully trained each year.

Three different approaches to estimating these costs of implementing the early mathematics curriculum could be used if centers are randomly assigned to either the enhancement or control group:

- Identify the costs associated with the new curriculum (described above) and ask center directors for this information.
- Obtain the full budget for the enhancement center for the year prior to curriculum implementation and for the current year and estimate the cost of the curriculum as the difference in budgets (adjusting for normal cost inflation from one year to the next).

- Obtain the full budget for the enhancement centers and the control-group centers and estimate the cost of the curriculum as the difference between enhancement and control-group center costs.

The first approach is the least burdensome for the enhancement centers (and eliminates burden for the control centers) but could miss costs associated with the curriculum. The second approach is more burdensome for the enhancement centers than the first, and does not require a response from the control centers, but if anything other than the implementation of the curriculum changes from one year to the next, the estimate of the cost of the enhancement will be inaccurate. The third approach is the most burdensome, involving extensive collection of cost information from both enhancement and control centers, but would provide the most accurate estimate of the costs of implementing the enhancement.

If classrooms within centers are randomly assigned to the enhancement or control groups, the first or the second approach to estimating the costs of the curriculum would have to be used. These approaches would provide estimates of the cost of implementing the curriculum in just a few of the classrooms, and would have to be adjusted using reasonable assumptions regarding fixed costs (costs that would remain the same if one or more classrooms in the center implemented the curriculum) and variable costs (costs that would increase if one more classroom implemented the curriculum).

Information about these costs can be obtained from semistructured interviews with program and center directors, from the curriculum developer and training/technical assistance supervisory staff, and from program records. All cost information must be obtained in dollars; however, to ensure that the dollar values collected at various time points represent the same value, the dollar values collected from informants and records should be either inflated or deflated, using the Consumer Price Index, to represent dollar values in a single target year (for example, the analysis and reporting year).

Finally, cost information is sometimes obtained using two perspectives. First, the actual dollar costs of the curriculum must be obtained. Second, a measure of cost to society would place a value on any volunteer labor or donated space and materials and would add those costs to the actual expended costs. Both cost perspectives would be used in the analysis of benefits and costs or cost-effectiveness of the curriculum.

E. ANALYSIS AND REPORTS

Reports based on the evaluation should report the estimated impacts of the mathematics curriculum on teachers' practices, classroom activities, and children's outcomes. They also should describe the implementation experience and should discuss how the curriculum could be implemented on a broad scale, using the Head Start training and technical assistance system for support. Reports should be written for a broad audience, with stand-alone summaries that can be understood by program staff and policymakers, and more-detailed reports that summarize the research design, sample characteristics, analytic

approaches, and findings in a clear and accessible way. In this section, we discuss the approach to estimating impacts and conducting the cost-effectiveness analysis.

1. Estimating Impacts of the Mathematics Curriculum

Using a random assignment evaluation design means that fairly simple estimation methods can be used to determine the impacts of the quality enhancements at a point in time. Under random assignment of classrooms, center-level impacts are calculated as the simple differences in the mean value of outcomes for children in the enhancement classrooms and the control classrooms, and then the center-level impacts are averaged for the overall impact estimate.[12] This provides an unbiased estimate of the impact of the enhancement compared to the status quo. Under random assignment of centers, the center-level mean outcomes are estimated, then, separately for the enhancement and control group, the center mean outcomes are averaged over all centers in that group. The simple difference in the means between enhancement and control centers is the impact estimate.

More-precise estimates can be obtained by estimating regression models. Regression procedures can improve the precision of the estimates and adjust for any residual differences in the observable characteristics of program and control group members due to random sampling and interview nonresponse. Regression models take the following form:

$$(1) \qquad Y = \alpha + X\beta + \gamma T + \varepsilon ,$$

where Y is an outcome variable; X is a vector of explanatory variables; T is an indicator that equals 1 for members of the enhancement group and 0 for members of the control group; α, β, and γ are parameters to be estimated; and ε is a random-error term. The estimate of the parameter γ is the estimated impact of the quality enhancement compared with regular Head Start services.

Because random assignment will have been conducted at the classroom or center level (depending on the design), the regression adjustment takes the form of a hierarchical linear model of child development consisting of two nested levels. By specifying the model at each level, we can conduct analyses for the appropriate units of analysis and can conduct statistical hypothesis tests that correctly account for the clustering of children within classrooms. While not strictly necessary for conducting impacts because the evaluation is based on a random assignment design, adjusting the impacts with an HLM model can help increase the precision of the estimates.

For the design involving random assignment of centers, the analytic model is the following, where the variables are indexed by child (*i*) and centers (*j*):

[12] Center-level impacts are averaged so that larger centers do not dominate the impact estimates.

Child-level model

$$(2) \qquad Y_{ij}(t) = \alpha Y_{ij}(0) + \beta X_{ij} + c_j + \varepsilon_{1ij}.$$

Center-level model

$$(3) \qquad c_j = \gamma T_j + \delta Z_j + \varepsilon_{2j}.$$

For the design involving random assignment of classrooms, the analytic model is the following, where the variables are indexed by child (*i*) and classrooms (*k*):

Child-level model

$$(4) \qquad Y_{ik}(t) = \alpha Y_{ik}(0) + \beta X_{ik} + \ell_k + \varepsilon_{3ik}.$$

Classroom-level model

$$(5) \qquad \ell_k = \lambda T_k + \psi W_k + \varepsilon_{4k}.$$

where *Y(t)* is the outcome at follow-up period *t; X* is a set of child characteristics, such as gender; *T* is a variable indicating whether the child was in an enhancement classroom or center; *Z* is a set of center-level variables, such as whether classes are full-day; *W* is a set of classroom-level variables, such as the teacher's education level; and ε_1, ε_2, ε_3, and ε_4 are disturbance terms assumed to have a mean of zero and to be uncorrelated with each other. Parameters to be estimated include α and β, vectors of coefficients on the child baseline characteristics; *c*, the center effect; γ or λ, the effect of the mathematics curriculum in the two models; δ, the coefficients on the center variables; and ψ, the coefficients on the classroom variables.

The statistical techniques used to estimate regression-adjusted impacts in equations (2) and (4) will depend on the form of the outcome, *Y*. If the dependent variable is continuous (such as the score on the Children's Math Assessment), ordinary least squares methods produce unbiased estimates of the parameter γ or λ. However, if the dependent variable is binary (such as whether the child's score on the Woodcock-Johnson Applied Problems subtest was one or more standard deviations below the norm), logit or probit maximum-likelihood methods will be used to obtain consistent parameter estimates.

The estimation models presented here assume that all centers or classrooms are weighted equally. Thus, in the case of center-based random assignment, the average outcome measure for each treatment center is averaged with those of all other treatment centers to obtain the mean outcome for all treatment centers. Larger centers thus do not receive greater weight in influencing the overall mean score for treatment (or control) centers. If the enhancement were implemented in a few centers with six or eight classrooms and several others with two or three classrooms, we would not want the results in the larger

centers to overwhelm the results for all treatment centers. Averaging results across all centers regardless of center size addresses the question, "Does the enhancement work in the average center?" This approach is appropriate because the purpose of the evaluation at Stage 2 is to measure how well the enhancement works in a collection of centers overall. The same is true if random assignment takes place at the classroom level. The Stage 2 evaluation will address the question of whether the enhancement works in the average classroom, without allowing larger classrooms to have a greater influence on the results than smaller classrooms.

2. Subgroup Analyses

Because the effectiveness of the mathematics curriculum may differ by program setting or by characteristics of the children served, it would be useful to determine the groups for which the curriculum is most effective. This type of subgroup analysis would then enable individual programs to decide which enhancements might be useful to them. For example, analysis may demonstrate that the mathematics curriculum is effective only in programs that offer full-day services or that its impacts are stronger when teachers have higher educational credentials at the outset of training.

The subgroups of interest will depend on the quality enhancement to be tested. Examples of center- or classroom-level characteristics that can define subgroups include:

- Full-day or part-day program
- High-fidelity implementation or incomplete implementation
- Teachers' qualifications
- Center or program size

Examples of categories of child and family characteristics (measured prior to the experience of the quality enhancement) for subgroup analysis include:

- Child's gender
- Child's English proficiency
- Mother's education level
- Level of family income
- Parents' employment status

Subgroup estimates can be obtained using the same procedures described above for calculating overall impacts,[13] but these calculations are made for particular subgroups. Regression-adjusted subgroup estimates can be obtained by introducing an interaction term that is the product of the treatment indicator and an indicator of membership in the subgroup of interest. This term is entered into the appropriate model shown above for the level of random assignment and for whether the subgroup is defined at the child level or at the classroom or center level.

Unless the overall sample is very large, however, it will be possible to detect impacts only for large subgroups of the population, as subgroup estimates, which are based on only part of the full sample, are less precise than are full-sample estimates. For example, in the center random assignment design with minimal baseline data, our sample of 1,233 children (in 74 centers) is sufficient for detecting impacts with effect sizes of .20 (one-fifth of a standard deviation on the outcome measures) or more. However, for a subgroup that includes 50 percent of the children across all the centers (for example, children whose mothers have less than a high school diploma), impacts with effect sizes of .25 or more can be detected. For a subgroup that includes all of the children in half the centers (for example, full-day programs), impacts with effect sizes of .29 or more can be detected.

3. Cost-Effectiveness Analysis

A cost-effectiveness framework should be used to evaluate the costs and benefits of the mathematics curriculum. This type of analysis does not attempt to place a dollar value on impacts. Instead, impacts are measured in a common unit, such as an effect size (the impact divided by the standard error of the outcome measure). The impact in effect-size units is compared with costs measured in dollars. For each quality enhancement, an effect size per dollar spent on the enhancement can be calculated. For example, if the mathematics curriculum were to produce an impact on mathematical skills of 0.3 in effect-size units, and if the cost was estimated to be $10 per child, then the cost-effectiveness of the curriculum would be 0.03 per dollar. Measuring cost-effectiveness in this way enables program administrators to compare the cost of producing impacts they consider important using the same metrics, so that enhancements can be assessed in terms of their ability to provide the most "bang for the buck."[14]

The enhancement is likely to have impacts that vary in size across the outcomes measured. Therefore, the estimate of cost-effectiveness will depend on the outcome used to

[13] For classroom-level random assignment, calculate mean impacts first at the center level, and then average across centers. For center-level random assignment, calculate center-level mean outcomes and average across centers in the enhancement and control groups.

[14] In contrast, a cost-benefit analysis requires researchers to convert impacts into "benefits" that are valued in dollars. This task is complicated by the fact that the evidence linking differences in child assessment scores with future employment and earnings, involvement with the criminal justice system, and use of public assistance programs is quite tenuous. Accordingly, we do not recommend conducting a cost-benefit analysis based on outcome data collected from one year in Head Start.

measure it. Researchers can report the range of cost-effectiveness estimates using the impacts on outcome measures considered to be most important. For example, the outcomes most important for a mathematics curriculum are children's developing mathematical skills. The cost-effectiveness of several enhancements designed to influence children's early mathematics skills could be compared using a common outcome measure of those skills.

CHAPTER V

SMALL-SCALE EVALUATION: APPROACHES TO SUPPORTING CHILDREN'S SOCIAL-EMOTIONAL WELL-BEING

Children's development proceeds along multiple dimensions simultaneously, with gains in one domain supporting gains in another. For example, children's ability to persist in a task and to manage frustration helps them to master new cognitive tasks, such as learning the sounds of letters and sounding out new words. Similarly, children's growing language skills enable them to dispel or avoid frustration by articulating their needs clearly and negotiating conflict, rather than by throwing objects or hitting others. Accordingly, young children's social-emotional development is important in supporting their cognitive development, with the cognitive gains, in turn, supporting prosocial behavior (Raver 2002; Raver and Knitzer 2002; Shonkoff and Phillips 2001).

Key aspects of children's social-emotional development (for example, enthusiasm and curiosity about new activities) are recognized by kindergarten teachers as critical elements of children's readiness for school (Heaviside and Farris 1993). Head Start teachers also have noted that even one or two disruptive childrenmake it difficult to create a learning environment for the class as a whole. Research also supports the notion that children's emotional health is positively correlated with the children's early school success (Raver 2002). Despite the current emphasis on preparing Head Start children for kindergarten through early language and literacy activities, many teachers are concerned that significant behavioral issues must be addressed first.

The Head Start community, consistent with the program's comprehensive approach to enhancing children's development, has taken important steps toward addressing these issues in the classroom. As one of the national consultants that comprise part of the Head Start Training and Technical Assistance System, the Center for the Social and Emotional Foundations of Children's Learning has developed web-accessible guides for creating classroom environments that promote positive behavior, minimize disruptive behavior, and address other behavioral issues. All Head Start programs are required to arrange for the provision of screening and treatment services from mental health professionals in their communities, but some have contracted for more-intensive, on-site, ongoing services. One of the Quality Research Center (QRC) Consortium research teams is evaluating a curriculum

that promotes positive social-emotional behavior as well as language and literacy through the reading and discussion of children's books describing situations calling for impulse control, empathy, or anger management (Administration for Children and Families 2004; Kupersmidt and Bryant 2001). A comprehensive intervention targeting both teachers' and parents' skills in managing children's behavior and children's social and behavioral skills has been implemented in many preschool and Head Start classrooms; evaluations of the intervention have found improvements in children's behavior and in the classroom climate (Webster-Stratton 1998; Webster-Stratton et al. 2001).

This chapter describes approaches to evaluating enhancements designed to support preschool children's social-emotional development, and to minimize disruptive behavior in the classroom, using Stage 2 evaluation designs. As discussed in the previous chapter, in contrast to nationally representative Stage 3 designs, these designs are called "small-scale" designs; in general, however, the sample sizes for these evaluations will be larger than those used in either the Preschool Curriculum Evaluation Research (PCER) project or the QRC Consortium projects. We focus on an approach to supporting children's social-emotional development that includes two elements: (1) teacher training to encourage positive child behavior, and to manage the classroom to minimize disruptions; and (2) an intervention with individual children (and their parents) who exhibit conduct problems (including high levels of aggression, defiance, and oppositional and impulsive behaviors). This chapter goes beyond the discussion in Chapter IV because the evaluation design must include intermediate measures of the intervention's teacher/classroom and parent components, and must measure changes in the well-being of the specific children receiving intensive services and of children who are members of the class more generally.

The approaches we describe to Stage 2 evaluations in this chapter and the previous one can be easily extended to evaluate other quality enhancements that can be implemented at the center or classroom level. For example, initiatives to improve children's health and fitness could be evaluated using these approaches. The steps in planning the evaluation are laid out in these chapters; the specific decisions regarding how to implement the enhancement; how to ensure fidelity; and how to measure outcomes will be specific to the enhancements. Those decisions, in turn, should grow out of prior work to develop the enhancement idea and pilot test implementation procedures.

A. ALTERNATIVE APPROACHES

A diverse set of strategies for supporting children's social-emotional well-being have been developed. The strategies can be classified in part by the breadth of their focus on children: Universal approaches involve all children in the classroom, while individually focused interventions target particular children exhibiting disruptive behavior or behavior indicative of withdrawal. The strategies also can be classified by the adults involved: Some interventions are implemented at the level of the teacher and classroom and might not involve parents to a significant degree; others involve both parents and teachers to improve the consistency and supportiveness of the home and classroom environments; still other interventions involve a mental health specialist working intensively with a child, the parent,

and the teacher. The following examples illustrate the diversity of approaches that have been developed:

- ***First Step to Success*** (Walker et al. 1998). First Step to Success is both a universal and an individual-child-focused intervention, and while mainly emphasizing the classroom, includes the parents of individual children who are targeted. The enhancement consists of training and technical assistance for teachers provided by a skilled clinician. Teacher training includes classroom management, the teaching of social skills to children, and positive and proactive discipline. Students requiring more-targeted intervention are provided a two-month collaborative home and school intervention program delivered by the clinician.
- ***Preschool Behavior Project*** (Kupersmidt and Bryant 2001). The Preschool Behavior Project is a universal intervention that consists of training teachers to establish proactive teaching and behavior management practices in the classroom; to build strong relationships with children; and to teach children to solve social problems constructively through dialogic reading of selected books with themes tied to the Second Step program, which addresses empathy, anger management, and problem solving (Grossman et al. 1997; Whitehurst et al. 1994). Researchers have implemented two versions of the program: one involving highly trained, supervised clinical consultants working with teachers and parents, and a second involving teacher training by less-specialized technical assistance staff from within the program.
- ***The Incredible Years: Parents, Teachers, and Children Training Series*** (Webster-Stratton 1998; Webster-Stratton et al. 2001). This program targets parents, teachers, and children. The program is directed toward individual children with conduct problems (by training parents and working with children) as well as more universally, by training teachers who implement a child training curriculum universally in the classroom. Most evaluations have been based on parent- or family-level random assignment and have focused on the parent and child components. An evaluation of the teacher-training component within Head Start classrooms was combined with the parent-training component, so that the independent contribution of the teacher-training component could not be measured. The evaluations found favorable impacts on the behavior of parents, teachers, and children.
- ***Partnership-Directed, School-Based Approach to Child Physical Abuse and Neglect*** (Fantuzzo et al. 1996). The "Play Buddy" intervention focuses on socially withdrawn victims of child physical abuse or neglect. "Play buddies," who are other children in the classroom with exceptionally positive peer interaction skills, are coached by the teacher and parents to initiate and engage in positive interactions with socially withdrawn peers.
- ***Social-Skills Curricula for Head Start*** (Conduct Problems Prevention Research Group 1999; Domitrovich and Greenberg 2001; Frey et al. 2000; Serna et al. 2000). While still in the planning and development stages, several

classroom-wide curricula are being developed to promote positive social behavior, encourage impulse control, and improve children's ability to communicate about their feelings.

- ***Starting Early, Starting Smart*** (Springer et al. 2003; Karoly et al. 2001). This intervention integrates behavioral health services in either primary health care settings or early childhood settings to promote access to family/parenting education, mental health services; and substance abuse services; improve parenting skills and family well-being, and strengthen child development. In early childhood settings, the initiative typically involves placing a mental health consultant in the settings, although the primary strategy involves intensive work with families. The initiative is being evaluated in 12 sites based on random assignment of families to an intervention and a control group.

For a few of the strategies described here, rigorous designs have been used to conduct the evaluations; for most of them, however, evaluations using rigorous designs and sufficient sample to detect moderate levels of impact have not been completed. Some of the enhancements have only recently been adapted for and implemented in Head Start programs. Others have rigorous evaluations in progress, but the evaluation results and critical features of the sampling plan and research design are not yet available. Most of the interventions described above could benefit from further Stage 1 development activities.

Nevertheless, we discuss evaluation designs for two examples—the First Step to Success and the Incredible Years—because they introduce a design issue not present in the curriculum example discussed in Chapter IV. This design issue is the dual focus of these enhancements on classroom-management strategies (which can be evaluated using an approach similar to that discussed in Chapter IV) and on intervention with individual children and their parents (which requires a design that identifies and follows individual children and their parents). The other enhancements that are directed toward all children in the classroom can be evaluated using an approach similar to that discussed in Chapter IV, while those with an individual-child focus can include an evaluation component that examines a subset of children likely to receive those services, as we discuss in this chapter.

B. STUDY DESIGN

The fundamental issues to be addressed by an evaluation of a social-emotional behavioral intervention are whether the classroom environment is managed so that behavioral issues are minimized; whether children's behavior is less disruptive and more socially competent; and whether children in the classroom make more progress in language development, early literacy skills, and early mathematics relative to children in classrooms that are not implementing the interventions. For the intervention focusing on individual children receiving intensive services and their families, the research questions include addressing whether parenting practices and the home environment are more positive and minimize negative behaviors, and whether the behavior and academic progress of the children identified for the intervention are improved. This section discusses key elements of research designs to address those questions. We begin by discussing the quality

enhancement and its counterfactual, the research questions, sampling strategies, random assignment plans, and sample sizes.

1. The Quality Enhancement and Its Contrast

Implementing positive classroom management strategies and providing special services for children from chaotic homes who are acting out in class are not new ideas for Head Start. Teachers with an early childhood education background are likely to have heard about many of the strategies for positive management of children's behavior. In addition, the requirement that Head Start programs forge links with community-based health services may lead to cooperative agreements that provide staff with access to mental health professionals who can assist children needing special services. In the 2002-2003 program year, eight percent of Head Start families accessed mental health services, six percent received child abuse and neglect services, four percent received domestic violence services, and four percent received services to prevent or treat substance abuse (Hart and Schumacher 2004). Two percent of Head Start children were referred for mental health services in the same program year. Many programs have access to mental health professionals on site; these professionals can provide consultation to Head Start staff about individual children and can meet with the staff, children, and parents about specific issues.

However, despite the access to mental health professionals for consultation, and services to address mental health and family crisis needs provided to many Head Start families and children, many Head Start teachers still cite children's behavior as a significant issue that takes time away from language, early literacy, and other activities that are expected to occur in the classroom (Paulsell et al. 2004). Several reasons may explain why staff continue to perceive a strong need for behavior management strategies. First, although most teachers may be aware of some of the principles of positive management of children's behavior, they may also have difficulty implementing these principles in the classroom without "hands-on" assistance. Moreover, a new classroom of children can present one or more new behavioral issues, ranging from hyperactivity, to oppositional defiant disorder, to aggression, and to withdrawal, which cannot be resolved quickly. Teachers must have strategies for carrying on classroom activities over time while individual behavioral issues are being addressed. Finally, some programs do not have access to mental health professionals or services, and the ones that do have access might not receive as much help as they feel they need to address behavioral issues in their classrooms. Thus, the counterfactual for an evaluation of an enhancement addressing children's social-emotional well-being would probably be highly variable classroom management practices and a smaller proportion of individual children receiving more-intensive services to address more-serious behavioral issues.

The enhancements proposed for Stage 2 evaluation have two components that are basic to both enhancements: (1) management of the classroom to minimize opportunities for disruptive behavior, and (2) strategies for addressing children's disruptive behavior when it occurs. To implement the enhancements, teachers must be trained and given technical assistance so that the principles are translated into the teachers' arrangement of the classroom, scheduling of activities, and behavior in the classroom. Because the extension of

these enhancements to also provide specialized, intensive services to individual children who exhibit more-serious problems adds complexity to the evaluation design, we will discuss this extension as another optional design. The enhancements and counterfactuals to be discussed include the following:

- The first design option will compare the basic enhancement (teacher training and classroom management) with regular Head Start services. As we discuss below, centers participating in the evaluation will be randomly assigned to implement the enhancement or not, and the analysis will compare outcomes based on a sample of children in the centers.
- The second design option will compare the basic classroom enhancement plus individual services for children needing specialized social-emotional support (1) with the basic classroom enhancement, and (2) with regular Head Start services. Centers participating in the evaluation are randomly assigned to implement the basic classroom enhancement, the classroom enhancement plus individual services, or basic Head Start services. Children are then randomly assigned to the different types of centers, so that those with the greatest behavioral needs are not placed predominantly in the enhancement centers. The research will focus on a random sample of children in the classroom. This design will provide an estimate of the impact of the full package of services compared with regular Head Start services, the impact of the basic classroom enhancement compared with regular Head Start services, and the "value added" by the individual-child services piece.
- As a third design option, children likely to be referred for intensive services could serve as the specific focus of additional sampling. Using a definition of children with very high levels of behavior problems (based on maternal report using a behavior problems scale at baseline), approximately 10 to 20 percent of the children in each center could be included in the sample. This group would overlap somewhat with a sample of children chosen at random from the center as a whole. By measuring outcomes for this subgroup of children, researchers would be able to measure the impact of the two enhancement options compared with regular Head Start services on children with higher levels of behavioral problems at enrollment, a group that is likely to include most of the children referred for intensive services.

To clarify subsequent discussion of the three plans in this chapter, key features of the three designs are summarized in Table V.1. We discuss the third design as a stand-alone option in order to focus on its unique details; however, because policymakers interested in these questions also likely would be interested in the more general questions of the impacts on all children in the classroom, the samples described in Options 2 and 3 would be combined. The remainder of this chapter discusses each design feature in more detail and provides the rationales for the decisions summarized in the table.

Table V.1. Essential Features of the Three Optional Research Designs for Evaluating Enhancements to Support Children's Social-Emotional Well-Being

Option 1	Option 2	Option 3
Intervention		
Teacher training to improve classroom management and teacher-child interactions to promote positive child behavior and minimize disruptive behavior	Teacher training to improve classroom management and teacher-child interactions to promote positive child behavior and minimize disruptive behavior Child-level intensive intervention to address significant behavioral issues	Teacher training to improve classroom management and teacher-child interactions to promote positive child behavior and minimize disruptive behavior Child-level intensive intervention to address significant behavioral issues
Random Assignment		
Centers One enhancement One control	Centers One enhancement One control	Centers One enhancement One control
Sampling		
15 children per center	15 children per center	5 or 10 children per center identified by baseline parent report
Data Collection		
Consent form with demographic information Direct child assessment Teacher self-administered questionnaire Classroom Observation	Consent form with demographic information Direct child assessment Teacher self-administered questionnaire Classroom observation	Consent form with demographic information Direct child assessment Teacher self-administered questionnaire Classroom observation Parent report at baseline Parent interviews Child observations

SAQ = self-administered questionnaire.

2. Research Questions

The first set of research questions about the basic classroom-level enhancement (without the intensive intervention for specific children) focuses on direct impacts of the enhancements on classroom activities and child outcomes. The first evaluation option would address the following questions:

- Do the classroom management and teacher-child interaction strategies reduce classroom disruptions and reduce time spent correcting children's behavior?
- Do the classroom management and teacher-child interaction strategies increase the amount of classroom time spent on language development, early literacy, and early mathematics activities?

- Do the classroom management and teacher-child interaction strategies improve children's cooperative and prosocial behaviors and reduce behavioral problems?
- Do the classroom management and teacher-child interaction strategies lead to improved language, early literacy, and early mathematics achievement for children in the classroom?

An evaluation of the classroom-level enhancement with the intensive intervention for specific children (the second evaluation option) would address the following questions:

- Do the classroom strategies with individual-child intervention reduce classroom disruptions and reduce time spent correcting children's behavior relative to regular Head Start services? Compared with the classroom strategy alone, does this more intensive option reduce classroom disruptions and reduce time spent correcting children's behavior?
- Do the classroom strategies with individual-child intervention increase the amount of classroom time spent on language development, early literacy, and early mathematics activities relative to regular Head Start services? What is the value added of the more intensive approach compared with the classroom-level approach by itself?
- Do the classroom strategies with individual-child intervention improve children's cooperative and prosocial behaviors and reduce behavioral problems relative to regular Head Start services? What is the value added of the more intensive approach compared with the classroom-level approach by itself?
- Do the classroom strategies with individual-child intervention lead to improved language, early literacy, and early mathematics achievement for children in the classroom relative to regular Head Start services? What is the value added of the more intensive approach compared with the classroom-level approach by itself?

By Stage 2, we expect that implementation is likely to be successful because enhancement developers will have honed their techniques for implementing the intervention in Head Start settings. Nevertheless, understanding the impacts of the enhancement on children's development and on classroom activities hinges on whether or not the training and technical assistance process led teachers to implement the enhancement with high fidelity. The larger number and diversity of centers and classrooms involved in the Stage 2 study might lead to new challenges, either for the "fit" between the enhancement and the program or for the training and technical assistance procedures. The evaluation should therefore address the following questions to examine whether the enhancement was implemented to high fidelity, what steps were taken to reach that point, and what was learned from the process:

- Were teachers able to implement the classroom management techniques and teacher-child interaction strategies fully and with a high degree of fidelity? For the second, more intensive child-focused option, were mental health professionals able to respond to teachers' requests for assistance with particular

children, and did these children receive a level of services to meet their social-emotional needs?

- What strategies were used to implement the enhancement? How much training was provided, and over what time period? What amount and types of technical assistance were provided? How much time did the mental health professional spend on site, and how many children and families could be served?
- What challenges were encountered, either for the "fit" between the classroom management techniques and teacher-child interaction strategies and the program or for training and technical assistance procedures, and how were they resolved? What challenges were encountered in providing individual-child services?

Finally, although the evaluation will examine whether the behavioral enhancement is effective overall, the Head Start community also will be interested in whether it is effective across different subgroups of children and families, and across different program designs. To examine effectiveness in this way, the following questions would be addressed:

- What were the impacts of the classroom behavior enhancement on key subgroups of children; for example, children with lower verbal skills at baseline? What were the impacts for Head Start programs with differing characteristics; for example, programs in which teachers have higher levels of education?

Some caution should be exercised in conducting subgroup analyses in Stage 2; if many subgroups are examined, some impacts will emerge simply by chance. Moreover, because the samples of grantees and centers are not representative of the broader Head Start population, the subgroups are similarly not representative. Drawing conclusions about how well the curriculum works in specific subgroups would require that the sample frame be designed to include a representative sample from those subgroups. However, this approach would not be consistent with the Stage 2 design, which is to conduct an evaluation that yields internally consistent impact estimates for a group of program grantees and centers that volunteer to participate in the study. Impact estimates for subgroups in a Stage 2 design are valid for the subgroups attending the centers included in the sample. Obtaining a representative sample of specified subgroups would increase the sample size requirements of the study. Thus, serious investigation of subgroup impacts must wait until a Stage 3 evaluation, which will examine a sample that is representative of Head Start programs at the regional or national level.

One subgroup of obvious interest is the children who are identified as needing intensive behavioral services. To address research questions about the impact of the intervention on children who needed intensive services during the year, we have included the third design option that focuses sampling on children identified at baseline as having high levels of externalizing behavioral problems. In the absence of such a strategy, measuring impacts for the subgroup of children in the sample who received intensive behavioral services during the Head Start year would be challenging; to do so, the researcher would have to identify the children in the control group who would have received intensive services had they been in the intensive enhancement group. Because the sample of children for Options 1 and 2 is

randomly drawn from all children in the center, it will not necessarily include the children receiving intensive services, as those children are expected to comprise between 10 percent and 20 percent of the population of children in the center. Option 3 is designed to address questions about that subgroup.

3. Major Activities and Timetable for the Evaluation

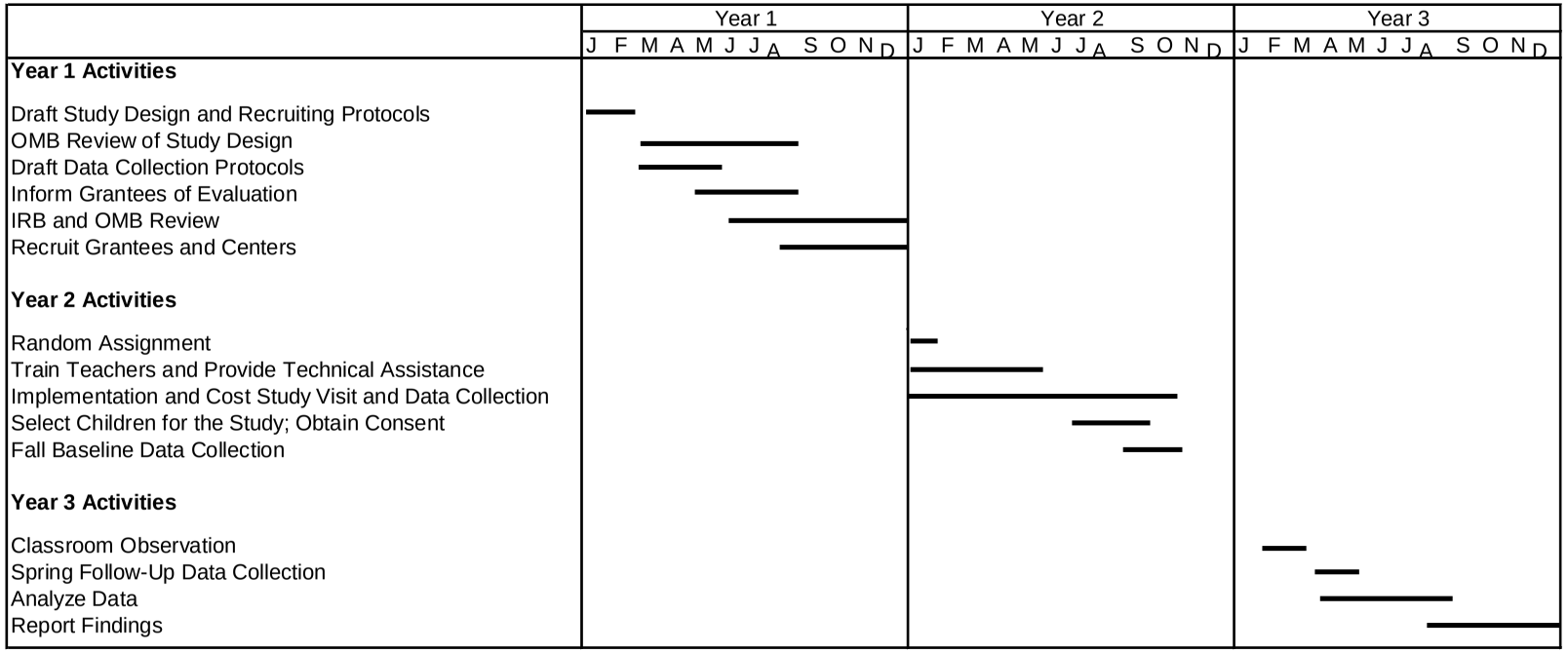
Several activities must be performed to carry out the evaluation:

- Draft study design and protocol for recruiting program grantees and centers and submit for Office of Management and Budget (OMB) review
- Identify programs and centers to participate
- Develop data collection instruments and prepare and submit review packages for the Institutional Review Board (IRB) and OMB
- Conduct and monitor random assignment
- Train teachers to implement the enhancement in randomly selected centers; provide technical assistance and additional services and supplies needed for implementation; monitor fidelity to implementation
- Identify a sample of children who have been given consent to participate
- Collect data from the enhancement and control groups
- Analyze the data and report findings

We recommend that the first four tasks occur during the first year of the evaluation, with implementation, sampling of children, and collection of data during the second year, and additional data collection, analysis, and reporting during the third year (see Figure V.1). Because the Head Start year typically runs from August or September through May or June, the timing of activities will proceed most smoothly if the evaluation activities begin in January or February. We discuss the steps in more detail in the rest of this chapter.

The first year of evaluation activities will be dominated by OMB review and by recruitment of grantees and centers. OMB review of the study design and protocols for recruiting grantees and centers must be completed before researchers and curriculum developers can obtain information about prospective centers or negotiate agreements to participate. We have estimated two months to draft and submit a package to OMB that includes the study design and recruiting protocols, and six months for OMB review. During the OMB review period, researchers would be able to conduct activities that do not involve information-gathering from prospective centers. For example, data collection protocols could be developed and submitted for OMB review, and the study design and data collection plans could be submitted for IRB review. Data systems for tracking children and managing evaluation data could be developed. The Administration for Children and Families (ACF) could inform the Head Start community about the pending evaluation and encourage

Figure V.1. Schedule of Activities for Small-Scale Evaluation of a Classroom Behavioral Enhancement Three-Year Study Beginning in January



IRB = Institutional Review Board; OMB = Office of Management and Budget.

participation. After receiving OMB clearance, the curriculum developer and researchers would be able to contact interested grantees and centers, and to begin the recruiting process. We estimate that recruitment, executing agreements, and randomly assigning centers will require approximately four months.

The main activities in the second evaluation year (beginning in January or February) will consist of training teachers to implement the classroom management techniques and teacher-child interaction strategies to high fidelity, and conducting the implementation study. In the fall, the main tasks will include ensuring children are placed in centers without regard to behavioral issues, obtaining consent for children to participate in the study, sampling children, and collecting baseline data. Ideally, teacher training and implementation will occur in the spring so that, when the new Head Start year begins, children enter a classroom that is managed and organized according to the techniques for maximizing positive behavior and minimizing disruptive behavior. Third-year evaluation activities will include collecting follow-up data in the spring on classrooms and children, analyzing the data, and reporting the results.

4. Sampling, Random Assignment, and Sample Sizes

Stage 2 designs generally include grantees and centers that have agreed to participate in the study and were selected because of their location or interest rather than their ability to represent Head Start centers and grantees regionally or nationally. At this stage, the evaluation focuses mainly on whether the quality enhancement *could* be effective in a set of those grantees and centers. As we discuss in more detail in Appendix B,[1] because the emphasis is on internal validity, or a rigorous test of the curriculum within the set of grantees and centers in the evaluation, the sample size requirements are lower than if grantees and centers were selected to represent all Head Start grantees and centers. For Stage 2 designs, a set of program grantees and centers interested in participating must be identified, random assignment completed, and classrooms and children selected to participate in data collection.

Identification of Grantees and Centers. Various methods could be used to recruit grantees. Methods of recruiting a broad set of programs include working through the Head Start Bureau to communicate with all regional offices, sending faxes to all programs, and sending emails to program directors. However, Stage 2 evaluations ideally should be geographically focused to support careful implementation of the enhancement, so more-targeted methods probably would be more effective than would broadly disseminated flyers or similar methods. With support from the Head Start Bureau and selected regional offices, the enhancement developer and research team also could approach directors of Head Start programs within a geographic area and could ask colleagues in other areas who might help with implementation to contact program directors in their own areas. If a broader call for participation is made first, interested programs could be asked to help the enhancement developer to approach other programs in their areas. The initial contact materials should

[1] If impact estimates were expected to be externally valid (representative of all Head Start centers), then sample sizes would have to be larger; see Appendix B for details.

briefly explain the study's focus (in this case, an approach to classroom management and addressing children's behavioral problems) and should briefly summarize the benefits of participation. The contact materials also might indicate that all centers in the program will have an equal chance of participating in the study, but that only some of the ones in each program will be chosen to implement the enhancement. Those not chosen to implement the enhancement in the first year will be given priority to implement it, if desired, after the follow-up data have been collected. Programs will receive information about the study's findings, and will be partners in the research study. We expect that several programs will want to learn more about the evaluation as a result of this recruiting effort.

Ideally, program grantees and centers chosen for the study would have the following characteristics:

- Centers are clustered in a small number of geographic locations to simplify implementation.
- Centers are not already implementing one of the social-emotional enhancements described earlier in the chapter but are interested in doing so. For the intervention that includes intensive child-focused services, participating centers do not currently have access to intensive, on-site mental health services for children.
- Grantees and centers are willing to cooperate with the study's random assignment and data collection requirements.

Recruitment of grantees and centers will be easier if their staff believe they will gain more from participation than they might lose. Accordingly, after the initial contact has been made, the benefits to sites must be explained in detail, and concerns about study burden must be discussed. Enhancement developers and research staff should visit program directors and other relevant administrative staff to discuss the enhancement, its expected benefits to children, what will be involved in implementing it, and the research aspects of the evaluation.

Researchers will explain the program's role in ensuring that the study yields useful information about the effectiveness of the classroom-wide intervention and the intensive child-focused intervention. For example, researchers will have to work with program staff to obtain information required to implement random assignment (for example, the number of centers and classes in each center, teachers' names, class sizes, and children's ages, as well as procedures followed for application to the program and placing children in centers). Program staff will have to maintain random assignment statuses (for example, by ensuring that teachers in the intervention group do not share information about the curriculum with control-group teachers). Researchers will have to monitor the integrity of random assignment over time, a key piece of information to demonstrate the reliability of the study.[2]

[2] If teachers in the enhancement group discuss the enhancement strategies with control-group teachers, the control group could implement a version of the enhancement, and the evaluation would not provide a valid

One important aspect of monitoring random assignment for this enhancement is to ensure that children with behavioral problems are not disproportionately assigned to centers offering the enhancement. Researchers also will have to work with program staff to schedule child assessments and classroom observations. These evaluation-related requirements are balanced by the chance to implement a classroom management and behavioral approach that could benefit children, and by the chance to work with the enhancement developer to ensure that the approach fits the needs and interests of Head Start classrooms, families, and children, and that it can be incorporated easily into Head Start program activities. Benefits to the control group are more challenging to identify, but an important benefit that could be offered for control-group participants in a Stage 2 evaluation is first priority to implement the enhancement after the children participating in the evaluation have finished their Head Start year.

A program's agreement to participate should include a memorandum of understanding that describes the benefits to the program, and that specifies the respective responsibilities of the enhancement developer, researchers, and program in this joint research undertaking. A similar agreement should be developed with each participating center. This level of detail ensures that misunderstandings that have the potential to threaten the success of the research study do not arise after it is too late to resolve them. A memorandum of understanding also offers a useful way of informing new staff about the study.

Other Program, Center, and Child Sample Selection Criteria. Because group-randomized designs require relatively large sample sizes to detect impacts of moderate size (20 percent of the standard deviation of the outcome measure), it can be costly to expand the sample to provide sufficient power to detect similar levels of impact in subgroups. For this reason, the programs, centers, and children to be included in the sample should be defined carefully, with the most important research questions in mind when doing so.

One sample-definition strategy is to include a broad set of programs with different characteristics. This approach would address the question of the enhancement's effectiveness across the range of Head Start programs and families. In Stage 2 evaluations, however, the programs will not have been sampled randomly from all Head Start programs, so the sample will not truly reflect the diversity of programs. If the evaluation finds that the enhancement is not effective, it will be difficult to determine whether it could be effective in some program subgroups.

A more useful strategy is to target programs with a narrower set of characteristics in common, so that the evaluation measures the effects of the enhancement in a more defined set of circumstances. Nevertheless, for the classroom management and child behavioral

(*continued*)
test of the impacts of the enhancement. Impact estimates would be biased downward, making it more difficult to conclude that the enhancement was effective. To avoid such contamination of the control group, we recommend implementing enhancements at the classroom level only if the potential for spillover from enhancement to control-group classrooms is small.

enhancement, diversity in some key areas could be useful. About half the children served by Head Start attend half-day programs; therefore, we recommend that the group of participating centers include an approximately even split between full-day and half-day programs. Children's age is another characteristic that varies across (and within) Head Start classrooms. We recommend including both three- and four-year-old children in the evaluation, as the enhancement is designed for both age groups, and most behavioral measures can be used and compared across both groups.

The first and second design options can be implemented using a random sample of three classrooms per center and eight children per classroom.[3] This sample would address classroom-level research questions (for example, whether the average level of behavioral problems is lower; positive social behavior increases; and children are progressing further in language, early literacy, and early mathematics than in the absence of the enhancement). Although not all of the classrooms and children in the center would be part of the research sample, the enhancement would be implemented in every classroom in a center assigned to the enhancement group.

For the third design option, which focuses on the well-being of the children who receive intensive behavioral services, researchers will have to identify (in both the enhancement and control groups) children who are likely to need intensive services to address behavioral issues (as this group is likely to be only about 10 percent of all children in the center). To do so, we recommend that researchers obtain parent reports about children's behavior using a psychometrically strong scale measuring children's behavioral problems. The parent report would be completed early in the Head Start year (as part of the consent process or at program application). The subgroup of children for the research sample in the third option will then be identified based on higher-than-average levels of parent-reported externalizing behavioral problems (including aggression and hyperactivity).

We recommend using parent reports so that the group of children can be identified at baseline, before they have experienced different classroom environments. Using teacher reports instead could delay identification of children, as it might take teachers as much as two months to distinguish children who are acting out because they are unfamiliar with the Head Start classroom from children who need intensive services. Moreover, the enhancement classrooms and regular Head Start classrooms might lead the same child to behave differently, thus influencing the selection of this subgroup. Thus, although parent reports will not enable researchers to perfectly identify children who would be referred by teachers during the year for intensive behavioral services, the overlap should be substantial, and this strategy sidesteps the issues that might arise when asking teachers to select this group.

Because teacher reports suggest that approximately 10 percent of Head Start children exhibit disruptive behavior, two or three children per class, or four to six children per center,

[3] Assuming a 90 percent response rate to the spring follow-up child assessment, nine children per class would have to be sampled at the beginning of the year.

would likely be identified for services (Kupersmidt et al. 2000). If referral services are available, teachers might approach a professional about more children than they would identify as having significant behavioral issues in the absence of services. Thus, our strategy for including children in this subsample should be somewhat more inclusive to ensure that nearly all children receiving intensive services are included. We recommend including in the sample five children per classroom (approximately 25 percent of all children in the class) who have the highest incidence of parent-reported externalizing behavioral problems. With a 90 percent response rate, we assume that four children per classroom will be available for follow-up assessment in the spring.

The evaluation could combine the sampling strategies of Options 2 and 3 to include this sample of children at high risk for needing intensive behavioral services, as well as a random sample of all children in the center. The two samples are likely to have little overlap. Combining the sampling approaches would enable the evaluation to address both research questions focusing on children likely to need the intensive services and the broader questions of whether the behavioral interventions lead to a greater focus on language, literacy, and mathematics in the classroom, and whether corresponding gains for children occurred in those areas.

Random Assignment. We recommend center-level random assignment for the classroom management and child behavioral interventions discussed in this chapter because the likelihood of spillover, or nonrandom allocation, of children to classrooms is too great if classrooms are randomly assigned. Children's behavioral problems are among the greatest challenges for preschool teachers, so it is highly likely that teachers assigned to the enhancement group would give good advice to teachers in the control group who are struggling with difficult behavioral issues. Moreover, within a center, directors allocate children based on several factors, including gender balance, classroom schedule, and the fit between a child's temperament and the teacher's strengths. Directors who are aware of a child's behavioral problems would be strongly tempted to place the child in one of the classrooms receiving enhanced behavioral services, which would threaten the validity of the study. We expect that strategic assignment of children is less likely to be an issue if centers were randomly assigned, as decisions about children's placement in centers are generally based on such factors as center locations and schedules relative to the family's residence and schedule.

If centers are randomly assigned, the sample size requirements (and thus, evaluation costs) will be higher than if classrooms were randomly assigned, as randomly assigning larger units generally decreases the power of the design (see Appendix B). Moreover, we give up two other advantages of classroom-level random assignment if centers are randomly assigned instead. First, classrooms within one center are more likely to be similar in population served and classroom management strategies than are two centers. Random assignment at the classroom level also might offer a slight advantage over center-level random assignment at the recruiting stage. Center directors might prefer the 100 percent chance that at least some classrooms in their centers could implement the enhancement under classroom-level random assignment over the 50 percent chance that no classrooms would implement the enhancement under center-level random assignment. Furthermore,

having both enhancement classes and control classes within their centers enables directors to gain a sense (although not rigorous evidence) of how the new approach seems to be working.

Nevertheless, the potential for spillover and strategic assignment of children to classrooms outweighs the potential benefits of classroom-level random assignment. In addition, center-level random assignment has two other potential advantages. First, teachers in the enhancement group within a particular center would be able to discuss the new techniques without fear of revealing ideas to control-group teachers. These discussions could enhance the overall success of implementation. Moreover, program developers often view full-center implementation as more realistic than an implementation in which enhancement and control classrooms are in the same center because they have the ability to work with everyone in the center and staff can talk freely about their implementation experiences. Second, classroom configurations and teachers change from year to year—and even from May to August—as enrollments are finalized, available space in the preschool is negotiated, and teachers' plans become settled. These changes would present fewer difficulties for the evaluation if an entire center were to either follow the enhancement or continue its usual practices than if individual teachers or classrooms were randomized to one group or the other.

Random assignment should take place during the year before outcome data are collected, so that enhancement developers can fully implement the classroom management and teacher-child interaction strategies and can ensure fidelity.[4] Under this schedule, if initial implementation begins in March or April of the program year, teachers will have time to learn and practice the new techniques before the next school year begins.

If centers are to be randomly assigned, participating programs will have to give the researchers a list of participating centers, so that random assignment can begin. Within a single grantee, centers would be randomly assigned to the control or treatment group(s). For Option 1, an even number of centers is needed (half for the enhancement group and half for the control group). For Options 2 and 3, the number of centers should accommodate assignment of one-third of the centers to the control group, and one-third to each of the two enhancement groups.

Selection of Classrooms and Children into the Research Sample After Random Assignment. Although the evaluation design may call for random assignment of centers, it will not be necessary to collect data from every classroom and every child in the center. Including three classrooms per center and eight children per classroom will provide a sample large enough to estimate average outcomes under random assignment for centers. Beyond that number, additional children add data collection costs at a constant rate but add less to the power of the design to detect impacts. As mentioned, the design for Option 3 would

[4] We discuss implementation further in Section C of this chapter.

focus sampling on approximately four children per classroom who are most likely to be referred for intensive, individual services to address behavioral issues.[5]

An alternative sampling approach that increases the power of the design is to sample children from the center without regard to classroom (the approach described in Chapter IV). This approach provides a sufficient sample of children to measure child outcomes, but the sample of classrooms (used for teacher- and classroom-level analyses) would not strictly be representative of the center's classrooms. Instead of randomly selecting classrooms for the classroom-level observations, researchers could observe the three classrooms per center that include most of the children selected for the child sample. The advantage of this approach is that the sample size of children needed to detect the same levels of impact at the same level of power is approximately 25 percent lower and the number of centers required is approximately 10 percent lower than under the previously discussed design, which would randomly select classrooms and then children within classrooms.

Sample Sizes. Table V.2 shows the number of centers, classrooms, and children required for the evaluation of a child behavioral enhancement under alternative designs. Contrasting two different enhancements to regular Head Start services or to each other requires a larger sample than does a simple comparison of one enhancement with regular Head Start services. Focusing the sample on children with higher levels of parent-reported behavioral problems reduces the number of children per center, requiring a slight increase in the number of participating centers but an overall decrease in the number of children if assumed levels of precision (15 percent to 20 percent of a standard deviation of the outcome measures) are to be preserved.

Under the first design option, which contrasts the classroom-level enhancement with regular Head Start services, if the desired level of precision in detecting impacts were .20 (one-fifth of a standard deviation on the outcome measures), researchers would randomize 92 centers (46 to each group) and would then randomly select 1,656 children from those centers, assuming the availability of only minimal baseline data.[6] Minimal baseline data would include demographic information about the family collected as part of the consent form and children's scores from the fall Head Start National Reporting System (NRS) assessments.[7] Baseline demographic data allows us to control for these characteristics statistically when we estimate impacts, which reduces the variance of the impact estimate and thereby increases the power of a particular sample size to detect impacts. (Section E describes the statistical models that could be used in the analysis of impacts.) Sample sizes

[5] The plan for selecting classrooms and children after random assignment of centers is discussed in Section D.

[6] We assume that the evaluation will include two classrooms per center and nine children per classroom in the initial sample, with a 90 percent response rate to the spring follow-up data collection.

[7] Currently, NRS assessment scores are available only to individual programs aggregated to the program level. Making the scores available for evaluation in the way that we describe would require approval from ACF and consent from the families.

Table V.2. Sample Sizes Required for a Stage 2 Evaluation of a Classroom Management and Child Behavioral Intervention Under Alternative Designs and Random Assignment of Centers

	Total Number of Centers	Total Number of Classes	Initial Sample of Children
Option 1: Classroom-Level Intervention Only; Sample from All Children			
MDE = 0.15			
No baseline	204	408	3,672
Minimal baseline	164	328	2,952
MDE = 0.20			
No baseline	116	232	2,088
Minimal baseline	92	184	1,656
Option 2: Classroom-Level and Individual-Child Interventions; Sample from All Children			
MDE = 0.15			
No baseline	306	612	5,508
Minimal baseline	246	492	4,428
MDE = 0.20			
No baseline	174	348	3,132
Minimal baseline	138	276	2,484
Option 3: Classroom-Level and Individual-Child Interventions; Sample of Children with Behavioral Problems			
MDE = 0.15			
No baseline	420	840	4,200
Minimal baseline	336	672	3,360
MDE = 0.20			
No baseline	237	474	2,370
Minimal baseline	189	378	1,890

Note: Sample size calculations assume that two classrooms per center are randomly selected for the research sample. We also assume that the research sample for options 1 and 2 initially includes 9 children per classroom and 90 percent of the initial sample is available at followup. We assume that the research sample for option 3 initially includes 5 children per classroom and that 4 children per classroom are available at followup. "Minimal baseline" means that demographic information is collected as part of the consent form, and that NRS fall outcome data are available so that the R^2 for the regression-adjustment of the impact estimates is .20.

Sample size calculations also assume a two-tailed test of statistical significance with 80 percent power and a 95 percent confidence level. The sample size calculations do not include an adjustment for the design effect of weighting for sample nonresponse. Appendix B provides details about the calculations.

MDE = minimum detectable effect; NRS = National Reporting System.

would have to be substantially larger if the desired precision level for impact estimates were .15, as opposed to .20.

Under the second design option, which includes a second enhancement group that adds to the basic classroom enhancement the offer of intensive services to children with more-serious social-emotional issues, the sample of centers and children expands by one-third. If the desired level of precision in detecting impacts were .20, researchers would randomize

138 centers (46 to each of three groups) and would randomly select 2,484 children from those centers if minimal baseline data were available.

Under the third design option, which examines the same two enhancements and control group as under the second design option but focuses on a sample of children identified by parent reports at baseline as having higher-than-average behavioral problems, the number of centers is higher, but the number of children required for the sample is lower. We expect that only five children per classroom will be identified as having high levels of behavioral problems so have assumed four children per classroom in the final analysis sample (after nonresponse) under this design, rather than eight. The clustering of the sample means that including more centers adds more to the power of the design than is lost by reducing the number of children per classroom and per center. Thus, if the desired level of precision in detecting impacts were .20, researchers would randomize 189 centers (63 to each of three groups) and would randomly select 1,890 children from those centers if minimal baseline data were available.

Alternatively, researchers could combine Options 2 and 3 while keeping the number of centers at Option 2 levels and accepting the loss of power to detect impacts in the subsample of children with high levels of behavioral problems. Thus, if an Option 2 study were conducted using 138 centers, two classrooms per center, and eight children per classroom, which can detect effect sizes of at least .20 if minimal baseline information is available, the subsample of children with behavioral problems would be sufficient to detect effects greater than .20 but less than .25, which seems adequate.

C. IMPLEMENTING THE CLASSROOM AND CHILD-FOCUSED INTERVENTIONS

The goals of implementation in Stage 2 are twofold. First, the enhancement should be implemented to high fidelity in the classrooms within the centers that were assigned to the enhancement group. High fidelity ensures that the evaluation measures the impact of the enhancement as designed, rather than measures a watered-down version that was implemented incorrectly. Second, procedures and manuals should be clear enough and should cover a broad enough range of Head Start program situations so that, in the future, it would be feasible to implement the enhancement on a broader scale, using Head Start training and technical assistance staff, with only minimal assistance from the enhancement developer. To achieve this goal, the procedures and manuals will have to be revised as necessary after implementation on the basis of information obtained while implementing the enhancement in the centers. Issues that arise during initial training or during the technical assistance period will have to be addressed and documented so that the training materials subsequently can be revised in a way that reflects what has occurred in the classroom during implementation.

At the beginning of Stage 2, the plans are available for training teachers on principles of classroom arrangement, scheduling activities, and interacting with children to promote positive behavior and minimize opportunities for disruptive behavior, as these plans will have been established during the enhancement development period (Stage 1). At the start of

implementation, a clear plan to implement the classroom-based aspects of the enhancement to high fidelity will therefore already be in place. Manuals will be available that describe room arrangements in detail; give teachers examples of activities; provide numerous examples of how to deflect or defuse conflict situations, and how to identify and acknowledge positive behavior; and offer trainers a plan to work with teachers during an intensive initial period and over time, while the teachers are trying the new methods in their classrooms. The manuals should indicate the intensity of initial training and ongoing technical assistance, the duration of the training and technical assistance, and the training and technical assistance staff's qualifications. These manuals do not guarantee that the implementation will proceed smoothly in all centers, but they do provide a basic plan that will have been refined in several Head Start centers. Challenges encountered and resolved during implementation of a Stage 2 evaluation will form the basis for modifying the enhancement or the training and technical assistance.

For a small-scale evaluation (consisting of 50 to 150 centers, depending on whether one or two versions of the enhancement are tested, whether the sample is drawn from all children in a center or focuses on children with high levels of behavioral problems, and whether minimal baseline data have been collected), we expect that the enhancement developer will coordinate training for the teachers in centers assigned to the treatment group. Depending on the design, the training and technical assistance staff will work with teachers in either one-half or two-thirds of the centers. An evaluation of this size could operate in approximately five to ten geographic locations, each with 5 to 15 enhancement centers and an appropriate number of control-group centers. The enhancement developer should have a site coordinator in each location and a staff of trainers to offer initial training and on-site technical assistance during both the implementation period (March through May of the second year) and the evaluation year (September through December of the second year and January through June of the third year). The enhancement developer should visit the sites periodically to conduct some of the large-group training, address any issues that arose during training and technical assistance, and monitor implementation.

The intensive child-focused intervention will require that a mental health professional be available to each center assigned to the treatment group that combines the classroom-level and individual-child services. The mental health professional will consult with teachers about individual children whose behavior raises concerns and, as appropriate, will observe the children in the classroom, provide strategies to the teachers for addressing behavioral challenges, and meet with parents to discuss issues in the home that may be contributing to the observed behavior. The mental health professional also will provide intensive, one-on-one services for approximately two months to children who can benefit from learning and practicing skills in communicating feelings, empathy, anger management, conflict resolution, and developing friendships.

Prior to implementation of the classroom-based social-emotional enhancement, the enhancement developers should communicate with the leadership and stakeholders of the participating programs at all the levels involved, including the program director, education coordinators, center directors, parent Policy Councils members, and others, as appropriate. Care must be taken to ascertain the most efficient selection of staff to train. For example,

training all members of a teaching team, including teaching assistants and aides, as well as the lead teacher, should help to ensure optimal implementation of the classroom management ideas and principles for working with children, as well as the continuation of the approach in the event of teacher turnover. All relevant staff should at least be informed about the goals and approaches of the behavioral enhancement.

Implementing a classroom management enhancement is likely to require two or three days of initial training in a specialized setting (such as a school or university that has facilities for both large and small groups). Training should include an overview of the classroom management approach and video footage of teachers practicing the skills involved in this approach to classroom management in classrooms of preschool children, so that teachers in the enhancement group can visualize what the approach looks like in practice. Training should then cover a series of modules that describe each strategy. Practical exercises involving role playing and hands-on experience with the major strategies should be offered so that the teachers can practice what they have learned. Techniques for direct and observational assessments of children and links between behaviors that these techniques are promoting and required elements of the Head Start Child Outcomes framework also can be covered in training sessions. The training should wrap up with summaries, questions and answers, and discussion of the technical assistance plan.

After training has been completed, technical assistance staff will visit classrooms periodically to observe how teachers are practicing the techniques, and to discuss any questions or difficult behavioral issues that teachers are facing. The staff should visit once per week at first, for about one to two hours, and could then taper the visit schedule to every other week and, eventually, to once per month. During the visits, the staff should discuss with the teachers any questions or issues faced while practicing the techniques. They should observe the classrooms; complete measures of fidelity; and discuss with the teachers what they have observed, what aspects of implementation are going well, and what steps the teachers can take to improve their techniques. In the fall, refresher training just before classes begin would likely be helpful to teachers. Trainers also should provide intensive training to new teachers who have joined enhancement centers, measure fidelity in all enhancement classrooms, and provide targeted technical assistance to teachers who need additional help with implementation.

1. Measuring Implementation

Researchers will have to monitor random assignment throughout the implementation period to ensure that the enhancement and control groups remain separate. During this time, they also will measure the process of implementation and fidelity of the ssocial-emotional enhancement. During the evaluation year, they will continue to monitor implementation and fidelity and, as discussed in the next section, will collect measures of teachers', classrooms', and children's outcomes for the impact analysis.

Measures of Implementation and Fidelity. Researchers will visit a subset of approximately 30 percent of the enhancement centers to measure implementation processes and fidelity to the enhancement.[8] The number of centers included in the implementation study is generally determined as a compromise between the evaluation budget and the need to measure implementation experiences in several of the centers. (A list of topics, by data collection mode, is shown in Table V.3.) Since the implementation study is intended to provide insights into what implementation strategies worked well and what did not, the implementation staff should help classify programs by their implementation experience (for example, how well and how easily implementation occurred) and centers can then be randomly selected from each of these groups to participate in the implementation study. Visits should take place during the spring of the implementation year (year 2), after teachers have received training and technical assistance, and again in the fall, approximately one month after classes begin. Researchers will conduct classroom observations to measure how closely the classroom practices match the enhancement's design, and how well the services of the mental health professional meet teachers' and children's needs. Researchers will conduct semistructured interviews with training and technical assistance staff and with key program staff, including directors, education coordinators, and health coordinators, as well as focus groups with teachers to explore the content and quality of training, the fit between the enhancement and the teachers' classroom experiences, and the content and quality of technical assistance. Questions will focus on what aspects of training and technical assistance worked well; what challenges were encountered; how well the approach fits with the program; and what aspects of the enhancement, training, and technical assistance might have to be changed.

Although enhancements that are evaluated at Stage 2 are expected to have well-developed implementation plans and thus a high degree of success in implementing to high fidelity in the participating centers, it is not possible to anticipate every challenge to implementation. Accordingly, measures of implementation and fidelity are required both to improve the next generation of implementation materials (as described above) and to ensure that the evaluation can measure the effects of a well-implemented enhancement (discussed in Section E). Measures of fidelity to the enhancement design and the quality of implementation obtained from classroom observations as part of the impact study will be used to define subgroups that enable researchers to estimate impacts of the classroom management approach among centers that implemented the enhancement with high fidelity. Because fidelity and the quality of implementation are not experimentally determined (relative to centers that do not implement to high fidelity), these estimates will have to be interpreted with caution, but they will provide a measure of the effectiveness of the enhancement approach in "high-fidelity" sites relative to the overall impact.

[8] Note that, as part of the impact evaluation, researchers will observe two classrooms per center to measure fidelity and classroom processes; these observations will be conducted in all enhancement and control centers. The implementation study will provide in-depth information about implementation experiences in a sample of enhancement centers.

Table V.3: Potential Implementation Study Topics for an Evaluation of Social-Emotional Behavioral Intervention

Implementation Study Topics	Data Collection Methods				
	Direct Observation	Program Records and Documents	Semi-Structured Interviews with Key Program Staff	Semi-Structured Interviews with T/TA Providers	Focus Groups with Teachers
Initial Staff Training					
What was the content of training?		X		X	X
What were the qualifications of trainers?				X	
What was the intensity and duration?				X	
How well did training prepare staff to implement the classroom management techniques?	X		X	X	X
How well did the approach support other classroom activities?	X		X	X	X
How could training be improved?			X	X	X
Enhancement					
Were the principles of classroom arrangement followed in classrooms observed?				X	X
Were classroom activities structured to minimize disruptive behavior?				X	X
Do teachers have more class time available for language, literacy, and mathematics activities than before?			X	X	X
Are teachers using positive behavioral strategies with children?	X		X	X	X
Are teachers using appropriate strategies to minimize negative behavior?			X	X	X
Technical Assistance and Support					
Who provided technical assistance and support? Enhancement development staff? Health coordinators? Others?			X	X	X
What types of support did staff receive in providing enhanced services? Observation and feedback? Modeling? Conferences between teachers and technical assistance staff? Written reports?			X	X	X
How often was this assistance provided?			X	X	X
What topics were covered?			X	X	X

Table V.3 *(continued)*

Implementation Study Topics	Data Collection Methods				
	Direct Observation	Program Records and Documents	Semi-Structured Interviews with Key Program Staff	Semi-Structured Interviews with T/TA Providers	Focus Groups with Teachers
What types of questions were raised? What support Was needed?			X	X	X
How helpful was the technical assistance?			X	X	X
What other types of support did staff need?			X	X	X
Lessons Learned					
Which aspects of training went well?			X	X	X
What factors associated with programs or teachers made implementation go well?			X	X	X
What challenges were encountered in training? In technical assistance?			X	X	X
What factors associated with programs or teachers made training or technical assistance more challenging?			X	X	X
What strategies were used to overcome the challenges?			X	X	X
How well has the classroom management approach supported other program activities?			X	X	X
How well has the approach seemed to support the needs of Head Start children and families?			X	X	X
How did teachers address any issues regarding the fit of the approach with the Head Start program?			X	X	X
What did Head Start families think of the classroom management approaches?			X	X	X
How well did the mental health professional work with teachers? What challenges, if any, were encountered in this partnership, and how were they resolved?					
How well did the mental health professional work with parents? What challenges, if any, were encountered and how were they resolved?					
How well did the mental health coordinator work with individual children? What challenges, if any, were encountered and how were they resolved?					
How can the enhancement be improved?			X	X	X
How can T/TA be improved?			X	X	X

T/TA = training and technical assistance.

D. OUTCOMES MEASUREMENT AND DATA COLLECTION PLANS

A comprehensive assessment of the impact of enhancements to support children's social-emotional development would examine the impact of the intervention on children, teachers, and the overall structure of classroom activities. If the enhancement is successfully implemented, it should enable the teacher to manage the classroom so that all children are actively engaged in language development, early literacy, early mathematics, and play, and so that the teacher does not have to spend disproportionate amounts of time with children who require additional support.

Outcome measures selected for the evaluation should have the following properties:

- ***Relevance to School Readiness Goals and the Head Start Child Outcomes Framework.*** The Head Start Child Outcomes Framework (COF) outlines eight domains of early learning and development (see Appendix A), including social and emotional development, approaches to learning, and the development of skills in language, early literacy, and early mathematics, all relevant to this enhancement.
- ***Sensitivity to Intervention.*** Measures should reflect outcomes malleable to intervention, as opposed to more trait-like qualities (such as temperament).
- ***Appropriateness for a Culturally Diverse, Low-Income Population.*** The selected measures should adequately assess school readiness of low-income 3- to 5-year-olds from diverse cultural populations, including those who do not speak English in the home.
- ***Adequate Psychometric Properties.*** Measures selected should be reliable and valid. Reliability means that they measure the same construct across various settings (for example, the classroom or the home), on repeated occasions (*test-retest reliability)*, when administered or rated by different interviewers/observers (*interrater reliability*), and when subsets of items are administered to identical samples (*split-half reliability*). Validity refers to the degree to which the measure taps the underlying construct it purports to measure, keeping in mind linguistic and cultural considerations.[9]
- ***Valid and Reliable for Intended Mode of Administration.*** A measure might not be valid or reliable as reported if it is used with a group for whom it was not designed or with a mode of administration for which its reliability and validity have not been tested.

[9] *Content validity* indicates that the set of items comprising a measure is a good representation of the construct being measured. *Concurrent validity* refers to sufficiently large correlations between the measure and another measure of the same construct (measured contemporaneously in the same sample). *Predictive validity* refers to a sufficiently large correlation between the child outcome measure and a subsequently measured construct that is theoretically associated with the child outcome.

- ***Prior Use in Large-Scale Surveys and Intervention Evaluations.*** Prior use is helpful because it suggests that the measure is practical and feasible to use in large-scale research and because it provides a benchmark for the scores of children participating in the evaluation.
- ***Cost and Burden.*** The cost or burden of data collection strategies, including training requirements, respondent burden, and program administration burden, should be minimized.

Table V.4 describes the classroom and child outcomes to be measured and the recommended measures of each one. Measures of teachers' knowledge and attitudes about developmentally appropriate practice can be taken from the Family and Child Experiences Survey (FACES). Measures of the classroom environment will include practices that support social and cognitive well-being as well as overall classroom quality. Measures of children's behavior will be based on teacher reports and direct observation. Measures of children's language, early literacy, and mathematics ability could rely on the NRS assessments, as those outcomes are more distal from the behavioral outcomes targeted by the intervention. For Option 3, which focuses on outcomes for children likely to receive intensive services for serious behavioral issues, a parent interview or combined interview and home visit will be important for gauging the impacts of the services.

To obtain data for the impact analysis, a baseline assessment of children's behavior will be obtained in the fall from teacher reports and time-sample observations; a similar follow-up assessment will be conducted in the spring. Children's progress in language and mathematics will be assessed by the fall and spring NRS. Because three-year-old children in the sample will not receive an NRS assessment, we recommend using the NRS instrument for three-year-olds in the spring followup only. A teacher interview, also conducted during the fall and spring, will obtain information on teachers' attitudes about and knowledge of developmentally appropriate practice. The classroom observations (in the enhancement and control classes), which needs to be conducted only during the spring, will measure overall classroom quality and the classroom management strategies that teachers learned as part of enhancement training. Taking stock of the overall quality of the classroom is an important step in helping to understand the pathways by which the classroom management approaches may be changing practice and potentially affecting child learning. Children's demographic characteristics will be obtained from an information sheet completed by parents as part of the consent process; for Option 3 (which includes a sample of children with high levels of reported behavioral problems), this form also will include parents' reports of their children's behavior. A parent interview conducted in the spring, either by telephone or in person, will provide information on parenting practices, the home environment, and the parent's perspective on the child's behavior.

Before any data are collected, the research team must draft instruments, obtain IRB and OMB approval, obtain parents' consent for their children's participation, and randomly select children to participate based on their eligibility status and their consent status. We discuss these steps in the remainder of this section.

Table V.4: Measures of Intermediate and Child Outcomes Associated with a Social-Emotional Behavioral Intervention

Outcome	Recommended Measure	Type of Measure
Teachers' Knowledge		
Attitudes and knowledge about developmentally appropriate practice	Teacher Beliefs Scale (Burts et al. 1990)	Teacher survey
Classroom Processes		
Materials and teacher activities to promote learning	Early Childhood Environment Rating Scale—Revised (Harms et al. 1998)	Observation and Teacher survey
Classroom practices that promote positive behavior and minimize disruptive behavior	Adaptation of the Inventory of Practices for Promoting children's Social and Emotional Competence (Center on the Social and Emotional Foundations for Early Learning 2003).	Observation
Children's Development		
Behavioral problems	Child Behavior Checklist (Achenbach and Rescorla 2000)	Teacher report
Social competence	Social Competence and Behavior Evaluation (LaFreniere and Dumas 1995)	Teacher report
Approaches toward learning	Preschool Learning Behaviors Scale (McDermott et al. 2000)	Teacher report
Child behavior toward peers and adults in the classroom[a]	Howes Peer Play Observation Scale adapted for FACES	Observation
Mathematics ability	Woodcock-Johnson Applied Problems (Woodcock et al. 2001)	NRS assessment
Language ability	Peabody Picture Vocabulary Test, 3rd Edition (Dunn and Dunn 1997)	NRS assessment
Early literacy	Woodcock-Johnson Letter-Word Identification (Woodcock et al. 2001)	NRS assessment

[a]Because this measure is very intensive and focuses on one child at a time, it is recommended for use only as part of an evaluation of Option 3, which focuses services on individual children most likely to be referred by teachers for intensive, individualized services to address behavioral issues.

FACES = Family and Child Experiences Survey; NRS = National Reporting System.

Develop Data Collection Instruments. Data collection instruments will be written during the first year of the study. The instruments will include the consent form, with a form requesting demographic and child behavior information; the teacher survey; the child assessment and observation protocol; and the classroom observation protocol. Many of these instruments will include standardized assessments that will have to be formatted to simplify administration by a trained assessor. Others (such as the teacher questionnaire) will rely on questions that have been used in other studies. After the data collection instruments have been developed, they will be pretested to ensure that respondents understand the questions, that the flow proceeds logically and smoothly, and that the time required to complete them is reasonable.

IRB and OMB Research Review. Research on human subjects must be reviewed by an IRB, which considers the benefits of the research to society, the programs, and the participating families and weighs those benefits against the cost of the research to the families and program staff. The IRB also reviews protection of research participants from harm by ensuring that confidentiality is maintained. In addition, if the evaluation is federally funded, data collection instruments and the research plan must be approved by OMB. The data collection instruments are reviewed to ensure that they do not overlap with other ongoing federal data collection efforts, and that burden is not excessive. These reviews will be conducted during the nine months preceding the start of the evaluation year, while the intervention is being implemented.

Consent. Active consent for children to participate in the research must be obtained from parents or guardians. The parent consent form will clearly inform parents (guardians) about the duration of the study, the types of assessments that will be administered, and the voluntary nature of participation. An information sheet will be included in the consent package to collect basic demographic information about the family, such as age, race and ethnicity, and family structure, as these variables will be needed to improve the precision of impact estimates, as well as to define subgroups. If the sample is to include children with high levels of reported behavioral problems, the consent form and information sheet will include a questionnaire for the children's parents that asks about child behavioral problems as well as positive social behavior, for balance.

The consent process could proceed more smoothly if it is incorporated into the home visits that many programs make to families. During these visits, which typically occur just before children attend class for the first time, teachers bring forms that parents must complete before the start of the school year. The teachers are available to explain the forms, and to ensure that they are completed correctly. If the study's consent is part of this process, the teachers would be able to explain the nature of the study, what will happen if the children participate in the research, and the voluntary nature of the children's participation.

Child Eligibility Criteria for the Study. Because the sample of children for Option 1 or Option 2 is intended to represent all three- and four-year-old children in each center, all children participating in Head Start should be eligible for the study. Although teacher training will begin in the final months of the preceding Head Start year, we expect that children attending Head Start the next year will have very little exposure to a fully

implemented enhancement. We therefore recommend including all Head Start children in the potential sample for the study, rather than limiting the sample to children new to Head Start during the year in which fall and spring data are collected.

For Option 3, researchers need a sample of children with high levels of externalizing behavioral problems who are thus most likely to be referred for intensive, child-focused services during the year. These children will be selected based on parent reports at baseline of high levels of aggressive behavior, hyperactive behavior, oppositional defiant behavior, and emotional reactivity. Based on previous studies of Head Start classrooms, we expect approximately 10 to 20 percent of children to meet the threshold for this sample.

Selection of Classrooms and Children for the Study. When classrooms and children's enrollments are established in August, researchers will work with center directors to obtain a list of teachers (in three- and four-year-old classrooms) and the number of children enrolled in each classroom. Two classrooms from each center will be selected with probability of selection proportional to class size. Parent consent forms will be distributed to parents of children in those classrooms. Returned consent forms and associated information sheets will be sent by program staff to the researchers for processing. The researchers will enter data from the forms to classify children by center, by classroom, by consent status, and by demographic characteristics. A sample of nine children from the pool of eligible children in each research classroom will be randomly selected for the data collection. For Option 3, the information from parent reports of children's behavioral problems will be entered and assessed. Approximately five children per classroom who meet a threshold cutoff for high levels of behavioral problems will be included in the sample.

Parent consent rates typically vary across classrooms, as rates at which parents return the forms and consent can depend on how organized the parents are, their degree of connection with the classroom teacher, whether parents talk to each other about participating in the study, and how diligently the teacher follows up with parents to return the forms. However, the study will typically not have information about children without parental consent; in this case, the study will have to generalize the results based on those with consent to the entire class. However, if some aggregate, class-level data are available (such as demographic composition or scores from the NRS assessments), this information can be used to adjust classroom-based estimates for nonresponse by weighting the responders to reflect the true aggregate classroom-level composition.

Teacher Self-Administered Questionnaire. Teachers' attitudes about managing the classroom to promote positive behavior and to minimize opportunities for negative behavior can influence how well the teachers implement the enhancement approach. We recommend giving teachers a short (no more than 15-minute) self-administered questionnaire during the fall, while the child assessments are in progress. This questionnaire could be adapted from the FACES teacher interview and could include questions about the teacher's background, attitudes, and beliefs about developmentally appropriate practice, as well as a short behavior problems scale for children in the research sample. During the spring data collection, the teacher questionnaire will omit the teacher background questions (unless the teacher is new) but will include a short behavior problems scale and a short scale measuring approaches

toward learning for each child in the research sample. The behavior problems scale and the social skills rating scale will provide additional information on children's behavior based on the teachers' observations during the year. The measure of approaches toward learning will help to identify positive factors, such as curiosity and attentiveness, that are associated with academic success.

Classroom Observation. Measuring the overall quality of the classroom with a commonly used observational protocol will enable researchers to understand how the classroom management techniques have contributed to the overall quality of the classroom environment. Thus, we recommend using the Early Childhood Environment Rating Scale-Revised (Harms and Clifford 1998) to measure classroom quality. In addition, a measure of the fidelity of teacher practices to the enhancement is important for gauging whether the measured impacts correspond to generally high-fidelity implementation, or whether the enhancement was not fully implemented. We recommend that the evaluation team work with the Center on the Social and Emotional Foundations for Early Learning to adapt its Inventory of Practices for Promoting Children's Social and Emotional Competence into a briefer observation tool. This tool would examine the ways in which a teacher builds positive relationships with children, creates a supportive environment, and promotes social and emotional well-being through his or her teaching strategies. We recommend that the classroom observations be conducted during the two weeks before the spring follow-up assessments begin.

Parent Interview and Home Visit. This portion of the data collection plan is an option if the sample includes children likely to be referred for intensive, individualized services. These services are likely to involve not only the child, but also the teachers and parents, who will be shown how to interact more positively with the child, and how to create an environment at home and at preschool in which the child can thrive. Because of the greater cost associated with home visits, parent information could be collected via telephone interview; however, the home observations have greater validity. We recommend using the Home Observation for Measurement of the Environment (HOME; Caldwell and Bradley 1984), preschool form. Numerous studies have linked scores on the HOME with children's cognitive, social, and emotional well-being. A short version of the HOME has been developed for the National Longitudinal Study-Child Supplement, which includes items that can be asked during a telephone interview. We also recommend using the Parenting Stress Index-Short Form (Abidin 1995) to measure stress in the parent-child relationship; this parent-report measure could be administered either in person or by telephone. The parent interview also would include measures of the child's behavior from the parent's perspective, including the Child Behavior Checklist (Achenbach and Rescorla 2000) and the Social Competence and Behavior Evaluation (LeFreniere and Dumas 1995) or Social Skills Rating Scale (Gresham and Elliott 1990).

Child Assessment. The primary hypotheses about the impacts of the improved classroom management techniques are whether the enhancement (1) improves children's social behavior, and (2) reduces behavioral problems. Accordingly, we recommend that the child assessments focus on children's social behavior. The Howes Peer Play Scale has been adapted for the FACES project to enable observers to measure the activities and behaviors

of as many as six children per classroom. This approach would accommodate assessment of the sample of children proposed for this evaluation. The measure could be adapted to yield measures of children's prosocial behavior, conflicts with other children, and other important outcomes. As a secondary effect, the classroom management techniques in this evaluation are intended to enable teachers to spend more productive classroom time supporting children's language, early literacy, and early mathematics skills. To address the research question about children's progress in those areas, we recommend using the NRS assessments. NRS assessment data potentially could be available for all of the four-year-old children in the sample, and they also could be administered easily to the sample's three-year-old children during the spring followup. Using the NRS assessments would save the expense and the burden on programs and children of collecting an additional 20 to 30 minutes of assessment data on children's language, early literacy, and mathematics skills. The child assessment protocol could then focus on obtaining child behavioral measures more central to the evaluation.[10]

An intervention to influence children's behavior is highly unlikely to have an immediate—or even near-term—effect on the children. Moreover, some children may act out at the beginning of the Head Start year while they are adjusting to new groups of children, to their classrooms and school schedules, and to their new teachers. Therefore, a very early baseline is unlikely to provide a valid measure of a child's behavior. Instead, we recommend a field period that starts at least one month after classes have begun at the Head Start center, with a six-week window for data collection at the sites. To assess how children's behaviors have been influenced by their experiences in the Head Start classroom and by what they have learned during the Head Start year, the follow-up assessment should be conducted as close to the end of that year as possible. We recommend that the follow-up data be collected during a six-week window that ends two weeks before the end of the year, and that data collection be matched to the timing of the fall assessment so that classes assessed early in the fall field period also are assessed early in the spring field period.

Costs of the Enhancement. Program administrators care not only about the effectiveness of the quality enhancements, but also about the costs of these enhancements over and above current expenditures. To support analyses of the cost of the enhancement relative to its impacts (for cost-effectiveness analysis) or benefits (for cost-benefit analysis), researchers will collect information on the costs of implementing the classroom behavioral approach and information on the additional cost of the intensive child-focused services relative to the cost of the program without either enhancement. Researchers should measure two types of costs: (1) the upfront, one-time costs of beginning implementation of the enhancement; and (2) the ongoing additional costs of the enhancement. Among the first set of costs are the cost of initial teacher training, including the enhancement developer's and training staff's time; the cost of substitute teachers hired to cover classrooms while the

[10] NRS data are not currently available below the program level, and the Head Start Bureau has not given permission for the programs to keep copies of children's NRS assessments. The Head Start Bureau would have to approve the use of NRS assessment data at the individual level to support evaluation of Head Start quality enhancements.

regular teachers attend training; and the cost of additional staff days for teachers who are paid for their days of training. The first set of costs also includes costs associated with the training staff's technical assistance visits and any costs associated with teachers attending extra training sessions outside their normal classroom duties (for example, periodic group discussions, if any). Time spent by teachers and trainers would be valued at these staff's hourly wage, including fringe benefits. If the training were to supplant the usual teacher training sessions conducted during the year, those routine training costs would be subtracted. Usually, however, new enhancement training is an add-on expense, rather than a substitute one. If space for training is rented, the cost of the rental is included as well. Materials, including documents, videos, and other training materials, should be valued at their cost.

The ongoing additional costs of the classroom-level behavioral enhancement would include the costs of ongoing technical assistance and costs to train new teachers. Teachers might require additional materials for the classroom, either to organize the space to prevent conflicts from occurring (for example, between children engaged in quiet activities and children engaged in more active, noisy activities), or to ensure sufficient supplies of popular items for all children who may want to use them. In addition, the cost of the mental health professional who consults with teachers about particular children and provides intensive services to children must be included in the cost of the enhancement; that individual's hourly rate would be used, with any corresponding reduction in other child mental health services used by the program subtracted. The cost of refresher training and ongoing technical assistance to teachers during a normal program year to ensure that the classroom management techniques continue to be implemented to high fidelity is another ongoing cost. For example, technical assistance staff might make two or three visits to each classroom during the program year to observe, measure fidelity, and meet with the teachers and with the Education Coordinator to discuss classroom practices and to respond to questions. Finally, overall teacher turnover in Head Start is approximately 15 percent per year (National Institute for Early Education Research 2003). Accordingly, an ongoing cost of the enhancement is the cost of training and technical assistance to 15 percent of the teachers in the enhancement group each year. Assuming that the evaluation includes 50 to 150 centers in the enhancement group, 8 to 23 teachers would have to be fully trained each year.

Three different approaches to estimating the costs of implementing the enhancement could be used:

- Identify the costs associated with implementing the classroom management approach and the services of the mental health professional for individual children and ask center directors for this information
- Obtain the full budget for the enhancement center for the year prior to enhancement implementation and the full budget for the current year and estimate the cost of the enhancement as the difference in the two budgets (adjusting for the normal cost inflation from one year to the next)
- Obtain the full budget for the enhancement centers and the full budget for the control-group centers and estimate the cost of the enhancement as the difference between enhancement and control-group center costs

The first approach is the least burdensome for the enhancement centers (and eliminates burden for the control centers) but could fail to include costs associated with the enhancement. The second approach is more burdensome for the enhancement centers than is the first one, and it does not require a response from the control centers. However, if anything other than the implementation of the enhancement changes from one year to the next, the estimate of the cost of the enhancement will be inaccurate. The third approach is the most burdensome, involving extensive collection of cost information from both enhancement centers and control centers, but it would provide the most accurate estimate of the costs of implementing the enhancement. If a formal benefit-cost analysis is to be performed, the more accurate cost data would be needed.

Information about these costs can be obtained from semistructured interviews with program and center directors, from the enhancement developer and training/technical assistance supervisory staff, and from program records. All cost information must be obtained in dollars; however, to ensure that the dollar values collected at various time points represent the same value, the dollar values collected from informants and records should be either inflated or deflated, using the Consumer Price Index, to represent dollar values in a single target year (for example, the analysis and reporting year).

Finally, cost information sometimes is obtained using two perspectives. First, the actual dollar costs of the enhancement must be obtained. Second, a measure of cost to society would place a value on any volunteer labor or donated space and materials and would add those costs to the actual expended costs. Both cost perspectives would be used in the analysis of benefits and costs or cost-effectiveness of the enhancement.

E. ANALYSIS AND REPORTS

Reports based on the evaluation should report the estimated impacts of the enhancement on teachers' practices, classroom activities, and children's outcomes. They also should describe the implementation experience and should discuss how the classroom management approach and child-focused services could be implemented on a broad scale, using the Head Start training and technical assistance system for support. Reports should be written for a broad audience, with stand-alone summaries that can be understood by program staff and policymakers, and more-detailed reports that summarize the research design, sample characteristics, analytic approaches, and findings in a clear and accessible way. In this section, we discuss the approach to estimating impacts and conducting the cost-effectiveness analysis.

1. Estimating Impacts of the Behavioral Enhancement

Using a random assignment evaluation design means that fairly simple estimation methods can be used to determine the impacts of the quality enhancements at a point in time. Under random assignment of centers, the center-level mean outcomes are estimated, after which, separately for the enhancement group and the control group, the center mean outcomes are averaged over all centers in that group. The simple difference in the means between enhancement and control centers is the impact estimate.

More-precise estimates can be obtained by estimating regression models. Regression procedures can improve the precision of the estimates and can adjust for any residual differences in the observable characteristics of program and control group members due to random sampling and interview nonresponse. Regression models take the following form:

(1) $Y = \alpha + X\beta + \gamma T + \varepsilon$,

where Y is an outcome variable; X is a vector of explanatory variables; T is an indicator that equals one for members of the enhancement group and zero for members of the control group; α, β, and γ are parameters to be estimated; and ε is a random-error term. The estimate of the parameter γ is the estimated impact of the quality enhancement compared with regular Head Start services.

Because random assignment will have been conducted at the center level, the regression adjustment takes the form of a hierarchical linear model (HLM) of child development consisting of two nested levels. By specifying the model at each level, we can conduct analyses for the appropriate units of analysis and can conduct statistical hypothesis tests that correctly account for the clustering of children within classrooms. Although not strictly necessary for conducting impacts (because the evaluation is based on a random assignment design), adjusting the impacts with an HLM model can help to increase the precision of the estimates.

For the design involving random assignment of centers, the analytic model is the following, where the variables are indexed by child *(i)* and centers *(j)* :

Child-level model

(2) $Y_{ij}(t) = \alpha Y_{ij}(0) + \beta X_{ij} + c_j + \varepsilon_{1ij}$,

Center-level model

(3) $c_j = \gamma T_j + \delta Z_j + \varepsilon_{2j}$,

where $Y(t)$ is the outcome at follow-up period t; X is a set of child characteristics, such as gender; T is a variable indicating whether the child was in an enhancement classroom or center; Z is a set of center-level variables, such as whether classes are full-day; and ε_1 and ε_2 are disturbance terms assumed to have a mean of zero, and to be uncorrelated with each other. Parameters to be estimated include α and β, vectors of coefficients on the child baseline characteristics; c, the center effect; γ, the effect of the enhancement; and δ, the coefficients on the center variables.

The statistical techniques used to estimate regression-adjusted impacts in equations (2) and (4) will depend on the form of the outcome, Y. If the dependent variable is continuous (such as the score on the NRS Math Assessment), ordinary least squares methods produce unbiased estimates of the parameter γ. However, if the dependent variable is binary (such as

whether a child was rated as having behavioral problems in the clinical range), logit or probit maximum-likelihood methods will be used to obtain consistent parameter estimates.

This estimation model assumes that all centers are weighted equally so that essentially, the average outcome measure for each treatment center is averaged with those of all other treatment centers to obtain the mean outcome for all treatment centers. Larger centers thus do not receive greater weight in influencing the mean score for treatment (or control) centers. If the enhancement were implemented in a few centers with six classrooms and several others with two or three classrooms, we would not want the results in the larger centers to overwhelm the results for all treatment centers. Averaging results across all centers regardless of center size addresses the question, "Does the enhancement work in the average center?" This approach is appropriate because the purpose of the evaluation at Stage 2 is to measure how well the enhancement works in a collection of centers overall. At Stage 3, when the results of the evaluation will be representative of all Head Start programs, it will be important to address the question of whether the enhancement worked for the average Head Start child, and as a consequence, the impacts measured in larger centers and program grantees should have greater weight than those measured in smaller centers and grantees.

2. Subgroup Analyses

Because the effectiveness of the enhancement targeting children's classroom behaviors may differ by program setting or by characteristics of the children served, it would be useful to determine the groups for which the enhancement is most effective. This type of subgroup analysis would then enable individual programs to decide which enhancements might be useful to them. For example, analysis may demonstrate that the classroom management approach is effective only when teachers have less than a bachelor's degree.

The subgroups of interest will depend on the quality enhancement to be tested. Examples of center-level characteristics that can define subgroups include:

1. Full-day or part-day program
2. High-fidelity implementation or incomplete implementation
3. Teachers' qualifications
4. Center or program size

Examples of categories of child and family characteristics (measured prior to the experience of the quality enhancement) for subgroup analysis include:

1. Child's gender
2. Child's English proficiency

3. Mother's education level
4. Level of family income
5. Parents' employment status

The same procedures for calculating overall impacts can be used to obtain subgroup estimates, with the calculations made for particular subgroups.[11] Regression-adjusted subgroup estimates are obtained by introducing an interaction term that is the product of the treatment indicator and an indicator of membership in the subgroup of interest. This term is entered into the appropriate model shown in Section E.1 for the level of random assignment, and for whether the subgroup is defined at the child level or at the center level.

However, unless the overall sample is very large, it will be possible to detect impacts only for large subgroups of the population, as subgroup estimates, which are based on only part of the full sample, are less precise than are full-sample estimates. For example, our sample of 1,490 children (90 percent of the initial sample in 92 centers) is sufficient for detecting impacts with effect sizes of .20 (one-fifth of a standard deviation on the outcome measures) or more. However, for a subgroup that includes 50 percent of the children across all the centers (for example, children whose mothers have less than a high school diploma), impacts with effect sizes of .23 or more can be detected. For a subgroup that includes all of the children in half the centers (for example, full-day programs), impacts with effect sizes of .28 or more can be detected.

3. Cost-Effectiveness Analysis

A cost-effectiveness framework should be used to evaluate the costs and benefits of the enhancement. This type of analysis does not attempt to place a dollar value on impacts. Instead, impacts are measured in a common unit, such as an effect size (the impact divided by the standard error of the outcome measure). The impact in effect-size units is compared with costs measured in dollars. For each quality enhancement, an effect size per dollar spent on the enhancement can be calculated. For example, if the classroom management approach were to produce an impact on children's cooperative behavior of 0.3 in effect-size units, and if the cost were estimated to be $10 per child, then the cost-effectiveness of the enhancement would be 0.03 per dollar. Measuring cost-effectiveness in this way enables program administrators to compare the cost of producing impacts they consider important using the same metrics, so that enhancements can be assessed in terms of their ability to provide the most "bang for the buck."[12]

---

[11] For center-level random assignment, calculate center-level mean outcomes and average across centers in the enhancement and control groups.

[12] In contrast, a cost-benefit analysis requires researchers to convert impacts into "benefits" that are valued in dollars. This task is complicated by the fact that the evidence linking differences in child assessment scores with future employment and earnings, involvement with the criminal justice system, and use of public assistance programs is quite tenuous. Accordingly, we do not recommend conducting a cost-benefit analysis based on outcome data collected from a single year in Head Start.

The enhancement is likely to have impacts that vary in size across the outcomes measured. Therefore, the estimate of cost-effectiveness will depend on the outcome used to measure it. Researchers can report the range of cost-effectiveness estimates using the impacts on outcome measures considered to be most important. For example, the outcomes most important for a social-emotional behavioral intervention are children's behavioral problems, social competence, behavior toward peers and adults in the classroom, and approaches toward learning. Alternatively, the cost-effectiveness of several enhancements designed to influence children's social-emotional development could be compared using a common outcome measure.

CHAPTER VI

FIELD TESTING QUALITY ENHANCEMENTS

Quality enhancements that have been carefully developed, replicated, and evaluated on a small scale are ready for the next step: a broad field test in a representative group of Head Start program grantees and centers. Whereas the small-scale evaluation is designed to address whether the enhancement can work in Head Start programs that are willing to try it, the field test is designed to address whether the enhancement works in a group of programs that are representative of all Head Start programs in selected regions or nationally. To extrapolate the results of the evaluation to the Head Start program nationally will require a very large sample because of clustering at the classroom, center, and grantee stages. A larger sample increases the costs of implementing the enhancement, either through higher implementation costs or through data collection costs (or both), which will limit the number of field tests that can be completed.

Funding for programs such as Head Start must be used wisely to ensure that the programs are as effective as possible at enhancing children's development. Thus, conducting research and evaluation as new ideas are implemented can help to ensure that the Head Start program continually identifies good ideas and refines them so that, over time, the Head Start community learns what works best for children. In this way, Head Start can reinforce its position as a national laboratory for good early childhood practice.

At times, policymakers might be ready to move to a larger-scale implementation of an idea that seems to address a pressing Head Start program need. It is likely that several relevant ideas would be in the development stages or beyond, and reports summarizing experiences and lessons as the idea is refined would help policymakers to consider and, perhaps, test alternatives. Even if ideas have not yet emerged from Stages 1 and 2, an evaluation still could be conducted if they are implemented strategically to support a rigorous evaluation approach.

The newly instituted Head Start National Reporting System (NRS), which provides assessment data covering vocabulary, letter recognition, and early mathematics for all four-year-old Head Start children, has increased the urgency of conducting a large-scale evaluation of quality enhancement ideas. Head Start programs recently received the results of the first year of Head Start NRS testing and many wondered what their next steps should be.

Companies are offering preschool curriculum materials and technical assistance to help programs to address perceived gaps in instructional quality. Although there is likely to be no shortage of ideas for helping Head Start programs to improve their NRS assessment scores, there also is little solid evidence of the efficacy of the options being offered to programs. Many programs will adopt whatever approach or curriculum that is presented directly or is well advertised. However, decisions about alternative approaches would be better informed if the Administration for Children and Families (ACF) could spearhead an effort to systematically test alternative promising approaches.

This chapter describes research designs that can provide nationally representative estimates of the effects of Head Start quality enhancement ideas on classroom environments, teacher and staff practices, and children's development. We first describe the general approach to these designs, including ways in which the Head Start training and technical assistance (T/TA) system can be strategically deployed and administrative data used to conduct the evaluation at reasonable cost. Some quality enhancement ideas will be best implemented at the grantee level, while many others will be implemented at the center level. We have selected two examples of quality enhancements, a grantee-wide self-study and improvement process and alternative literacy curricula, which can be implemented at the grantee and center level respectively. These examples illustrate the issues in evaluating quality enhancement ideas at the Stage 3 level. Section B describes the design of an evaluation of a grantee-wide quality enhancement initiative to use program data in a self-study process to identify and implement quality improvement efforts, and Section C describes the design of an evaluation of alternative literacy curricula, which would be implemented within centers.

A. APPROACH TO FIELD TESTING QUALITY ENHANCEMENTS

Obtaining measures of the impact of quality enhancements that are representative of Head Start programs regionally or nationally requires a large sample of program grantees, centers, and children that are participating in the evaluation. The larger scale of the evaluation is feasible because of two important Head Start program resources: (1) the Head Start T/TA system; and (2) administrative data, including the NRS. These resources have the potential to provide an important foundation for implementing new enhancements and collecting data to support an evaluation so that the Head Start community can test good ideas and determine what works best for children and families. Nevertheless, to support large-scale evaluations, the T/TA system and the NRS would have to be used in new ways.

This section provides an overview of our approach to field tests of quality enhancement ideas in Head Start. We first discuss how the strategic implementation of new ideas can provide fairly quick evidence about the ideas' effectiveness, or about the relative effectiveness of alternative approaches. We then discuss the T/TA system and how it might support strategic implementation to improve our understanding of what works in Head Start, and why. Finally, we discuss how the administrative data collected regularly from all Head Start program grantees might support a field test.

1. Field Test Design

A field test of a Head Start quality enhancement would have many similarities to a national implementation of a new program enhancement. Like a national implementation, a field test ideally would involve implementing the idea in all ACF regions as well as in a large number of diverse programs.

One important aspect of a field test would distinguish it from a national implementation. To measure the impact of the enhancement on children's progress, a contrast must be established between the enhancement and something else of interest. In most cases, the most interesting question is whether the quality enhancement being tested is an improvement over usual Head Start practice. To answer this question, some Head Start programs or centers must be randomly selected to implement the enhancement, and others selected to not implement the enhancement. In other cases, policymakers might want to know whether one enhancement approach is more effective than another. To answer this question, some Head Start programs or centers must be randomly selected to implement one enhancement, and others must be randomly selected to implement an alternative enhancement.

A field test of quality enhancements will be large and geographically dispersed, which will pose a challenge for implementation. Ensuring that the enhancement is implemented to high fidelity will require a strong T/TA plan and careful monitoring. In addition, researchers will have to minimize the potential for spillover of the enhancement into control-group sites (or of enhancement groups across one another, if alternative enhancements are evaluated) to ensure that a contrast can be made between the enhancement and control group or between different enhancements in a large-scale initiative. Spillover can be minimized by randomly assigning grantees or centers rather than classrooms, because the challenges posed by classroom random assignment in a small-scale evaluation (discussed in Chapters IV and V) will be even more difficult to address in a large-scale evaluation. The scale and geographic dispersion of a large-scale evaluation will make it more difficult to discourage enhancement-group teachers from discussing the enhancement with control-group teachers. It also will be more difficult to monitor the changing configurations of classrooms and the attrition of teachers so that random assignments are preserved from the initial implementation year into the data collection year.

Random assignment of either program grantees or centers to implement an enhancement, an alternative, or neither option must occur independently of any considerations of program desires or initial characteristics. For this reason, random assignment should be conducted and monitored by a research organization that is independent of the programs. That organization would work with the regional offices and T/TA staff to ensure that the enhancements are fully implemented in the selected program grantees or centers.

2. Implementation

At the field test stage, enhancements would be implemented using large-scale methods commonly used to implement initiatives program-wide, but implementation would proceed

in a structured way: Grantees or centers assigned to the enhancement group would implement the enhancement, and grantees or centers assigned to the control group would not. Implementing the enhancement strategically in randomly selected grantees or centers will enable the Head Start community to measure the impact of the enhancement by comparing the outcomes of children in the enhancement centers (or grantees) with those in control centers (or grantees). Any differences in children's outcomes could be attributed to the enhancement with a measurable degree of statistical certainty.

Clearly, an important foundation to guide implementation of quality enhancements on a large scale is documentation about the enhancement and manuals describing how to implement it that have been refined during the first and second stages of research. These materials and a good understanding of key steps in implementing the enhancement are critical to supporting high-fidelity implementation at Stage 3, when the enhancement developer cannot exercise direct control over the day-to-day progress of implementation. T/TA staff will have to be trained to assist with implementation, measure fidelity to implementation, and address issues that arise during implementation.

Training methods used to implement program-wide initiatives could include national or regional training of trainers and distance learning approaches that use web-based lectures with electronic questions and feedback. The T/TA system could provide critical support for implementation. Here, we discuss each of these training resources. In addition to these resources, documentation on the enhancement (printed documents, training videos, and other materials) would be provided to centers that have been randomly assigned to implement the enhancement.

Training of Trainers. In this model, one or two staff from a local program receives training at a national or regional training conference. Training focuses on the enhancement, the theory behind the enhancement, and strategies for offering practical training on key concepts. The staff person then becomes the "trainer" for all other staff at the local program. This approach was used to implement strategies for enhancing classroom literacy environments (Strategic Teacher Education Program, or STEP training), as well as to train program staff to administer the NRS child assessment. The effectiveness of this approach depends on the skill levels of the local trainers, the complexity of the enhancement, and the extent to which key ideas can be covered during the time allotted for training. After receiving STEP training in summer 2002, the new early literacy specialists were charged with providing training to teachers in their local programs on the techniques they had learned, in preparation for the 2002-2003 program year. Because the early literacy approaches were not fully implemented in all Head Start programs, in November 2002, the early literacy specialists attended a second training to learn mentor-coaching skills that would enable them to observe teachers, and to provide ongoing feedback throughout the year.

Like national and regional training, the train-the-trainers approach has the potential to offer training provided by skilled trainers during the first round of training, and to ensure a high level of consistency across many programs, as all local trainers will have received the same instruction and materials. The weakness of this approach lies in the loss of control over the quality of the second round of training, when the local trainers return to their

programs to work with local staff. The skills and qualifications of local trainers are likely to vary widely across programs, as are the intensity and duration of training provided locally. Ultimately, the quality of implementation and degree of fidelity achieved at the local program level depend heavily on the ability of trainees to provide effective local training, and to ensure that implementation proceeds according to the model. The weaknesses of the approach could be addressed by requiring a higher minimum initial skill level for local trainers, thus ensuring that training is designed so that essential information is covered; "certifying" trainers at the close of training by asking them to demonstrate mastery of essential skills; and prescribing the content, frequency, and intensity of local training. In addition, T/TA staff could monitor implementation and could provide additional assistance, where necessary, to ensure high-fidelity implementation.

Distance Learning. Especially when programs do not have access locally to the training they need, distance learning is an alternative strategy for training teachers to implement specific enhancements in the classroom. One example is HeadsUp! Reading, provided by the National Head Start Association. Under this approach, teachers gather for weekly two-hour classes that take place during a satellite broadcast. Trained facilitators are on site to lead discussions and other classroom activities. One advantage of this approach is that many teachers—even those who live in remote, rural areas— can participate in training provided by highly skilled trainers at lower cost than the cost of traveling to a national or regional conference. HeadsUp! Reading also incorporates call-in segments that enable teachers to interact with instructors.

Distance learning is likely to be more cost-effective than on-site training, because a single, highly skilled trainer can train teachers from many programs simultaneously. However, local programs must have access to satellite equipment, and local facilitators must be recruited and trained. In addition, distance learning may be better suited to lower-intensity, longer-duration training courses in which classes meet for several hours per week over a period of months, rather than to intensive pre-service training in preparation for implementing a new curriculum or enhancement approach.

An alternative to a satellite broadcast is a webcast, which the Head Start Bureau currently is using to communicate with program staff about the progress and results of the NRS assessments, as well as other topics. Compared with satellite equipment, the equipment necessary to participate in a webcast might be more readily available to Head Start program staff, which could make it a more useful vehicle for communicating information about a quality enhancement. However, it would be very difficult to ensure that only the centers or program grantees that had been randomly selected to implement the enhancement participate in the webcasts.

Role of the Head Start T/TA System. Head Start's new T/TA system, initiated on September 1, 2003, consists of 12 contracted centers that are managed by the ACF regional offices. In the 10 geographically based regions, two or three T/TA managers and several content experts (in such areas as early literacy, disabilities, health, and administration) work out of the regional offices, with the goal of facilitating closer communication and coordination among Head Start offices and staff. In addition, the T/TA specialists work

directly with a group of 12 to 15 Head Start grantees, under the supervision of the T/TA managers.

T/TA specialists could play a critical role, if not in actually providing training, then in assessing the fidelity of implementation, and of ensuring that additional training and technical assistance are obtained by staff who need it. T/TA staff periodically visit all programs, and this direct contact could be valuable in ensuring that enhancements are implemented to high fidelity, and that random assignment is not compromised. T/TA staff who work with the programs selected for the enhancement group could be trained to assist with implementation, and to collect data periodically on the fidelity of implementation.

Implementing the enhancement with high fidelity is critical if the evaluation is to provide useful information about the effectiveness of the enhancement. A weakly implemented enhancement is unlikely to show impacts, wasting evaluation resources and subjecting the enhancement to an unfair test. Accordingly, implementation will require careful training and monitoring of the T/TA staff and of any other individuals charged with training staff or assisting with full implementation. This effort should be intensive enough to ensure that the enhancement is implemented well across many Head Start programs.

In arguing for an intensive, serious approach to implementing enhancements for a Stage 3 evaluation, we are not suggesting that the effort go beyond a reasonable effort to implement enhancements nationally. However, it is possible that the current approach to national implementation is not sufficiently intense. If resources required for high-fidelity implementation for a Stage 3 evaluation go beyond what is customary for a national implementation, the approach to national implementation might merit further consideration. In a national implementation, we cannot determine whether the enhancement is more effective than usual practice because no rigorous contrast will have been established. Without the prospect of a rigorous basis for comparing children's outcomes with and without the enhancement, some might assume that implementing any version of the enhancement will improve outcomes for children. That assumption might be called into question when the quality enhancement idea is being evaluated, as planners will more realistically understand that a weak implementation will likely generate weak or no impacts. However, if our approach to implementation nationally is not as intense or serious as that effort would be if we were expecting the enhancement to be evaluated, we risk wasting implementation resources on an activity that induces teachers and programs to act differently but does not ultimately benefit Head Start children.

3. Measuring Outcomes and Other Family and Program Characteristics Using Head Start Administrative Data

Collecting data for a field test of Head Start quality enhancements could be greatly simplified if outcomes measured by the NRS can be used. The NRS, first implemented in the 2003-2004 program year, is an ambitious initiative to systematically assess the early literacy, language, and numeracy skills of all four- and five-year-old children enrolled in Head Start. The NRS aims to collect information on a standard set of child outcomes from all Head Start programs in a consistent manner. It includes a 15-minute child assessment

battery; a system for training staff from all Head Start grantees to administer the assessment; and a computer-based reporting system that programs use to report information on a limited set of characteristics of participating Head Start programs, teachers, and children. The NRS thus offers the potential to provide measures of outcomes of Head Start quality enhancements.

Even though the Head Start Bureau has decided to report average scores and other information only at the grantee level, the data offer the potential to link an individual child's outcome data with demographic information about that child, and with data about the child's teacher (for example, education level) the center, and the grantee. NRS data at the classroom and center levels could provide a basis for estimating the impacts of enhancements that target outcomes measured by the NRS. However, before evaluations of Head Start quality enhancements requiring NRS data can proceed, ACF would have to approve this new use of the data.

Because of the potential usefulness of these data to field tests of Head Start quality enhancements, we provide background on the content of the assessments, the reliability of the data, and the quality of administration of the assessments. We then describe two additional sources of administrative data that can provide background information on programs, centers, classrooms, teachers, and children for the analyses.

Content of the NRS. The current NRS assessment battery includes four components:

1. ***Comprehension of Spoken English.*** This component is an English-language screener to identify children whose English is insufficient to participate in the full assessment.[1] It consists of items from two subtests from the Oral Language Development Scale of the PreLAS 2000 (Duncan and DeAvila 1998). The first set of 10 items uses the "Simon Says" game to request that children follow simple commands, such as, "Touch your ear" and, "Point to the door." In the second set of 10 items, children are asked to name or describe the function of objects in pictures.
2. ***Vocabulary.*** This section, adapted from the Peabody Picture Vocabulary Test, third edition (PPVT-III), includes 24 items that represent a range of difficulty. The PPVT-III was shortened for this purpose using Item Response Theory (IRT) techniques based on data from the field test and from the Family and Child Experiences Study (FACES).
3. ***Letter Naming.*** This section, developed for the Head Start Quality Research Centers curriculum intervention studies, presents all 26 pairs of upper- and lowercase letters of the alphabet in three groupings. (The Spanish-language version of the assessment, described below, contains 30 letters.) Children are asked to name the letters they know.

[1]When children who do not speak English as a home language make 15 or more errors on the language screener, the English assessment is terminated.

4. ***Early Math.*** This section, adapted from the mathematics assessment used in the Early Childhood Longitudinal Study—Kindergarten cohort (ECLS-K), includes 17 items on number understanding, shape recognition, relative size judgments and measures, and simple word problems involving counting or basic addition and subtraction.

A Spanish-language version of the child assessment also was developed. All children whose home language is Spanish are assessed in both English and Spanish, provided that they pass the language screener for each version of the assessment.[2]

Most Head Start program grantees have decided that the NRS assessments can be conducted most efficiently by someone other than the child's own Head Start teacher. In the 2003-2004 program year, the child's own teacher conducted approximately 35 percent of the assessments, other Head Start professionals conducted approximately 40 percent, and about 25 percent were conducted by contractors/consultants or others.

The specific vocabulary and math items in a given round of data collection are drawn from a pool of items whose difficulty is ranked using IRT methods. The item selection work is designed to keep the level of test difficulty constant across the annual administrations (fall to fall and spring to spring). However, the items on the test differ little from year to year. The similarity between tests from year to year raises concerns that, over time, Head Start staff could, "teach to the test."

Reliability of the NRS Data in 2003-2004. Analyses of reliability of the 2003-2004 NRS data were based on data submitted on nearly 407,000 children from 1,766 programs.[3] The internal consistency reliability of the spring 2004 data at the child level was good (.76 for understanding English and .81 for understanding Spanish; .81 for English vocabulary and .81 for Spanish vocabulary; .93 for English letter naming and .93 for Spanish letter naming; and .82 for English early math skills and .83 for Spanish early math skills). The reliability of the data at higher levels of aggregation, such as at the program level, was higher (more than .90 at this level). Thus far, the NRS data demonstrate that the three assessments reliably tap the targeted child outcome domains.

Quality of Administration of the Assessments. During the 2003-2004 program year, Mathematica Policy Research, Inc. conducted an implementation study of the NRS based on site visits to a nationally representative sample of 35 Head Start programs. Site visitors observed a random sample of approximately 10 child assessments per program, interviewed key Head Start staff about NRS implementation, and held a focus group with staff conducting the assessments to learn about their experiences. Using the certification

[2] Ninety-one percent of the spring 2004 assessments were in English, 21 percent in Spanish, and 13 percent were in both English and Spanish.

[3] As of August 2004, the computer-based reporting system (CBRS) data indicated that another 10,000 assessments had been completed but were not yet received. Spring forms were not submitted by 132 programs.

procedures and criteria as a guide, site visitors coded the observed assessments for errors and computed a certification score for each one. In spring 2004, 87 percent of observed English assessments scored 85 or higher, the minimum score required for certification (93 percent of Spanish assessments scored 85 or above). The study also compared the certification scores on assessments conducted by teachers who assessed children in their own classrooms with those of other assessors, as well as scores for experienced and new assessors in spring 2004. There were no significant differences in mean certification scores for teachers versus others who conducted the assessments. The overall error rate was low in both fall 2003 and spring 2004, with assessors making fewer errors, on average, in the spring than in the fall.

Strengths and Weaknesses of NRS Data for Evaluating Head Start Quality Enhancements. Although the NRS data provide a ready source of data on important child outcomes, the current scope of the assessment is quite narrow. There are plans to field test teacher-reported social-emotional measures next year, but enhancement studies may still need to be augmented with additional outcome measures rather than relying on the NRS as the only source of child outcome data. Two strategies are possible. First, if Head Start program staff are willing, one or two brief outcome measures relevant to a particular enhancement might be included as part of the center's data collection activities at the time that the NRS is conducted. Second, centers from the enhancement and control groups could be sampled for more-intensive measurement of implementation, interim outcomes, and child outcomes, much as FACES currently does.

One of the challenges of using the NRS in addition to other outcome data collected specifically for an enhancement study is that the data collectors across the two types of measures would be different; Head Start staff would collect the NRS measures, and study staff would collect the additional measures. As described above, the NRS data are collected by a variety of staff members with different levels of familiarity to the child, including the child's teacher, another Head Start staff member, and an outside consultant hired to conduct the NRS assessments. Study staff members would be "strangers" to the children, which could lead some children to perform differently on the assessments than they would if they knew the staff. Analyses to examine whether children perform differently on the same FACES and NRS assessments could address the question of whether having people with different relationships to the child conduct assessments would affect the quality and comparability of the data.

Plans indicate that the NRS will continue to evolve, and, in the future, it may tap additional Head Start Child Outcomes Framework domain indicators (see Appendix A). Expansion of the NRS, particularly in concert with the directions of major field tests of Head Start quality enhancements, would increase its potential as a source of outcome data in large-scale enhancement studies.

The Computer-Based Reporting System. The Head Start Bureau has implemented the CBRS to collect background information on Head Start programs, teachers, and children; to facilitate the identification of eligible children; and to track completed assessments. The CBRS is a web-based system in which Head Start staff enter program-,

classroom-, and child-level data (Table VI.1). After programs enter these data, the CBRS assigns unique identification numbers to Head Start grantees, centers, assessors, classrooms, and eligible children. In a field test evaluation of Head Start quality enhancements, classroom-, teacher-, and child-level information included in the CBRS could be used to improve the precision of impact estimates, and to define subgroups.

The Head Start Family Information System. The Head Start Family Information System (HSFIS) is a computer-based management information system developed by the Head Start Bureau during the mid-1990s. The HSFIS stores information on family and child characteristics relevant to determining eligibility for Head Start, as well as for determining service needs. HSFIS also stores an ongoing record of program services received, including child development services, parent services, health and mental health services, and others. These data have the potential to provide information on intermediate outcomes of Head Start quality enhancements, as well as background variables on families and children that could help to improve the precision of the impact estimates.

The potential for using the HSFIS in an evaluation is diminished by the fact that many programs have not adopted the system. Nevertheless, these programs might be collecting similar data on a small number of alternative management information systems. Thus, it might be possible to use program management information system data in an evaluation if a sufficient number of useful variables can be obtained with reasonable amounts of effort.

In the following sections, we discuss research designs for field tests of two enhancements. The first is an evaluation of T/TA approaches to improving program quality through a year-long self-evaluation process. Because the self-evaluation process involves all levels of program staff, from the director, through education and other service coordinators, to the center and classroom levels, this enhancement would be implemented grantee-wide and would therefore involve random assignment at the grantee level. This research design would be applicable to any enhancement that is implemented grantee-wide because it influences levels of staff beyond the individual center. The second research design is an evaluation of alternative curricula to enhance language and literacy. These curricula could be implemented at the center level, so we describe a center-level random assignment plan. This research design is applicable to any quality enhancement implemented at the center level for which ACF is interested in trying alternative approaches.

B. TRAINING AND TECHNICAL ASSISTANCE TO SUPPORT PROGRAM ASSESSMENT TO IMPROVE QUALITY

With completion of NRS assessments of four-year-old children for fall 2003 and spring 2004, Head Start programs received their first "progress reports." These initial reports to programs on 2003-2004 assessment data provided results only at the program level (grantee or delegate agency); results were not broken out by center, classroom, or individual child. For each subscale of the assessment, programs received mean scores for their program and for all Head Start programs. Summary results for each of the eight assessments (four

Table VI.1. CBRS Data Elements

Program	Center
Program name[a]	Center name
Director name[a]	Center contact information
Director email[a]	Center type (Head Start center, family child care home, home visitor cluster, child care partner)
Agency description[a]	Enrollment year start date
Number of delegates[a]	Enrollment year end date
Number of centers[a]	Center NRS lead name and contact information
Program auspice[a]	
Number of family child care homes	
Number of home visitors	
Program NRS lead name and contact information	

Classroom	Teacher	Child
Teacher's name	Teacher's name	Child's name
Class session (am, pm, mwf, or tth)	Teacher's language fluency	Date of birth
Classroom type (5 days, 4 days, home-based option, family child care, locally-designed option)	Language in which teacher provides instruction	Child entry date into classroom
Day option (part or full)	Total years of teaching experience	Child exit date from classroom
Total enrollment	Total years of Head Start teaching experience	Child unique ID from center
Number of child development staff in addition to the lead teacher	Highest grade or year of school completed	Number of prior years in Head Start
Teacher date of entry to classroom	Highest degree in Early Childhood Education or related field	Disability status
		Other languages spoken
		English-speaking ability
		Primary language spoken at home
		Child race/ethnicity

Assessor	Assessment Information
Assessor's name	Child ID
Highest grade or year of school completed	Assessment date
Highest degree held in Early Childhood Education or related field	Completion status
Assessor's program position	Session (fall or spring)
Other comments	Language (English, Spanish, both)
	Assessor ID
	Whether assessor is child's teacher

Source: National Reporting system computer-based reporting system: User's Manual, July 2004.

[a]Imported from the Head Start Program Information Report (PIR).

English and four Spanish) were presented by reporting the distribution of results across six skill levels that ranged from lowest performance (none or the easiest questions answered correctly) to highest performance (many or the most difficult questions answered correctly).

Many programs questioned the meaning and implications of the NRS assessment results. If their program's average spring scores were lower than the average for all Head Start programs, or if the change in average scores from fall to spring was smaller than the average for all Head Start programs, what might they do to improve children's performance? Head Start programs collect data from many sources for management purposes, but only a few programs have developed systems and expertise to analyze the data to identify areas for improvement. Many programs would benefit from technical assistance to help them to understand how their data can illuminate program strengths and weaknesses, and how to use that information to identify ways to improve the program. Currently, the T/TA system has begun to work with grantees on a system for program self-evaluation based on a manual developed in Region I. This research design describes how an evaluation could be designed to accompany that effort.

Programs seeking to improve their services have many sources of data to consult in addition to the NRS, as several types of data must be collected regularly:

- Assessments of children conducted by the program during the year, using locally selected measures and gauging children's progress across a broad set of child outcomes
- Services received by children and families during the year; intensity; and participation rates
- Family and community service needs

Those data could be used to identify areas on which the programs could focus to improve the match of services and family needs, or to enhance children's outcomes in a particular area.

In addition, to identify areas that are strong and areas that could be improved, programs could collect data that are not required, but that would shed light on their operations:

- Observational assessments of the quality of classroom instruction, availability of materials to support learning, and adult-child interactions
- Staff satisfaction surveys
- Discussions with teachers, staff, parents, and partners in the community about service needs, adequacy, quality of current services, and areas for improvement

The additional sources of data would assess the quality of particular program service areas that teachers or managers identify as critical to overall service quality.

The idea of collecting and using data to monitor program operations and identify areas for improvement is not a new one. Some grantees already have implemented broad data collection efforts that include the areas identified above, and they have procedures and a schedule for analyzing the data, identifying areas for improvement, and setting priorities for the upcoming year. The process results in consistent measures of program strengths and weaknesses that can be tracked over time, as well as an action plan for improvement for the next year.

Nevertheless, although all Head Start programs collect a substantial amount of data on services and outcomes, many do not fully use the information to identify areas for program improvement. T/TA staff thus may be able to help program staff develop a self-study and individualized improvement process. A recent Quality Research Consortium project involved training Head Start teachers, education coordinators, and the director to conduct regular assessments of children and of classroom quality, and to analyze the data to identify areas in which practice could be improved. An interventionist helped the teachers and managers to understand what the assessment data indicated, and how it could translate into recommendations for improvements. In a broader implementation of this idea, the T/TA system could work with programs to select appropriate assessments, interpret the data collected, and identify areas for improvement. Over time, program staff would gain an understanding of the stages of data collection, analysis, and interpretation, and they would be able to manage the continuous program improvement feedback loop themselves.

1. Study Design

An enhancement to help Head Start program grantees to use data to improve program services so that children's outcomes are measurably improved could be evaluated based on random assignment of grantees and a comparison of children's progress toward academic and social competence. This section discusses key elements of a research design to evaluate the impacts of this enhancement on children's outcomes. We begin by discussing the quality enhancement and its counterfactual, the research questions, the sampling strategy, random assignment plan, and sample sizes.

The Quality Enhancement, Counterfactual, and Research Questions

At the core of this quality enhancement is a feedback loop that includes collecting data on program services and outcomes; analyzing the data to summarize performance, and to identify strengths and weaknesses; discuss progress and weaknesses; and identify priority areas for improvement and methods of addressing those priorities. Ideally, data will be collected at the grantee, center, classroom, and child levels, and measures will be chosen carefully to ensure reliability and relevance. Different program staff members can be responsible for collecting specific types of data. The full range of program stakeholders, including staff, parents, and community partners, should see the results of the performance assessments and should have an opportunity to suggest areas for improvement. The program management team would set priorities based on the ideas submitted by all stakeholders.

A reasonable approach to evaluating this quality enhancement idea would be to contrast the continuous feedback approach to current practice, which will be a minimal self-study process. Because a basic approach to self-study is about to be implemented nationwide, it might be useful to describe how that effort could be evaluated.[4] If ACF were interested in evaluating this initiative, a randomly chosen set of grantees in each region could be selected to not implement the procedures so that a contrast could be made in each region between grantees implementing and not implementing the self-study process. The implementation process would likely vary across regions, depending on the T/TA organization and the emphasis placed on this task by the ACF regional office. With sufficient sample, regional subgroup impacts could be estimated to examine these potential differences.

Research questions to be examined as part of the evaluation include whether the enhancement was implemented to high fidelity:

- Did staff implement all features of the enhancement, including data collection, data analysis to identify program strengths and weaknesses, identification of strategies for program improvement, and setting priorities for improvement plans for the coming year?
- What level of T/TA was needed to implement the enhancement? What were the education levels and experience levels of T/TA staff? How much T/TA was provided, and over what time period?
- What challenges were encountered in implementing the enhancement, and how were they resolved?
- To what extent are control-group programs collecting data, analyzing the data, and identifying areas for improvement?

Another set of research questions focuses on the impacts of the enhancement on the quality of program services:

- Is the quality of the classroom environment and learning activities higher in programs that implemented the enhancement relative to the control group?
- Is the quality of teacher-child interactions higher in programs that implemented the enhancement relative to the control group?
- Is parent participation in the program's parent education programs higher in programs that implemented the enhancement relative to the control group?

[4] Another evaluation design would evaluate alternative approaches to providing T/TA to staff that would result in full implementation of the measurement, analysis, and feedback procedures. For example, the level of education or amount of experience of T/TA staff who help with implementation could be systematically varied, as could the number and content of meetings with program staff to assist with implementation. Alternatives tested should have a strong likelihood of being effective at fully implementing the process.

- Are children who need health, mental health, and disabilities services linked with these services more effectively in programs that implemented the enhancement relative to the control group?

Whether the enhancement yields measurable program improvements is only part of the story. The reason for engaging in the continuous program improvement process is to improve the program in ways that enhance children's outcomes. Thus, a set of research questions focuses on children's development:

- Do children in programs using a continuous improvement process progress further in vocabulary, early literacy skills, and early mathematics skills than those in control-group programs?
- Do children in programs using a continuous improvement process show greater social skills and fewer behavioral problems than those in control-group programs?
- Do children in programs using a continuous improvement process show greater sustained attention to task, greater engagement of adults and peers, and more self-control than those in control-group programs?

Finally, although the evaluation will examine the effectiveness of the enhancement for children overall, the Head Start community also will be interested in whether the enhancement is effective across different subgroups of children and families, and across programs with different characteristics:

- What are the impacts of the continuous improvement process on outcomes for key subgroups of children? What are the outcomes for Head Start programs with different characteristics?

If many subgroups are examined, some impacts will emerge simply by chance, so some caution must be exercised in examining subgroup impacts. To guard against finding impacts by chance, researchers can implement a Bonferroni correction that essentially inflates the standard error of the impact estimates to correspond to the number of hypothesis tests. This provides a higher bar for deciding that a subgroup impact is significantly different from zero.

In a Stage 3 design that is nationally or regionally representative of all Head Start programs, subgroup analyses should be based on a representative sample of that subgroup in Head Start. Therefore, subgroups of interest must be identified in advance so that statisticians can design a sampling plan for grantees, centers, and children that ensures adequate representative samples of the subgroups of interest.

For an evaluation of a program self-assessment process, the ACF regions would be useful subgroups because the strategies for implementing the self-assessment process will vary somewhat across regions. Because there are 10 geographically based regions, they will have to be grouped for analysis to keep the sample size within reasonable bounds.

Obtaining information about each region's implementation strategies would help to group regions with similar strategies. Because approval from the Office of Management and Budge (OMB) is necessary before contacting regions and T/TA contractors about implementation strategies, the information required to form regional subgroups would be available only after the sample has been drawn. To permit flexibility, sampling would be designed to be representative of each region, but the precision of the estimates would be lower at the individual region level, under the assumption that at least two or, possibly, more than two regions would be grouped for analysis.

Another subgroup of interest is program schedule, particularly full-day versus part-day. Because of the cost differences between the two program types, policymakers are interested in learning how much more full-day programs can contribute to children's cognitive and social-emotional development relative to part-day programs. A related question is whether quality enhancements can contribute more to children's development when they are implemented in full-day programs than when implemented in part-day programs. Information about the proportion of children served in part-day and full-day programs is available at the grantee level; these proportions vary considerably by region (for example, 20 percent of children in the western regions are served in full-day programs compared with 74 percent of children in the southern regions; see Table II.1). Unless the sample is very large, it will not be possible to examine subgroups defined by region and program schedule; therefore, we recommend drawing a sample of grantees that is regionally representative, but using program schedule as an implicit stratification variable at a later stage of sampling (center-level).

Sampling, Random Assignment, and Sample Sizes

An enhancement that implements a program self-assessment process would be most efficiently and effectively implemented at the grantee level. Programs would be more likely to improve if all centers are engaged in the continuous feedback process, and staff of the entire program work together. Working with one or two centers per grantee to implement this process center-wide might be useful, but it would fail to generate the synergies that would result from program-wide implementation that would enhance the quality of implementation.

Because implementation would take place at the grantee level, random assignment to the enhancement or control group also would have to be at that level. Within strata defined by ACF regions, urbanicity, and percentage of children who are African-American and percentage who are Hispanic or Latino (large versus small percentage), a sample of program grantee or delegate agencies would be selected using the most recent Head Start Program Information Report (PIR) data as the sampling frame. The PIR contains data from annual reports submitted by all program grantees and delegate agencies about program size, schedule, and other characteristics. The selected grantees and delegate agencies would then be randomly assigned to the enhancement or control group. This design assumes that ACF is planning to implement this enhancement nationally while retaining a control group for evaluation purposes; therefore, the control group would not receive information or T/TA to

implement the self-assessment process, while the enhancement group and any programs other than the control group would implement the enhancement.

The least expensive source of outcome data for an evaluation at the grantee level is NRS data. NRS data are available for all four- and five-year-old children in every grantee; therefore, if they can be obtained and analyzed for the sample of grantees included in the evaluation, all centers and children in those grantees can be included in the evaluation as well. We assume that an average of eight centers per grantee and 45 children per center would be available, using NRS data for the evaluation. Table VI.2 shows the number of grantees required for the evaluation in order to detect impacts with a minimum effect size between 0.1 and 0.2 (that is, the impact as a proportion of the standard deviation of the outcome measure). To detect an effect size of at least 0.15, the evaluation would require either at least 88 grantees in the enhancement group and 88 in the control group (if no baseline data are available) or 70 grantees in each group (if some baseline data are available). Because NRS data are available on all children in the program, the number of centers that would be included in the evaluation is 560 per group (if 70 grantees are used and if each grantee has an average of eight centers) and 25,200 children per group (if each center has an average of 45 children in the follow-up sample).

Table VI.2. Sample Sizes for Grantees, Centers, and Children per Evaluation Group, Stage 3 Evaluation with Random Assignment of Grantees and Using NRS Data on All Children in All Centers

	Total Number of Grantees per Evaluation Group	Total Number of Centers per Evaluation Group	Number of Children in Final Sample per Evaluation Group
MDE = 0.10			
No baseline	197	1,576	70,920
Minimal baseline	158	1,264	56,880
MDE = 0.15			
No baseline	88	704	31,680
Minimal baseline	70	560	25,200
MDE = 0.20			
No baseline	49	392	17,640
Minimal baseline	39	312	14,040

Note: Sample size calculations assume that grantees are randomly assigned, all centers are included in the evaluation, each grantee includes 8 centers (on average), and each center includes 45 children (on average) in the final sample.

"Minimal baseline" means that demographic information and NRS fall test scores are available so that the R^2 for the regression adjustment of the impact estimates is .20.

Sample size calculations also assume a two-tailed test with 80 percent power and a 95 percent confidence level. Appendix B provides details about the calculations.

Sample size calculations assume that the sampling strategy used minimizes the design effect of weighting for the source of data. Thus, if the evaluation is based on NRS data on outcomes for all children in the grantee, sampling of grantees would be based on equal probability of selection for all grantees stratified by size. The sample size calculations do not include an adjustment for the design effect of weighting for sample nonresponse.

MDE = minimum detectable effect; NRS = National Reporting System.

The NRS currently provides only a limited set of outcome measures and no information on the quality of implementation. Consequently, a broader data collection effort that includes measures of children's social and emotional well-being as well as the quality of classroom environments and fidelity of implementation are important for understanding why the enhancement is or is not successful. To collect additional data will require drawing a more efficient sample from the participating grantees. Head Start classrooms average 17 children. A sample of two or three centers per grantee and 10 children per center would provide a reasonable basis for estimates of the effects of the enhancement in each grantee.

The optimal strategy for sampling grantees is different depending on whether the evaluation is based on NRS data on all children in the grantee or broader data on a sample of centers and children in the grantee. If NRS data are used, the sampling strategy that minimizes the design effect of weighting is equal probability sampling of grantees. If data from a sample of centers and children are used, the sampling strategy that minimizes the design effect of weighting is selecting grantees with probability proportional to size.

To sample centers, grantees and delegate agencies would be contacted for information about the number of centers and, within centers, the number of classes, by age of the children. Sampling would be based on probability proportional to size, with an implicit stratification procedure that would draw from a sorted list of centers, with sorting based on such characteristics as classroom schedule and percentage of non-English-speaking children. Drawing from a sorted list helps the sample to be more proportionally representative of the characteristics used in the sort.

Table VI.3 shows the numbers of grantees, centers, and children included in a sample that includes two centers per grantee and 10 children per center at the final followup. We use slightly larger minimum detectable effect (MDE) sizes in this table (0.15 to 0.25) because data collection is considerably more expensive, but also because the measures used may be more sensitive to the enhancement. Notably, the number of grantees that would be required to attain the same precision level as in the preceding example (.015) is higher (107 per group for this design, compared with 70 per group in the previous design), but the number of centers and children in the evaluation would be much lower. The two designs could be combined by selecting a target MDE level for the in-depth data collection, identifying the target number of grantees, and then using that sample of grantees for the NRS analysis as well, which would have more power to detect impacts because all centers and children would be included in the NRS outcome data. For example, if broader child assessment and classroom observation data are collected, the target MDE size could be set at 0.2, and 60 grantees per group would be required (assuming that the fall NRS data can be used to improve the precision of the impact estimates). If NRS data are obtained for all of these grantees, the precision levels for the impact analyses would be approximately .17 because the sample of children and centers within each grantee would be larger. A sample of centers and children could be drawn from these grantees for more-extensive data collection. Of course, if the two designs are combined, a single grantee sampling strategy would need to be chosen which would strike a compromise between the probability proportional to size and the equal probability sampling strategies. One possibility is to use the square root of the number of

Table VI.3. Sample Sizes for Grantees, Centers, and Children per Evaluation Group, Stage 3 Evaluation with Random Assignment of Grantees and Data Collection in a Sample of Centers

	Total Number of Grantees per Evaluation Group	Total Number of Centers per Evaluation Group	Total Number of Classrooms per Evaluation Group	Number of Children per Evaluation Group in Final Sample	Number of Children per Evaluation Group in Initial Sample
MDE = 0.15					
No baseline	134	268	804	2,680	2,948
Minimal baseline	107	214	642	2,140	2,354
MDE = 0.20					
No baseline	75	150	450	1,500	1,650
Minimal baseline	60	120	360	1,200	1,320
MDE = 0.25					
No baseline	48	96	288	960	1,056
Minimal baseline	39	78	234	780	858

Note: Sample size calculations assume that grantees are randomly assigned, two centers are selected randomly for the evaluation sample, and centers include an average of three classrooms. Within each center, eleven children are randomly selected for the evaluation sample and approximately ten children are available at follow-up.

"Minimal baseline" means that demographic information and NRS fall test scores are available so that the R^2 for the regression adjustment of the impact estimates is .20.

Sample size calculations also assume a two-tailed test with 80 percent power and a 95 percent confidence level. Appendix B provides details about the calculations.

Sample size calculations assume that the sampling strategy used minimizes the design effect of weighting for the source of data. Thus, if the evaluation is based on data from a sample of centers and children, sampling of grantees would be based on probability of selection proportional to size. The sample size calculations do not include an adjustment for the design effect of weighting for sample nonresponse.

MDE = minimum detectable effect; NRS = National Reporting System.

children as the measure of size at each stage of sampling, and then select with probability proportional to this measure of size. This strategy would result in some design effect of weighting in analyses based on both the NRS and the broader data collection, and this would decrease the power of the sample to detect impacts relative to the figures presented in Tables VI.2 and VI.3.

Recruiting Selected Grantees

Because the evaluation is based on the random assignment of grantees, sampling of grantees and random assignment can occur simultaneously. Subsequently, if the evaluation will rely on administrative data to measure outcomes, recruitment can involve sending letters to the enhancement grantees to inform them of the opportunity to work with a T/TA specialist to understand the data they collect, and to implement a continuous program improvement process. An enhancement like this one would be welcomed by grantees not already engaged in such a process. Grantees that are already conducting these activities would receive help to move the process to a higher level.

If more-extensive data will be collected from a sample selected from within the grantee, both enhancement and control grantees will have to be recruited to participate in the study, and to provide information on centers to support sampling and recruiting at that level. Grantees can be recruited to participate in the study using letters that ask them directly to participate according to their random assignment status. Grantees selected to implement the enhancement will be informed about the opportunity to learn about a continuous program improvement process and informed that some centers, classes, and children will be selected to participate in a study. Grantees selected for the control group will be informed that, because the Head Start Bureau is interested in learning more about how Head Start programs operate and about how children are faring, some centers, classes, and children will be selected to participate in a study. The control-group letter would be much like the current invitation to participate in FACES. Currently, program cooperation with the FACES is very high, so we anticipate no significant difficulties in gaining program participants.

After grantees have been selected and have agreed to participate, they will be asked to provide information on the centers, including contact information for the director; center characteristics (full-day or part-day services, and length of program year); and the number of classrooms, teachers, and students per class. Based on this information, a sample of centers can be selected for data collection beyond the NRS. The centers' directors would be contacted to arrange the sampling of classrooms and children for the evaluation, and to obtain teachers' and parents' consent to participate in the study.

2. Implementing the Quality Enhancement

The enhancement would be implemented using the manual developed by Region I, with numerous examples and resources for program staff, as well as on-site assistance from the T/TA system. The manual would provide ideas about measures to use, instructions for administering the measures, and information about how to report the results in light of Head Start Program Performance Standards. T/TA staff could help to orient staff, train the staff how to collect the data, train the staff to score and present results, and work with them to interpret the measures to understand program strengths and weaknesses. T/TA staff could then suggest ideas for program improvement, and program staff would identify high-priority areas. By going through the entire self-evaluation process once with a good set of measures and clear information to guide analysis of the data, program staff can experience how the process works. In the next year, they could either repeat the process with more confidence or modify the process by changing or adding some measures to cover areas in which they would like additional or different types of information.

The amount of time needed for implementation and evaluation of this enhancement is likely to be longer than for other enhancements discussed in this report because one cycle of the information-gathering and program improvement process takes a full program year. Meaningful program changes and effects on children cannot occur until one cycle of self-review has been completed, priorities have been established, and the first set of improvements has been implemented. Two consecutive cohorts of child outcome data collected after the first implementation could be even more useful in demonstrating the

effects of this approach, as successive years of the continuous feedback cycle should yield cumulative program improvements. Moreover, program staff will make the process more informative and useful over time as they become more familiar with it, identify new areas to measure, and otherwise tailor the process to their needs. A longer implementation period will be difficult for the control group, which will want to implement the enhancement more quickly. A timeline of activities associated with evaluating this enhancement based on NRS data is shown in Figure VI.1.

Measuring fidelity of implementation is important so that the evaluation measures the impacts of the enhancement as designed, rather than a pale substitute. Because this enhancement has not emerged from a Stage 1 Development phase, measures of fidelity have not been developed. However, measures of adherence to a program measurement, data analysis, stakeholder consultation, and priority-setting process could be developed with reasonable effort. Such measures should tap key steps in the process as well as the intensity and breadth of the process.

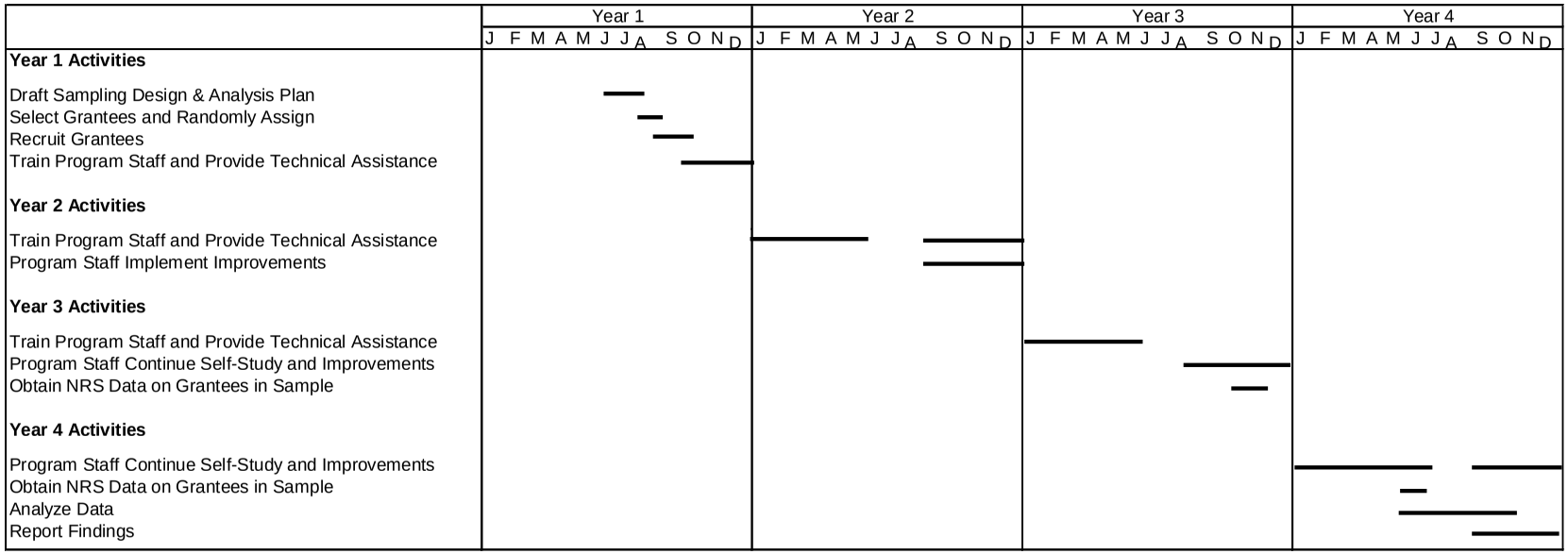
T/TA staff can measure these aspects of implementation as they visit programs to provide technical assistance. Results of fidelity measurement should help them to determine priorities for consultation, training, and assistance. Ratings of fidelity to implementation can also be used in the evaluation, if measurement is comparable across raters. Comparability can be improved if T/TA staff are carefully trained to use the assessment protocol and, where items involve some judgment on the part of the rater, if steps are taken to check the reliability of coding. Reliability can be checked initially and over longer periods by asking T/TA staff to videotape the situation they are coding, and to send the videotape and coding results to the research organization to check. Research staff would then view the videotapes, code independently, compare coding, and provide feedback to the T/TA staff to help improve their understanding of any items for which there was substantial disagreement.

3. Outcomes Measurement and Data Collection Plans

We discuss two options for measuring impacts of the enhancement on children's outcomes. First, the evaluation can rely on data from the NRS. This strategy would provide information on vocabulary, letter recognition, and early mathematics skills for four- and five-year-old children in the program grantees that were randomly assigned. Alternatively, the evaluation can include data collected from a sample of centers and children in the participating program grantees.

One challenge facing any measurement strategy for this enhancement is the potential breadth of the strategy's effects. The continuous program improvement feedback loop will highlight different areas for improvement for different programs. For example, one program might decide that staff education levels need the most attention in the coming year, another might implement a mathematics curriculum, and a third might institute a set of interventions to address children's social-emotional development and behavioral issues. Thus, programs might make very different decisions about what steps to take to increase quality, and to improve compliance with the Head Start Program Performance Standards.

Figure VI.1. Schedule of Activities for Field Test of a Continuous Program Improvement Initiative Using NRS Data Three and One-Half Year Study Beginning in June



These changes in different program areas will generate impacts on different aspects of child development. As a result, the average impact of the enhancement on any single area of child development might be quite small, as only a subset of programs might be implementing a program change to improve that area. Although the areas addressed by programs might become more similar over time, during the initial years, we would expect several different areas of focus across the grantees that implement the continuous program improvement approach.

Use of Head Start Administrative Data

NRS data on children's vocabulary, letter recognition, and early mathematics skills can support estimation of the impacts of the enhancement on these outcomes. As discussed in Section A, the NRS measures have good reliability, and they provide a measure of three important areas of children's academic progress in Head Start. The NRS currently does not measure children's social emotional well-being, so any changes in this area would not be measured if the evaluation relied on administrative data.[5] Moreover, the NRS data do not cover three-year-olds in the program, but children in this age group constitute about one-third of all Head Start children.

Information on children, classrooms, teachers, and centers can be taken from the CBRS; for the most part, this information would offer control variables to improve the precision of the impact estimates (see Table VI.1). However, one item (teachers' education levels) could be considered an intermediate outcome of this enhancement, because programs might decide to focus on increasing the proportion of lead teachers with master's degrees as one strategy to improve instructional quality. Grantee-level information on location, size of the program, and aggregate child and family characteristics can be obtained from the PIR and can be used as control variables to improve precision, or to define subgroups for analysis. Information about parents' education levels and learning activities in the home would not be available; nor would any measures of classroom quality.

Broader Data Collection from a Sample of Children Within Grantees

As an option, ACF might decide to collect more-extensive outcome data and information on intermediate outcomes from a sample within the grantees participating in the study. The number of children per center (five) would be much less than under the design based on administrative data. The number of centers per grantee (two) and number of classrooms per center (two) also would be much less than under the administrative-data design. Table VI.3 summarizes the sample sizes required to detect effect sizes (impacts as a percentage of the standard deviation of the outcome measure) of 0.15, 0.20, and 0.25.

Under this option, researchers would supplement the NRS outcomes with intermediate outcomes measuring directors', teachers', and other staff members' knowledge of the

[5] Measures of children's social-emotional well-being are a major area of interest for future NRS assessments. An Advisory Group will consider alternative measures and other issues over the coming months.

process of gathering information, analyzing program strengths and weaknesses, and recommending steps toward improvement, as well as their knowledge of the management climate (see Table VI.4). Additional intermediate outcomes would focus on the classroom, including the overall quality and the quality of instructional practices in reading, language development, and early mathematics. Measures of children's development would be supplemented by measures of phonemic awareness and writing, as well as aspects of social-emotional development, including aggressive behavior, hyperactivity, prosocial behavior, and approaches toward learning.

Expanding data collection beyond Head Start administrative data would increase the costs and complexity of the study beyond the obvious need to develop data collection instruments and then visit centers to collect classroom-level and child-level data. The study design and data collection instruments would have to be approved by both OMB and an Institutional Review Board (IRB); centers, classrooms, and children would have to be sampled; and center staff and parents would have to consent to participate. Because of the need to obtain OMB clearance and to recruit centers, teachers, and families, the timeline for the study would have to be longer as well. We discuss these tasks and their timing below. Figure VI.2 summarizes the activities and their timing if additional data are collected from a sample of centers and children within the grantees.

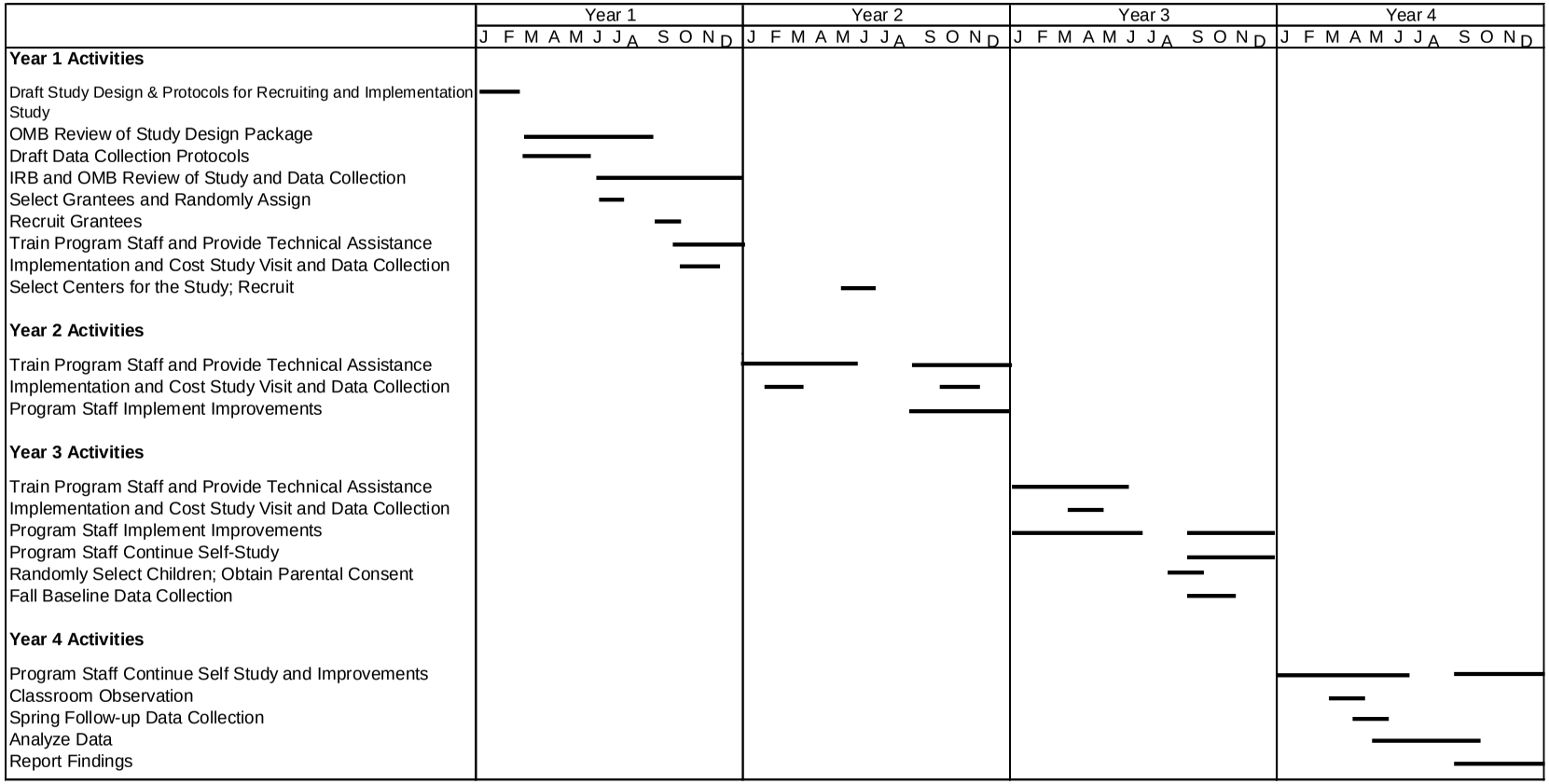
Develop Data Collection Instruments. Data collection instruments, including the consent form with a form requesting demographic information; the director, Head Start staff, and teacher surveys; the child assessment and observation protocol; and the classroom observation protocol; will be developed during the first year of the study. Many of these instruments will include standardized assessments that will have to be formatted to simplify administration by a trained assessor. Others (such as the teacher questionnaire) will rely on questions that have been used in previous studies. After the data collection instruments have been developed, they will be pretested to ensure that respondents understand the questions, that the question flow proceeds logically and smoothly, and that the time required to complete the questions is reasonable.

IRB and OMB Research Review. Research on human subjects must be reviewed by an IRB, which considers the benefits of the research to society, the programs, and the participating families and weighs them against the cost of the research to the families and program staff. The IRB also ensures that research participants are protected from harm by determining that confidentiality is maintained. In addition, if the evaluation is federally funded, data collection instruments and the research plan must be approved by OMB. The data collection instruments are reviewed to ensure that they do not overlap with ongoing federal data collection efforts, and that burden is not excessive. These reviews will be conducted during the first year of the study and will have two parts: the first review will be initiated quickly and will cover the study design, recruiting protocols, and implementation study protocols. The second review will be conducted during the second half of the first year and will pertain to the data collection protocols.

Table VI.4: Measures of Intermediate and Child Outcomes Associated with a Continuous Program Improvement Enhancement

Outcome	Recommended Measure	Type of Measure
Directors' Knowledge and Practice		
Using data to identify program improvements	Questions about key steps in the process; who participated; what recommendations were made; how priorities were set; what changed as a result	Director survey
Management climate	Policy and Program Management Inventory (Lambert, Abbott-Shim, and Oxford-Wright 2004)	Directory survey
Head Start Staff Knowledge and Practice		
Using data to identify program improvements	Questions about key steps in the process; who participated; what recommendations were made; how priorities were set; what changed as a result	Staff survey
Management climate	Policy and Program Management Inventory (Lambert, Abbott-Shim, and Oxford-Wright 2004)	Staff survey
Classroom Processes		
Materials and teacher activities to promote learning	Early Childhood Environment Rating Scale – Revised (Harms et al. 1998)	Observation Teacher survey
	Teacher Behavior Rating Scale (CIRCLE 2005)	Observation
	Classroom Assessment Scoring System (subscales measuring Emotional Climate and Instructional Climate) (CLASS; LaParo, Pianta, and Stuhlman 2004)	Observation
Teachers' Knowledge and Practice		
Attitudes and knowledge about developmentally appropriate practice	Teacher Beliefs Scale (Burts, Hart, Charlesworth and Kirk 1990)	Teacher survey
Using assessment data to individualize instruction	Questions about key steps in the process; what assessments were used; how priorities were set; how instruction is individualized	Teacher survey
Management climate	Policy and Program Management Inventory (Lambert, Abbott-Shim, and Oxford-Wright 2004)	Teacher survey
Children's Development		
Phonemic awareness	Woodcock-Johnson III Sound Awareness	Assessment
Writing, small motor skills	Woodcock-Johnson III Spelling	Assessment
Behavioral problems	Child Behavior Checklist (Achenbach and Rescorla 2000)	Teacher report
Social competence	Social Skills Rating Scale (Gresham and Elliott 1990)	Teacher report
Approaches toward learning	Preschool Learning Behaviors Scale (McDermott et al. 2000)	Teacher report

Figure VI.2. Schedule of Activities for Field Test of a Continuous Program Improvement Initiative with Additional Data Collection
Four-Year Study Beginning in January



Timetable for Implementation. The enhancement will be implemented in the beginning of the Head Start year, nine months after the study begins. Implementation requires working with the program for a full Head Start year through the full cycle of data collection, analysis, identification of strengths and weaknesses, and development of recommendations for program improvements in the coming year. During the second year, programs will go through the self-study cycle again but will also implement some of the recommendations for program improvement. Because the self-study process by itself is not expected to produce measurable improvements in children's well-being, we recommend that data on children's outcomes be collected starting during the middle of the third year, after programs have had a year to implement program improvements based on their self-study process.

Selection of Centers and Children for the Study. When classrooms and children's enrollments are established in August of the third year, researchers will work with program grantee directors to obtain a list of centers, teachers, and the number of children enrolled in each classroom. Two centers will be selected for the research sample. Parent consent forms will be distributed to parents of children in those centers. Returned consent forms and associated information sheets will be sent by program staff to the researchers for processing. The researchers will enter data from the forms to classify children by center, by consent status, and by demographic characteristics. A sample of 10 children from the pool of eligible children in each research center will be randomly selected for the data collection.

Consent. Consent for children to participate in the research must be obtained from parents or guardians. The parent consent form will clearly inform parents (guardians) about the duration of the study, the types of assessments that will be administered, and the voluntary nature of participation. An information sheet will be included in the consent package to collect basic demographic information about the family, such as age, race and ethnicity, and family structure; these variables will be used to improve the precision of impact estimates as well as to define subgroups.

The consent process could proceed more smoothly if it is incorporated into the home visits that many programs make to families. During these visits, which typically occur just before children attend class for the first time, teachers bring forms that parents must complete before the start of the school year. The teacher is available to explain the forms, and to ensure that they are completed correctly. If the study's consent is part of this process, the teacher would be able to explain the nature of the study, what will happen if the child participates in the research, and the voluntary nature of the child's participation.

Child Eligibility Criteria for the Study. Because the sample of children is intended to represent all three- four-, and five-year-old children in each center, all children participating in Head Start should be eligible for the study. Although implementation of the enhancement will occur before the sample of children is drawn, we recommend including all Head Start children in the potential sample for the study, rather than limiting the sample to children new to Head Start during the year in which fall and spring data are collected.

Directors', Staff, and Teachers' Self-Administered Questionnaires. All levels of staff will be asked about the continuous improvement process, how the process is working

to improve quality of the management climate, and their satisfaction with their work. Teachers' attitudes about developmentally appropriate classroom practices will be tapped, as will ways in which the teachers are intentionally instructing children on language arts, early reading, and early mathematics skills. In addition, the questionnaire will include questions about the teachers' background, and about their attitudes and beliefs about developmentally appropriate practice. During the spring data collection, the teacher questionnaire will omit the teacher background questions (except in the case of teachers who are new), but it will include a short behavior problems scale, social skills scale, and approaches toward learning scale for each child in the research sample. The behavior problems scale and the social skills rating scale will provide additional information on children's behavior based on the teachers' observations during the year. The approaches toward learning scale will help identify positive factors, such as curiosity and attentiveness, that are associated with academic success.

Classroom Observation. Measuring the overall quality of the classroom using a commonly used observational protocol will enable researchers to understand how the continuous program improvement process has contributed to the overall quality of the classroom environment. Thus, to measure classroom quality, we recommend the use of the Early Childhood Environment Rating Scale-Revised (Harms et al. 1998). We also recommend including subscales of the Teacher Behavior Rating Scale (CIRCLE 2005) that tap the quantity and quality of instruction in language development, early reading skills, and early mathematics skills. In addition, we recommend using two subscales of the Classroom Assessment Scoring System (CLASS; LaParo et al. 2004): (1) the Instructional Climate subscale, and (2) the Emotional Climate subscale. We recommend that the classroom observations be conducted during the two weeks before the spring follow-up assessments begin.

Child Assessment. The impacts of the quality improvement process on children will depend on what program improvements are implemented. The measurement plan would include measures of domains not included in the NRS; thus, we recommend measures of phonemic awareness, writing, and social skills to supplement NRS measures of language development, letter recognition, and early mathematics skills.

The importance of obtaining a true baseline child assessment, ideally before the children have had any experience in the Head Start classroom, must be balanced against both the need to control data collection costs by assessing children in the Head Start center and the two- to three-week period required to stabilize enrollment in Head Start classes. We recommend a field period that starts approximately two to three weeks after classes begin at the Head Start center, with a six-week window for data collection in the sites. To assess how children's outcomes have been influenced by their experiences in the Head Start classroom and by what they have learned during the Head Start year, the follow-up assessment should be conducted as close to the end of the year as possible. We recommend that the follow-up data be collected during a six-week window that ends two weeks before the end of the year, and that data collection be matched to the timing of the fall assessment so that classes assessed early in the fall field period also are assessed early in the spring field period.

C. APPROACHES TO ENHANCING EARLY LITERACY

Head Start has invested considerable resources into strengthening the support for early literacy instruction in Head Start classrooms. Toward this end, the Head Start Bureau developed the Strategic Teacher Education Program (STEP) to train all Head Start teachers in techniques that support children's emergent literacy. Research suggests that children progress further in reading if they have strong vocabulary skills (so that the words they read make sense) and strong decoding skills (so that they recognize letters, recognize the sounds of individual letters or groups of letters, and are able to combine the sounds to form words) (Whitehurst and Lonigan 1998). The vocabulary skills are referred to as "outside-in" skills because knowledge of the context and vocabulary supports the ability to read specific words, and the decoding skills are referred to as "inside-out" skills because knowledge of sounds and letters is used to read words and sentences.

Based on the reading research, STEP training and the curricula we discuss below include a combined focus on (1) building children's vocabulary and familiarity with children's books; and (2) teaching letter recognition, rhyming, the sounds of words, and the relationship between letters and sounds. Each Head Start grantee identified one staff member to attend intensive STEP training during the summer of 2002. After receiving the training, the 2,500 early literacy specialists returned to their Head Start programs to train teachers locally in preparation for the 2002-2003 program year. A system of mentor-coach training was designed to provide individualized follow-up training for teachers. Early literacy specialists provided one-on-one training to teachers, offering individual assistance to tailor the early literacy teaching techniques to each teacher, and to respond to the diverse backgrounds and needs of children in the classroom. To further reinforce the techniques and provide resource materials, the Head Start Bureau developed a website containing resource materials for teachers.

Given the level of investment in language development and emergent literacy initiatives already made in Head Start classrooms, in-depth examination of alternative approaches in this area might seem unnecessary. However, it is possible that some programs have not fully implemented the STEP techniques, so that following a well-documented, well-implemented approach would make a difference for children. Even if ACF is not interested in alternative literacy curricula at this time, this design offers a template for evaluating competing approaches to enhancing children's development in the same domains as are covered in STEP (language and early literacy). Many of the early literacy curricula are at a stage of development that would enable them to be widely implemented, making them good candidates for a Stage 3 evaluation. This enhancement example is distinct from the previous one because it can be implemented at the center level, and, because we expect little spillover across centers, random assignment can take place at that level. Other curriculum innovations or changes that are implemented primarily at the classroom level, such as a mathematics curriculum or approaches to enhancing children's social-emotional development, would be good examples of enhancements that would fit this design.

Several approaches to enhancing preschool children's language development and early literacy skills have been developed. Some approaches are embedded within a broader preschool curriculum, and some present guides to classroom activities in large or small

groups to promote the desired outcomes; another one is a computer-based approach. Manuals, training guides, and materials have been developed for most of the approaches, and most of them have been implemented in many preschool classrooms. We include these as illustrative examples, although they might not all be ready for Stage 3 evaluation:

- ***Ladders to Literacy.*** This program consists of 60 teacher-led activities organized into three units: (1) print awareness (for example, understanding the parts of a book and reading from left to right), (2) phonological awareness (understanding rhymes, understanding syllables, and linking sounds with letters), and (3) oral language (vocabulary). Activities require varying amounts of preparation, but all of them can be easily included in classroom routines.
- ***Creative Curriculum Approach to Literacy.*** This program shows teachers how to incorporate language development and early literacy activities into the Creative Curriculum, one of two comprehensive preschool curricula commonly used in Head Start. Ideas are provided for activities that fit into all 11 of the curriculum's interest areas, and they can be used throughout the daily program schedule.
- ***Waterford Early Reading Program.*** The first level of this reading program focuses on developing phonological awareness; letter recognition; and understanding of print concepts, such as reading from left to right. Phonological Awareness and Writing offer companion modules for children at the initial levels of the program. This curriculum is computer software-based.
- ***Breakthrough to Literacy.*** This curriculum includes ideas for books to be read aloud and discussed in the classroom; small-group instruction in language, phonological awareness, letter recognition, and words and sentences; language and literacy centers for independent activity; and writing instruction.
- ***Let's Begin with the Letter People.*** This curriculum uses songs, stories, and puppets representing letters in an engaging approach to learning letter recognition and phonological-awareness skills.

Systematically implementing these five curricula in different Head Start centers would provide a test of which ones are more effective for Head Start children, and whether any or all of them improve children's emergent literacy skills relative to current Head Start practice. We describe the study design next.

1. Major Activities and Timetable for the Evaluation

Several activities must occur to carry out the evaluation:

- Draft study design and protocol for recruiting program grantees and centers and submit for OMB review
- Develop data collection instruments and prepare and submit review packages for IRB and OMB review

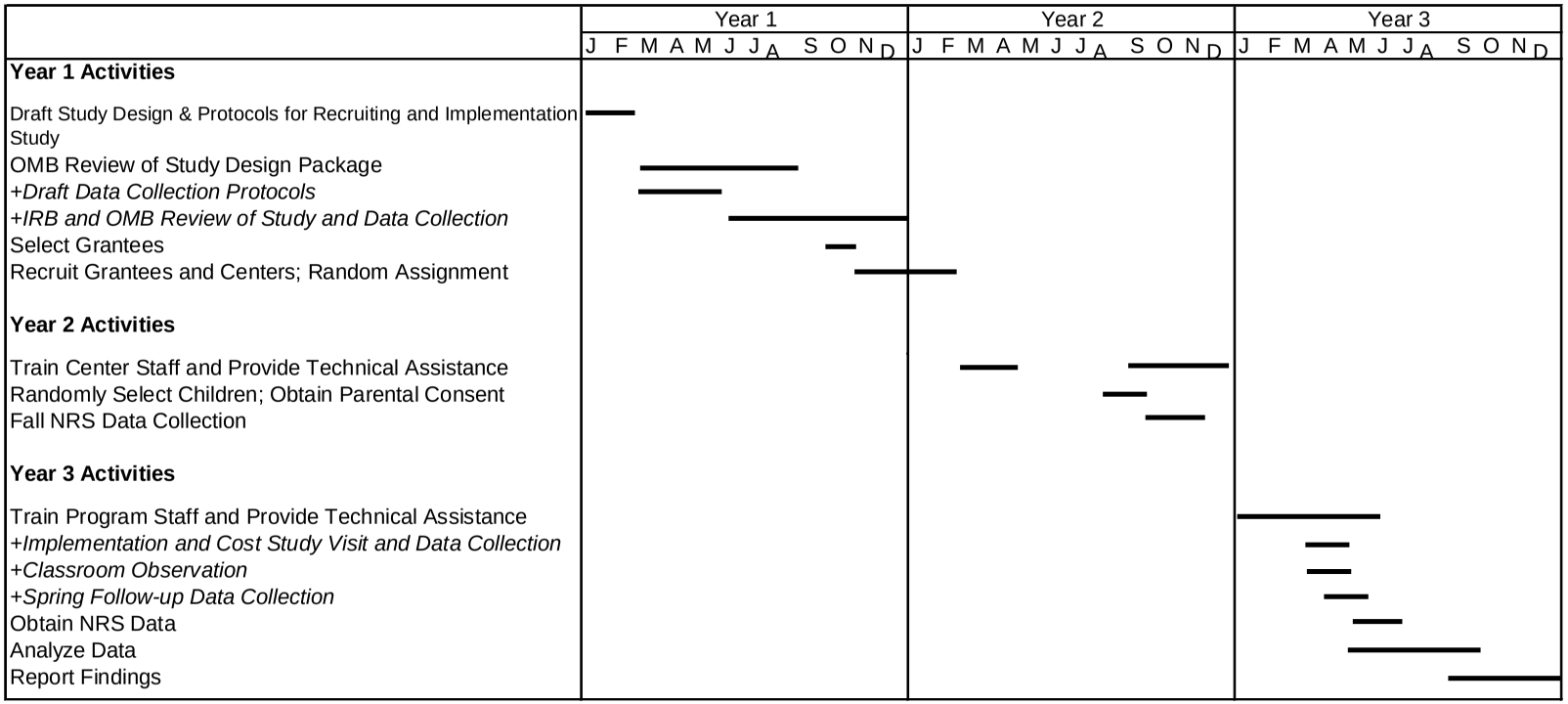
- Select grantees, recruit grantees and centers, and randomly assign centers
- Implement the curriculum in randomly selected centers and conduct implementation and cost study site visits
- Obtain parents' consent and select a sample of children for the research sample
- Collect data from enhancement and control groups during the fall and spring of the Head Start year; observe classrooms during the spring
- Analyze the data and report findings

We recommend that the first three tasks be conducted during the first year of the evaluation, with implementation, sampling of children, and collection of baseline data during the second year, and follow-up data collection, analysis, and reporting during the third year (see Figure VI.3). Because the Head Start year typically runs from August or September through May or June, the timing of activities will proceed most smoothly if the evaluation activities begin in January or February. We discuss the steps in more detail in the rest of this chapter.

The first year of evaluation activities will be dominated by OMB review and by the sampling and recruitment of grantees and centers. OMB review of the study design and protocols for recruiting grantees and centers must be completed before researchers and curriculum developers can discuss the study with grantees, obtain information about prospective centers, or negotiate agreements to participate. We have estimated two months to draft and submit a package to OMB that includes the study design and recruiting protocols, and six months for OMB review. During the OMB review period, researchers can conduct other evaluation-related activities that do not involve data collection from prospective grantees and centers. For example, data collection protocols can be developed and submitted for OMB review, and the study design and data collection plans can be submitted for IRB review. Data systems for tracking children and for managing evaluation data can be developed as well. In addition, ACF can inform the Head Start community about the pending evaluation. After receiving OMB clearance, grantees can be sampled, and the curriculum developer and researchers can contact the selected grantees and centers and can begin the recruiting process. We estimate that recruitment, executing agreements, and randomly assigning centers or classrooms will require approximately five months.

The main evaluation activities during the second year consist of implementing the curricula to high fidelity, conducting the implementation study, obtaining consent for children to participate in the study, sampling children, and collecting fall baseline data. Ideally, implementation will occur during the spring so that the Head Start classes can benefit from a fully implemented curriculum during the entire Head Start year that follows. Third-year evaluation activities will include collecting follow-up data in the spring on classrooms and children, analyzing the data, and reporting the results.

Figure VI.3. Schedule of Activities for Field Test of Alternative Early Literacy Curricula
Three-Year Study Beginning in January



+ = Additional tasks associated with the design option that includes data collection in the spring of the Head Start year.

2. Study Design

An evaluation of the literacy curricula described here would be based on random assignment of centers and a comparison of children's progress toward language development and emergent literacy skills. This section discusses key elements of a research design to evaluate the impacts of this enhancement on children's outcomes. We begin by discussing the quality enhancement and its counterfactual, the research questions, sampling strategy, random assignment plan, and sample sizes.

The Quality Enhancement, Counterfactual, and Research Questions

The quality enhancement to be evaluated is a curriculum to support children's language development and the acquisition of early literacy skills. ACF could decide to study one curriculum alone or multiple curricula. If more than one curriculum is studied, the outcomes of children using the different curricula can be compared to test whether any of the curricula are more effective than the others. Because implementing each curriculum is likely to increase the focus on language development and early literacy skills, it is likely that the differences in outcomes between children using the alternative curricula will be small. However, the evaluation still can establish whether any (or each) of the curricula is more effective than current practice. For this reason, under any evaluation design, it would be useful to include a control group that does not implement any of the curricula in order to permit comparisons of children under each curriculum with those not implementing any of the curricula.

The Head Start STEP training provided a foundation for language and literacy activities in each classroom. We have no information on how well STEP training supported better language development and early literacy activities in the classroom. Some information on the quality of implementation of the STEP approaches could be obtained from measures of language and literacy activities in the control-group classrooms in an evaluation of alternative curricula.

Research questions to be examined as part of this evaluation include whether each curriculum was implemented to high fidelity:

- Did staff implement the curriculum fully? Did they implement it with a high degree of fidelity?
- What strategies were used to implement the curriculum? How much training was provided, and over what time period? What amount and types of technical assistance were provided, and over what time period? What were the education levels and experience of T/TA staff?
- What challenges were encountered in implementing the curriculum, and how were they resolved?

- To what extent are control-group classrooms providing language environments and early literacy skills instruction similar to those provided by classrooms using the enhancement curricula?

The most important impacts of the language and literacy curricula are those pertaining to the quality of the classroom language and literacy environment, and to children's language development and emergent literacy skills. Thus, the following research questions are critical ones:

- Does the use of a language/literacy curriculum increase the amount of time and quality of related activities, such as reading books, introducing new words, discussing the content of books, and engaging in activities to promote letter recognition and phonemic awareness?
- Does the use of a language/literacy curriculum reduce the amount of time spent in play, on gross motor activities, and on cooperative activities?
- Do children in centers using a language/literacy curriculum progress further in vocabulary, book knowledge, and early literacy skills, such as letter recognition and phonological awareness, than those in control-group programs?
- Do children in centers using a language/literacy curriculum make less progress in social-emotional development, including increases in cooperative behavior, initiative, persistence, and self-control (relative to children in centers without that curriculum)?

Finally, although the evaluation will examine the effectiveness of the curricula for children overall, the Head Start community also will be interested in whether the curricula are effective across different subgroups of children and families and across programs with different characteristics:

- What were the impacts of the language/literacy curriculum on outcomes for key subgroups of children? What were the outcomes for Head Start programs with different characteristics?

Some caution must be used when examining subgroup impacts, because, if many subgroups are examined, some impacts will emerge simply by chance. To guard against finding impacts by chance, researchers can implement a Bonferroni correction that essentially inflates the standard error of the impact estimates to correspond to the number of hypothesis tests. This method provides a higher bar for deciding that a subgroup impact is significantly different from zero.

In a Stage 3 design that is nationally or regionally representative of all Head Start programs, subgroup analyses should be based on a representative sample of that subgroup in Head Start. Thus, subgroups of interest must be identified in advance so that statisticians can design a sampling plan for grantees, centers, and children that ensures adequate representative samples of the subgroups of interest. Perhaps the most interesting subgroups in a curriculum study are program schedule (part-day versus full-day), because of the tradeoff

between program costs and outcomes; race/ethnicity, because of concerns about gaps in educational progress; and home language, because a curriculum may or may not work well for children whose home language is not English.

Sampling, Random Assignment, and Sample Sizes

Alternative curricula would be most efficiently and effectively implemented at the center level. Implementing the curricula center-wide would enable teachers within each center to discuss how to approach various activities, trade ideas for books to read to the class, and share other good practices. It is unlikely that spillover would occur across centers, as the curricula involve classroom-based activities. However, some communication across centers does occur, and if this could lead to some spillover, it needs to be monitored and measured.

Because implementation would take place at the center level, random assignment to the enhancement or control group should be at the center level as well. However, researchers would have to draw a first-stage sample of program grantees, delegate agencies, and then sample centers within these agencies. The sample of program grantee or delegate agencies would be drawn within strata defined by Census regions, urbanicity, and percentage of children who are African-American and percentage who are Hispanic or Latino (large versus small percentage). Selection would be based on probability proportional to the number of children in the program, using the most recent PIR data as the sampling frame. The PIR contains data from annual reports submitted by all program grantees and delegate agencies about program size, schedule, and other characteristics.

NRS data are available for all four- and five-year-old children in every center within every grantee. If these data can be obtained and analyzed for the sample of centers included in the evaluation, evaluation costs could be reduced. In the previous example, in which the enhancement is implemented at the grantee level, the availability of NRS data on every child in every center leads to a strategy of minimizing the number of grantees included in the evaluation by including in the evaluation sample every center from the selected grantees. However, in the example of a curriculum implemented at the center level, it makes sense to minimize the number of centers included in the evaluation even if NRS data are available on all children, as implementation requires additional resources per center. Compared to using four centers per group from each grantee, if we randomly assign two centers per grantee to each curriculum group or to the control group, we will slightly increase the number of grantees to be recruited, but substantially decrease the number of centers that would have to implement the curriculum as part of the evaluation. Because many grantees have eight or more centers, assigning two centers from each grantee to each evaluation group also enables the evaluation to expand to include more than one or two curricula, if that is desirable.

We assume that, with NRS data used for the evaluation, an average of two centers per evaluation group from each grantee would produce outcome data on an average of 45 children per center. The sampling strategy that minimizes the design effect of weighting is equal probability sampling of centers. Table VI.5 shows the numbers of grantees and centers per evaluation group that would be necessary to detect impacts with a minimum

Table VI.5. Sample Sizes for Grantees, Centers, and Children per Evaluation Group, Stage 3 Evaluation with Random Assignment of Centers and Using NRS Data on All Children

	Total Number of Grantees per Evaluation Group	Total Number of Centers per Evaluation Group	Number of Children per Evaluation Group in Final Sample
MDE = 0.10			
No baseline	232	464	20,880
Minimal baseline	185	370	16,650
MDE = 0.15			
No baseline	103	206	9,270
Minimal baseline	82	164	7,380
MDE = 0.20			
No baseline	58	116	5,220
Minimal baseline	46	92	4,140

Note: Sample size calculations assume that grantees are randomly selected and two centers per grantee are randomly assigned to each evaluation group. Within each center, 45 children on average are included in the follow-up data collection.

"Minimal baseline" means that demographic information and NRS fall test scores are available so that the R^2 for the regression adjustment of the impact estimates is .20.

Sample size calculations also assume a two-tailed test with 80 percent power and a 95 percent confidence level. Appendix B provides details about the calculations.

Sample size calculations assume that the sampling strategy minimizes the design effect of weighting for the source of data. Thus, if the evaluation is based on NRS data on outcomes for all children in the center, sampling of grantees would be based on probability proportional to size, but sampling of centers within grantees would be based on equal probability of selection for centers stratified by size. The sample size calculations do not include an adjustment for the design effect of weighting for sample nonresponse.

MDE = minimum detectable effect; NRS = National Reporting System.

effect size between 0.1 and 0.2 (that is, the impact as a proportion of the standard deviation of the outcome measure). We present sample size requirements for an effect size of 0.1 because the NRS measures are expected to be less sensitive to intervention and more "noisy" (to have higher measurement variance) than measures we would use as direct assessments. However, to obtain a precision level of 0.10, sample sizes must be very large. To detect an effect size of at least 0.15, the evaluation would require at least 206 grantees in each group (if no baseline data are available) or 164 grantees in each group (if some baseline data are available). Because NRS data are available on all four- and five-year-old children in the program, 164 centers per group would be included in the evaluation (if 82 grantees are used and if each grantee contributes an average of 2 centers to each evaluation group), with 7,380 children per group (if each center has an average of 45 children in the follow-up sample).

Because the NRS currently provides only a limited set of outcome measures and no information on the quality of implementation, a broader data collection effort that includes measures of children's expressive language ability, phonemic awareness, and social-emotional well-being would provide a stronger basis for drawing conclusions about whether the curricula have important impacts on children's development. Measures of the quality of

classroom environments, quantity and quality of language development, and quantity and quality of early literacy activities in the classroom are critical for understanding whether the curricula have impacts on classroom learning environments.

Collecting additional data will require drawing a more efficient sample of children from the participating grantees and centers. To sample centers, grantees and delegate agencies would be asked to provide lists of centers containing information about enrollments, by age of child; the number of classes and class sizes; and directors' contact information. Sampling would again be based on probability proportional to the number of children in the center and on an implicit stratification procedure that would draw from a sorted list of centers, where sorting is based on characteristics, such as classroom schedule, size, and percentage of non-English-speaking children. Drawing from a sorted list helps obtain a sample that is more proportionally representative of the characteristics used in the sort.

Table VI.6 shows the number of grantees, centers, and children to be included in a sample consisting of two centers per group from each grantee, two classrooms per center, and six children per class at the final followup. We use slightly larger minimum detectable effect sizes in this table (0.15 to 0.25) because data collection is considerably more expensive and the measures used likely to be more sensitive to the enhancement.than in the previous example. Notably, the number of grantees that would be required to attain the same precision level as in the previous example (.015) would be higher (105 per group for this design, compared with 82 per group for the previous design), but the number of children in the evaluation would be much lower. The strategy of increasing the number of grantees while reducing the number of children in the sample is efficient because assessing children is very expensive. The tradeoff produces a much lower sample size of children for a relatively modest increase in the number of grantees to be recruited. The increase in the number of grantees reduces the effect of clustering in the sample of children. The two designs could be combined by selecting a target MDE level for the in-depth data collection, identifying the target number of grantees and centers, and then using the sample of grantees and centers for the NRS analysis as well (in other words, using all the children in the centers, rather than only the children sampled for broader data collection). A sample of children could then be drawn from the centers for more-extensive data collection. Of course, sampling of centers for this design would need to be based on a compromise between probability proportional to size and equal probability sampling. One possible compromise is to sample based on probability proportional to size, but to use the square root of the sample size as the measure of size.

Recruiting Selected Grantees and Encouraging Centers to Participate

The selected grantees will be directly recruited to participate in the evaluation. However, given that the grantees' willingness to participate will likely depend on the willingness of the centers' directors and staff to become involved, meetings should include the directors and key staff fairly early in the process. Researchers should call each grantee to describe the study, and to explain its importance. With support from the Head Start Bureau and regional offices, researchers should then schedule a meeting with the grantee executive

Table VI.6. Sample Sizes for Grantees, Centers, Classrooms, and Children per Evaluation Group, Stage 3 Evaluation with Random Assignment and Data Collection in a Sample of Centers

	Total Number of Grantees per Evaluation Group	Total Number of Centers per Evaluation Group	Total Number of Classrooms per Evaluation Group	Number of Children per Evaluation Group in Final Sample	Number of Children per Evaluation Group in Initial Sample
MDE = 0.15					
No baseline	131	262	524	3,144	3,668
Minimal baseline	105	210	420	2,520	2,940
MDE = 0.20					
No baseline	74	148	296	1,776	2,072
Minimal baseline	59	118	236	1,416	1,652
MDE = 0.25					
No baseline	47	94	188	1,128	1,316
Minimal baseline	38	76	152	912	1,064

Note: Sample size calculations assume that grantees are randomly selected for the evaluation and two centers from each grantee are randomly assigned to each evaluation group. Within each center, two classrooms and seven children per classroom are randomly selected for the evaluation sample (six per classroom, on average, are available for the follow-up).

"Minimal baseline" means that demographic information and NRS fall test scores are available so that the R^2 for the regression adjustment of the impact estimates is .20.

Sample size calculations also assume a two-tailed test with 80 percent power and a 95 percent confidence level. Appendix B provides details about the calculations.

Sample size calculations assume that the sampling strategy used minimizes the design effect of weighting for the source of data. Thus, if the evaluation is based on data from a sample of classrooms and children, sampling of grantees and centers would be based on probability of selection proportional to size. The sample size calculations do not include an adjustment for the design effect of weighting for sample nonresponse.

MDE = minimum detectable effect; NRS = National Reporting System.

staff and with as many center directors as possible to discuss in depth the goals of the study, the way that the study would be conducted in the centers, and the benefits and costs of participating. Ideally, the curriculum developers would accompany the researchers to the meeting to help generate excitement about the enhancement options.

The initial contact materials should briefly explain the study's focus (in this case, the early literacy curricula) and should briefly summarize the benefits of participation. The contact materials also might indicate that all centers in the program will have an equal chance of participating in the study, but that only some will be chosen to implement the curricula. Those not chosen to implement the curricula during the first year will be given priority to implement it, if desired, after the follow-up data have been collected. The materials also will explain that the programs will receive information about the study's findings, and that they will be partners in the research study.

Recruitment of grantees and centers will be easier if their staff believe that they will gain more from participation than they might lose. Accordingly, after the initial contact has been made, the benefits to sites must be explained in detail, and concerns about study burden

must be discussed. Curriculum developers and research staff should visit program directors, center directors, and other relevant administrative staff to discuss the curricula, their expected benefits to children, what will be involved in implementing it, and the research aspects of the evaluation. Researchers also will explain the program's role in ensuring that the study yields useful information about the curricula's effectiveness. For example, researchers will have to work with program staff to obtain information required to implement random assignment (for example, the number of classes in each center, teachers' names, class sizes, and children's ages). Program staff will have to maintain random assignment statuses (for example, by ensuring that teachers in the intervention group do not share information about the curricula with control-group teachers). Researchers will have to monitor the integrity of random assignment over time, a key piece of information to demonstrate the reliability of the study. They also will have to work with program staff to schedule child assessments and classroom observations. These evaluation-related requirements will be balanced by the opportunity to implement new curricula that could benefit children. Benefits to the control group are more challenging to identify, but an important benefit that could be offered for control-group participants is first priority to implement a curriculum after the children participating in the evaluation have finished their Head Start year.

Agreements by grantee and center directors to participate should include a memorandum of understanding that describes the benefits to the program and center, and that specifies the respective responsibilities of the curriculum developer, researchers, and Head Start staff in this joint research undertaking. This level of detail ensures that misunderstandings that have the potential to threaten the success of the research study do not arise after it is too late to resolve them. A memorandum of understanding also offers a useful way of informing new staff about the study.

Gaining cooperation of grantee and center directors over time can become easier as the Head Start community gains experience with the process of evaluating quality enhancements. Head Start directors have had positive experiences participating in FACES, the QRC Consortium, and the Early Head Start evaluation, and these positive experiences have made it easier for researchers to recruit programs to participate in subsequent rounds of FACES and in the NRS Quality Assurance study. Publishing results of the quality enhancement studies in outlets that are read by Head Start staff and presenting findings at professional meetings attended by Head Start staff (such as the National Head Start Association and National Association for the Education of Young Children) as well as at regional gatherings of Head Start staff would help to generate excitement, and to foster support of future research efforts.

After the selected grantee directors have agreed to participate, they will be asked to provide information about the centers, including contact information for the directors; the centers' characteristics (full-day or part-day services, and program year); and the number of classrooms, teachers, and students per class. Based on this information, a sample of centers will be selected and randomly assigned to one of the curriculum groups or to the control group. Directors of those centers would then be contacted to discuss plans for

implementing curricula, obtaining teachers' and parents' consent, and sampling classrooms and children for the evaluation.

3. Implementing the Quality Enhancement and Measuring Fidelity

The goal of implementation is to ensure that each curriculum is implemented to high fidelity in the classrooms of the designated Head Start centers. Clearly, high-fidelity implementation at Stage 3 must not require enormous resources; if it does, the curriculum would not be a good candidate for implementation in a large national program such as Head Start. Nevertheless, the task of implementing each curriculum should have strong leadership and the resources and resolve to reach high fidelity. T/TA staff periodically should measure teachers' practices and classroom environments, using the fidelity measure developed for that curriculum, and should use the measures as the basis for conferring with teachers about successful practices and about how to address weak areas.

Each curriculum will be implemented using the relevant curriculum manual, necessary materials, and T/TA staff who are trained to help teachers implement the curriculum to high fidelity. The manual should provide ideas for arranging materials in the classroom, leading and supporting large- and small-group activities, and individualizing instruction for children having difficulty with any parts of the curriculum. T/TA staff will provide initial training in a workshop format during the spring of the second evaluation year (see Figure VI.3) and over the next two months will regularly observe and measure performance as part of technical assistance visits to improve fidelity. When the Head Start program begins again in the fall, new teachers will receive the initial workshop training while experienced teachers will receive a shorter refresher. T/TA staff will then make periodic visits to centers during the fall and spring, with more frequent visits to teachers having more difficulty reaching or maintaining fidelity to the curriculum.,

Measuring fidelity of implementation is important so that the evaluation measures the impacts of the enhancement as designed rather than a pale substitute. One option for obtaining fidelity measures is to obtain the regular measures taken by T/TA staff as they conduct visits to work with teachers on implementation. If measurement is comparable across raters, these ratings of fidelity to implementation can be used in the evaluation. Comparability can be improved if T/TA staff are carefully trained to use the assessment protocol and, where items involve some judgment on the part of the rater, steps are taken to check the reliability of coding. Reliability can be checked initially and over time by asking T/TA staff to videotape the situations they are coding, and to send the videotapes and coding results to the research organization to check. Research staff will view the videotapes, independently code the results, compare coding, and provide feedback to the T/TA staff to help improve their understanding of any items for which the coding was discrepant.

Concerns that the validity of fidelity measures obtained for T/TA purposes could be compromised if they also are used for research may lead to an alternative strategy. A trained cadre of observers could visit each center once during the Head Start year to obtain fidelity measures. Visiting such a large number of centers and classrooms (assuming three or more curricula are being evaluated) will be fairly expensive. Nevertheless, if policymakers want

reliable, valid measures of implementation fidelity so that the results of the evaluation can be trusted as measures of the effectiveness of each curriculum in Head Start, one visit to each center to collect implementation fidelity measures should be strongly considered. If this visit is scheduled for the spring of the evaluation year, the observers also could be trained to collect a set of child outcome measures that reflects the breadth of the early literacy curricula beyond the areas currently covered by the NRS.

4. Outcomes Measurement and Data Collection Plans

We discuss two options for measuring impacts of the literacy curricula on children's outcomes. Under one option, the evaluation could rely on data from the NRS. This strategy would provide information on vocabulary, letter recognition, and early mathematics skills for all children in the centers that were randomly assigned.[6] Under the second, the evaluation could include data collected from a sample of children in the participating centers.

Use of Head Start Administrative Data

NRS data on children's vocabulary and letter recognition can support estimation of the impacts of the literacy curricula on those outcomes. As discussed in Section A, the NRS measures have good reliability, and they provide a measure of important areas of children's progress in Head Start. However, the literacy curricula will include a focus on phonological awareness, which is not measured fully by the letter recognition test, and on vocabulary, which could be more fully measured by adding an assessment of expressive vocabulary. The NRS currently does not measure children's social-emotional well-being, so any changes in this area would not be measured if the evaluation were to rely on administrative data.[7]

Information on children, classrooms, teachers, and centers can be taken from the CBRS; for the most part, this information would offer control variables to improve the precision of the impact estimates or information required to define subgroups (see Table VI.1). Grantee-level information on location, program size, and aggregate child and family characteristics could be obtained from the PIR and used as control variables to improve precision. Information on the fidelity of implementation obtained from the T/TA staff could be used to distinguish classrooms that reached a threshold for high-fidelity implementation from classrooms that did not reach the threshold.

Broader Data Collection from a Sample of Children Within Grantees

As an option, ACF might decide to collect more-extensive outcome data and information on intermediate outcomes from a sample of children in the centers participating in the study. The number of classrooms per center (two) and children per classroom (six)

[6] Using NRS data for an evaluation in this way would require approval of the Head Start Bureau.

[7] Measurement of children's social-emotional well-being is a key area of interest for future NRS assessments. An Advisory Group will consider alternative measures and other issues over the coming months.

would be much lower than under the design based on administrative data (shown in Table VI.5 and VI.6). Under this option, the NRS outcomes would be supplemented with intermediate outcomes measuring the overall quality of the classroom and the quality of instructional practices in language development and early literacy (see Table VI.7). Measures of children's development would be supplemented by measures of expressive vocabulary, phonemic awareness, and writing, as well as aspects of social-emotional development, including aggressive behavior, hyperactivity, prosocial behavior, and approaches toward learning. Because of the high cost of data collection when several curricula are evaluated, we recommend a single visit during the spring of the Head Start year to collect only follow-up child outcome data and classroom observation data. NRS data can serve as the baseline for all children in the sample.

Expanding data collection beyond Head Start administrative data would increase the costs and complexity of the study beyond the obvious need to develop data collection instruments and to then visit centers to collect classroom-level and child-level data. The study design and data collection instruments would have to be approved by OMB and an IRB; classrooms and children would have to be sampled; and center staff and parents would have to consent to participate. Here, we discuss the steps necessary to collect additional data beyond the NRS for this evaluation.

Develop Data Collection Instruments. Data collection instruments will be developed during the first year of the study. They will include the consent form with a form requesting demographic information for parents; the director, Head Start staff, and teacher surveys; the child assessment and observation protocol; and the classroom observation protocol. Many of these instruments will include standardized assessments that will have to be formatted to simplify administration by a trained assessor. Others (such as the teacher questionnaire) will rely on questions that have been used in previous studies. After the data collection instruments have been developed, they will be pretested to ensure that respondents understand the questions, that the question flow proceeds logically and smoothly, and that the time required to complete them is reasonable.

IRB and OMB Research Review. Research on human subjects must be reviewed by an IRB, which considers the benefits of the research to society, the programs, and the participating families and weighs them against the cost of the research to the families and program staff. The IRB also reviews protection of research participants from harm by ensuring that confidentiality is maintained. In addition, if the evaluation is federally funded, data collection instruments and the research plan must be approved by OMB. The data collection instruments are reviewed to ensure that they do not overlap with other federal data collection efforts, and that burden is not excessive. These reviews will be conducted during the first year of the study, in two parts: The first review will be initiated quickly and will cover the study design, recruiting protocols, and implementation study protocols. The second review will be conducted during the second half of the first year and will pertain to the data collection protocols.

Table VI.7: Measures of Intermediate and Child Outcomes Associated with Alternative Language/Literacy Curricula

Outcome	Recommended Measure	Type of Measure
Teachers' Knowledge and Practice		
Attitudes and knowledge about developmentally appropriate practice	Teacher Beliefs Scale (Burts, Hart, Charlesworth and Kirk 1990)	Teacher survey
Curriculum and support for language development and early literacy activities in the classroom	Questions about curricula used in the classroom; instruction on the curricula; materials available to support language development, book reading, and early literacy activities	Teacher survey
Daily activities, language activities, and early literacy activities	Questions about daily schedule and mix of language and literacy-oriented activities	Teacher survey
Classroom Processes		
Materials and teacher activities to promote learning	Early Childhood Environment Rating Scale – Revised (Harms et al. 1998)	Observation Teacher survey
	Teacher Behavior Rating Scale (CIRCLE 2005)	Observation
Children's Development		
Expressive language	Preschool Language Scale, 4th Edition (PLS-4; Zimmerman et al. 2002) or Oral and Written Language Scales (OWLS; Carrow-Woolfolk 1995)	Assessment
Phonemic awareness	Woodcock-Johnson III Sound Awareness (Woodcock, McGrew, and Mather 2001)	Assessment
Letter Recognition	Woodcock-Johnson III Letter-Word Identification (Woodcock, McGrew, and Mather 2001)	Assessment
Writing, small motor skills	Woodcock-Johnson III Spelling (Woodcock, McGrew, and Mather 2001)	Assessment
Behavioral problems	Child Behavior Checklist (Achenbach and Rescorla 2000)	Teacher report
Social competence	Social Skills Rating Scale (Gresham and Elliott 1990)	Teacher report
Approaches toward learning	Preschool Learning Behaviors Scale (McDermott et al. 2000)	Teacher report

Selection of Classrooms and Children for the Study. After classrooms and children's enrollments have been established (during August of the third year of the study), researchers will work with center directors to obtain a list of teachers and the number of children enrolled in each classroom. Two classrooms from each center will be selected with probability of selection proportional to class size. Parent consent forms will be distributed to parents of children in those classrooms (as detailed in the next paragraph). Returned consent forms and associated information sheets will be sent by program staff to the researchers for processing. The researchers will enter data from the forms to classify children by center, by classroom, by consent status, and by demographic characteristics. A sample of seven children from the pool of eligible children in each research classroom will be randomly selected for the data collection.

Consent. Consent for children to participate in the research must be obtained from parents or guardians. The parent consent form will clearly inform parents (guardians) about the duration of the study, the types of assessments that will be administered, and the voluntary nature of participation. An information sheet will be included in the consent package to collect basic demographic information about the family, such as age, race and ethnicity, and family structure, as these variables are necessary to improve the precision of impact estimates as well as to define subgroups.

The consent process could proceed more smoothly if it is incorporated into the home visits that many programs make to families. During these visits, which typically occur just before children attend class for the first time, teachers bring forms that parents must complete before the start of the school year. The teacher is available to explain the forms, and to ensure that they are completed correctly. If the study's consent is part of this process, the teacher would be able to explain the nature of the study, what will happen if the child participates in the research, and the voluntary nature of the child's participation.

Child Eligibility Criteria for the Study. Because children's language development and early literacy skills grow rapidly between three and five years of age, we probably would have to analyze three- and four-year-olds separately, thus reducing the effective analysis sample. We therefore recommend focusing the research sample on children who will be four years old in the fall of the Head Start year. Although teacher training will begin during the final months of the previous Head Start year, we expect that children attending Head Start that year will have had very little exposure to a fully implemented enhancement. Therefore, we recommend including all four-year-old Head Start children in the potential sample for the study, rather than limiting the sample to children who will be new to Head Start during the year in which fall and spring data are collected.

Teachers' Self-Administered Questionnaires. Teachers will be asked about the curricula used in the center and the support they receive for language and early literacy activities in the classroom through purchase of books and other materials and training. In addition, teachers' attitudes about developmentally appropriate classroom practices will be tapped, as will ways in which they are intentionally instructing children in vocabulary, letter recognition, and phonological awareness. Teachers also will complete questionnaires about every research-sample child from their classes ; the questionnaires will include a short

behavior problems scale, a social skills scale, and a scale measuring approaches toward learning. The behavior problems scale and the social skills rating scale will provide information on children's behavior based on the teachers' observations during the year. The scale measuring approaches toward learning will help identify positive factors, such as curiosity and attentiveness, that are associated with academic success.

Classroom Observation. Including a commonly-used observational protocol to measure the overall quality of the classroom will enable researchers to understand how the early literacy curriculum may have influenced the overall quality of the classroom environment. Therefore, we recommend using the Early Childhood Environment Rating Scale-Revised to measure classroom quality (Harms and Clifford 1998). We also recommend including subscales of the Teacher Behavior Rating Scale (CIRCLE 2005) that tap the quantity and quality of instruction in language development and early literacy skills.

Child Assessment. The early literacy curriculum is expected to support improvements in children's vocabulary, letter recognition, phonological knowledge, and writing ability. We recommend measuring these outcomes with well-established scales that have been used in populations similar to the children in the proposed enhancements. For expressive language ability, we recommend using the Preschool Language Scale (Zimmerman et al. 2002) or the Oral and Written Language Scales (Carrow-Woolfolk 1995). For phonological awareness, we recommend the Woodcock-Johnson III Sound Awareness task; for letter recognition, the Woodcock-Johnson III Letter-Word Identification task; and for writing ability, the Woodcock-Johnson III Spelling task (Woodcock et al. 2001).

D. ANALYSIS AND REPORTS

For any of the field test designs described in this chapter, the evaluation should culminate in a report that clearly describes the enhancement, the research design, and the estimated impacts of the enhancement on key outcomes. This report should be written for a broad audience. In addition, other reports and papers can be written to address other information needs:

- ***Implementation.*** This report would describe how the enhancement was implemented and the degree of fidelity to implementation.
- ***Impacts.*** A longer version of the report described above would address a technical audience and would describe the research design, sample characteristics, analytic approaches, and findings with sufficient detail for other researchers to evaluate the quality of the findings
- ***Cost-Effectiveness.*** This report would present estimates of the costs of the enhancement, describe how those cost estimates were obtained, and compare them with impacts on the outcomes most closely related to the enhancement, presented in effect size units.

Detailed and summary versions of each report should be written to address the needs of technical audiences as well as the Head Start community and other early childhood practitioners and policymakers.

In this section, we discuss the approach to estimating impacts when grantees are randomly assigned and when centers are randomly assigned. The following section describes the approach to the cost-effectiveness analysis.

1. Weighting the Sample

Analysis weights will be created to account for variations in the probabilities of selection, eligibility rates, and cooperation rates among those selected. First, the probability of selection should be calculated for each stage of sampling (grantee, center, classroom, and child), and within each explicit sampling stratum (for example, region). The inverse of the probability of selection at each stage is called the sampling weight. This stage of weighting would account for the probability proportional to size (PPS) sampling approach, any certainty selections, and any oversampling. At each stage, the sampling weights will be adjusted by the inverse of the weighted response rate among eligible cases at that stage.

If sampling is based on a PPS method at each stage, the children in the resulting sample will have roughly equal probability of selection. In this case, weights are unlikely to be very different for each child, and, therefore, the weighted analyses will not be very different from the unweighted analyses.

2. Estimating Impacts of the Quality Enhancement

Using a random assignment evaluation design means that fairly simple estimation methods can be used to determine the impacts of the quality enhancements at a point in time. Under random assignment of centers, the center-level mean outcomes are estimated, after which, separately for the enhancement and the control groups, the center mean outcomes are averaged over all centers in that group. An analogous approach is taken when grantees are randomly assigned. The simple difference in the means between enhancement and control centers (grantees) is the impact estimate.

More-precise estimates can be obtained by estimating regression models. Regression procedures can improve the precision of the estimates and can adjust for any residual differences in the observable characteristics of program and control group members due to random sampling and interview nonresponse. Regression models take the following form:

$$(1) \qquad Y = \alpha + X\beta + \gamma T + \varepsilon ,$$

where Y is an outcome variable; X is a vector of explanatory variables; T is an indicator that equals one for members of the enhancement group and zero for members of the control group; α, β, and γ are parameters to be estimated; and ε is a random-error term. The estimate of the parameter γ is the estimated impact of the quality enhancement compared with regular Head Start services.

Because random assignment will have been conducted at the grantee or center level (depending on the design) and sampling will have taken place at the grantee, center, and, sometimes, classroom level, the regression adjustment takes the form of a hierarchical linear model (HLM) of child development consisting of two or three nested levels (depending on whether classrooms are sampled). By specifying the model at each level, we can conduct analyses for the appropriate units of analysis and can conduct statistical hypothesis tests that correctly account for the clustering of children within classrooms, classrooms within centers, and centers within grantees. Although not strictly necessary for conducting impacts (because the evaluation is based on a random assignment design), using an HLM model to adjust the impacts can help increase the precision of the estimates.

For the design involving random assignment of grantees with sampling of grantees, centers, and children, the analytic model is the following, where the variables are indexed by child (*i*), centers (*j*) and grantees (*k*):

Child-level model

(2) $Y_{ijk}(t) = \alpha Y_{ijk}(0) + \beta X_{ijk} + c_{jk} + \varepsilon_{1ijk}$.

Center-level model

(3) $c_{jk} = \Ozeta Z_{jk} + gk + \varepsilon_{2jk}$.

Grantee-level model

(3) $g_k = \gamma T_k + \delta V_k + \varepsilon_{3k}$.

For the design involving random assignment of centers with sampling of grantees, centers, classrooms, and children, the analytic model is the following, where the variables are indexed by child (*i*), classrooms (*h*), centers (*j*), and grantees (*k*):

Child-level model

(4) $Y_{ihjk}(t) = \alpha Y_{ihjk}(0) + \beta X_{ihjk} + \ell_{hjk} + \varepsilon_{4ihjk}$.

Classroom-level model

(5) $\ell_{hjk} = \psi W_{hjk} + c_{jk} + \varepsilon_{5hjk}$.

Center-level model

(5) $c_{jk} = \lambda T_j + \psi \Ozeta Z_{jk} + gk + \varepsilon_{6jk}$

Grantee-level model

(3) $g_k = \delta V_k + \varepsilon_{7k}$.

In the models, *Y(t)* is the outcome at follow-up period *t*; *X* is a set of child characteristics, such as gender; *T* is a variable indicating whether the child was in the enhancement group or control group; *W* is a set of classroom-level variables, such as the classroom literacy environment; *Z* is a set of center-level variables, such as whether classes are full day; *V* is a set of grantee-level variables, such as the region or grantee director's education level; and ε_1 through ε_7 are disturbance terms assumed to have a mean of zero, and to be uncorrelated with each other. Parameters to be estimated include α and β, vectors of coefficients on the child baseline characteristics; *c,* the center effect; γ or λ, the effect of the quality enhancement in the two models; δ, the coefficients on the grantee variables; υ, the coefficients on the center variables; and ψ, the coefficients on the classroom variables.

The statistical techniques used to estimate regression-adjusted impacts in equations (2) and (4) will depend on the form of the outcome, *Y*. If the dependent variable is continuous (such as the score on the Woodcock-Johnson Letter-Word Identification assessment), ordinary least squares methods produce unbiased estimates of the parameters γ or λ. However, if the dependent variable is binary (such as whether the child's score on the Woodcock-Johnson Letter-Word Identification subtest was one or more standard deviations below the norm), logit or probit maximum-likelihood methods will be used to obtain consistent parameter estimates.

3. Adjusting for Participation

The research design is based on the assumption that random assignment will be implemented well so that a large proportion of children in the enhancement group remain in that enhancement center for the Head Start year and all children in control-group centers remain in those centers for the Head Start year. Nevertheless, children who start out in an enhancement center might move during the Head Start year to a different center that has not implemented the enhancement. At the same time, children in a control-group center might move during the Head Start year to a center that has implemented the enhancement.

To the extent that these movements across centers occur, the impact analysis described above will not measure completely the average impact of the quality enhancement on children, but rather the impact of being given the *opportunity* to attend a center with the enhancement; this is called the "intent to treat" impact estimate. The effect of both types of "crossovers" is to reduce the average impact because some children in the enhancement group did not receive the full effect of the enhancement and some children in the control group did. Researchers can adjust the impact estimates to account for such crossovers. The intent-to-treat estimate can be adjusted by dividing it by the difference between the proportion of students who remained in the enhancement centers and the proportion of students from the control centers who switched to an enhancement center.

4. Subgroup Analyses

Because the effectiveness of an enhancement may differ by program setting or by characteristics of the children served, it would be useful to determine the groups for which the enhancement is most effective. This type of subgroup analysis would then enable individual programs to decide which enhancements might be useful to them. For example, analysis may demonstrate that the early literacy curriculum is much more effective in programs that offer full-day services than in those offering part-day services, or that its impacts are stronger when teachers have higher educational credentials at the outset of training.

The subgroups of interest will depend on the quality enhancement to be tested. Examples of center- or grantee-level characteristics that can define subgroups include:

- Full-day or part-day program
- High-fidelity implementation or incomplete implementation
- Teachers' qualifications
- Center or program size

The program subgroups defined by whether the enhancement was implemented with high fidelity or not are perhaps the most important for understanding the impacts of the quality enhancement. High-fidelity implementation of the quality enhancement is critical to providing a fair test of its effectiveness. High-fidelity implementation should also be feasible by Stage 3, as the details of how to implement the quality enhancement in a variety of Head Start settings should have been developed by this point. Nevertheless, rather than assuming that implementation was successful, the Stage 3 evaluation plan should include measurement of fidelity.

If some centers or grantees participating in the evaluation did not reach critical thresholds for high-fidelity implementation, the impact analysis should include an examination of subgroups defined by degree of fidelity to the enhancement model. Constructing subgroups in both the enhancement and control groups for the analysis of impacts by implementation status poses a challenge because the control group should not have implemented the enhancement. Under center random assignment, the best approach is to define the implementation status for the two enhancement centers within a particular grantee. If both centers reached high-fidelity implementation, both the enhancement and control group centers from that grantee would be placed in the high-fidelity implementation group for analysis. An analogous procedure would follow if both centers did not reach high-fidelity implementation; the enhancement and control centers within a particular grantee would be classified according to the implementation status of both enhancement centers. If the implementation status of the two enhancement centers is divergent, however, some judgment would need to be made regarding the appropriate placement of the set of centers from that grantee. If both are fairly close to the threshold for full implementation but only one exceeded the threshold, they could both be placed in the high-fidelity category. If the statuses of the two centers are quite different, however, it might be best to omit the

enhancement and control centers from that grantee from the analysis. Under grantee random assignment, each grantee is selected from within sampling strata defined by region, grantee size, and other characteristics. These sampling strata could define "groups" of enhancement and control grantees that could then be classified according to the implementation status of all or most of the enhancement-group grantees.

Examples of categories of child and family characteristics (measured prior to the experience of the quality enhancement) for subgroup analysis include:

- Gender
- Children's English proficiency
- Mothers' education levels
- Family income levels
- Parents' employment status

Subgroup estimates can be obtained using the same procedures as the ones to be used for calculating overall impacts,[8] but these calculations are made for particular subgroups. Regression-adjusted subgroup estimates can be obtained by introducing an interaction term that is the product of the treatment indicator and an indicator of membership in the subgroup of interest. This term is entered into the model appropriate for the level of random assignment, and for whether the subgroup is defined at the child level or at the classroom, center, or grantee level.

However, unless the overall sample is very large, it will be possible to detect impacts only for large subgroups of the population, as subgroup estimates, which are based on only part of the full sample, are less precise than are full-sample estimates. For example, in the center random assignment design with minimal baseline data, our sample of 1,416 children (in 118 centers) is sufficient for detecting impacts with effect sizes of .20 (one-fifth of a standard deviation on the outcome measures) or more. However, for a subgroup that includes 50 percent of the children across all the centers (for example, children whose mothers have less than a high school diploma), impacts with effect sizes of .22 or more can be detected. For a subgroup that includes all of the children in half the centers (for example, full-day programs), impacts with effect sizes of .24 or more can be detected.

E. COST-EFFECTIVENESS ANALYSIS

Program administrators care not only about the effectiveness of the quality enhancements, but also the costs of these enhancements. An enhancement may not be implemented if its impacts are not viewed as commensurate with the amount of program

[8] For center-level random assignment, calculate center-level mean outcomes and average across centers in the enhancement and control groups.

resources it requires. Thus, program administrators must be provided with estimates of the quality enhancement's costs as well as estimates of the program impacts.

Relevant Costs of the Enhancement. To support analyses of the cost of an enhancement relative to its impacts or benefits, researchers ideally would obtain information during the implementation period on total program costs relative to the cost of the program without that enhancement. Researchers should measure two types of costs: (1) the upfront, one-time costs of beginning implementation; and (2) the ongoing additional costs resulting from the enhancement. Among the first set of costs are the costs associated with initial training, including the T/TA staff's time, the cost of substitute teachers hired to cover classrooms while regular teachers attend training, and the cost of additional staff days for teachers who are paid for the days that they attend training. The first set of costs also includes costs associated with the training staff's technical assistance visits and any costs associated with having teachers attend extra training sessions outside their normal classroom duties (for example, periodic group discussions, if any). Time spent by teachers and trainers would be valued at their hourly wage, including fringe benefits. If this training actually supplanted the usual teacher training sessions conducted during the year, those routine training costs should be subtracted. If space for training is rented, the cost of the rental will be included as well. Materials, including documents, videos, and other training materials, should be valued at their cost.

Ongoing additional costs resulting from enhancements would include the costs of materials, any computer-related costs, the cost of ongoing technical assistance, and costs to train new teachers. Teachers will have to maintain a ready supply of materials for any curriculum-related enhancement. Because broken or lost materials must be replaced, the cost of maintaining the supply of curriculum materials during the program year is part of its ongoing cost. Similarly, if a computer-based software program is part of the curriculum, the cost of setting up and maintaining the personal computers in the classroom with appropriate software is an additional cost. The cost of refresher training and ongoing technical assistance to teachers during a normal program year to ensure that the curriculum continues to be implemented to high fidelity is an ongoing cost as well. For example, T/TA staff might make two or three visits to each classroom during the program year to observe; measure fidelity; and meet with the teachers and with the education coordinator to discuss classroom practices, and to respond to questions. Finally, overall teacher turnover in Head Start is approximately 15 percent per year (National Institute for Early Education Research 2003). Accordingly, an ongoing cost of the mathematics curriculum is the cost of T/TA provided to 15 percent of the teachers in the enhancement group each year. Assuming that the evaluation includes 20 to 50 classrooms in the enhancement group, three to seven teachers would have to be fully trained each year.

Any of three different approaches to estimating these costs of implementing an enhancement could be used:

- Identify the costs associated with the enhancement and ask center directors and teachers to provide this information

- Obtain the full budget for centers (or grantees) assigned to the enhancement group for the year preceding enhancement implementation and for the current year and estimate the cost of the enhancement as the difference in budgets (adjusting for normal cost inflation from one year to the next)
- Obtain the full budget for the centers (or grantees) assigned to the enhancement group and for those assigned to the control group and estimate the cost of the enhancement as the difference between enhancement and control group costs

The first approach is the least burdensome for the enhancement centers or grantees (and eliminates burden for the control centers or grantees) but could fail to account for costs associated with the enhancement. The second approach is more burdensome for the enhancement group than is the first one, and it does not require a response from the control group; however, if anything other than the implementation of the enhancement changes from one year to the next, the estimate of the cost of the enhancement will be inaccurate. The third approach is the most burdensome, involving extensive collection of cost information from both enhancement and control groups, but it would provide the most accurate estimate of the costs of implementing the enhancement.

Obtaining Cost Information in a Field Test Evaluation. Information about costs can be obtained from semistructured interviews with program and center directors, from T/TA supervisory staff, and from program records. For a field test that relies on NRS data without making visits to programs to collect data, cost information could be obtained from telephone interviews with key T/TA supervisory staff about the number of hours spent at each program or center site, and from Head Start program cost records submitted annually to the federal government. However, Head Start programs often receive funding from multiple sources, so a particular program's federal Head Start budget could be an imperfect measure of program costs. Interviews with program and center directors about costs could be accommodated in a study that includes visits to programs to collect fidelity and child outcome data, but this approach would require extending the visits, thereby adding to evaluation costs.

All cost information must be obtained in dollars. However, to ensure that the dollar values collected at various time points represent the same value, the dollar values collected from informants and records should be either inflated or deflated, using the Consumer Price Index, to represent dollar values in a single target year (for example, the analysis and reporting year).

Finally, cost information sometimes is obtained using two perspectives. First, the actual dollar costs of the curriculum must be obtained. Second, a measure of cost to society would place a value on any volunteer labor or donated space and materials and would add those costs to the actual expended costs. The latter costs would be difficult to add to a study that relies on data obtained without visiting programs. Both cost perspectives could be used in the analysis of the cost-effectiveness of an enhancement, if researchers visit programs for the purpose of conducting data collection.

Approach to the Cost-Effectiveness Analysis. A benefit-cost analysis allows a direct comparison of impacts and costs by converting the impacts into "benefits" that are valued in dollars. A dollar value is placed on the impacts by estimating the market price for the change in resources that results from the impact. For example, a reduction in the use of special education services in the elementary school that results from the quality enhancement can be valued by estimating the resulting reduction in special education costs.

Impacts on the cognitive, social-emotional, and physical development of children, however, are very difficult to value. Children's development during the pre-school years may affect future employment and earnings, involvement with the criminal justice system, and use of public assistance programs—all of which can be quantified. However, as yet, insufficient evidence exists on the relationships between preschool child outcomes and these later outcomes to be able to put a dollar value on changes in the early outcomes. Accordingly, we do not recommend conducting a cost-benefit analysis based on outcome data collected from a single year in Head Start.

Instead, a cost-effectiveness framework can be used to evaluate the costs and benefits of the enhancement. This type of analysis does not attempt to place a dollar value on impacts. Instead, impacts are measured in a common unit, such as an effect size (the impact divided by the standard error of the outcome measure). The impact in effect-size units is compared with costs measured in dollars. For each quality enhancement, an effect size per dollar spent on the enhancement can be calculated. For example, if an early literacy curriculum were to produce an impact on vocabulary of 0.3 in effect-size units, and if the cost were estimated to be $10 per child, then the cost-effectiveness of the curriculum would be 0.03 per dollar. Measuring cost-effectiveness in this way enables program administrators to compare the cost of producing impacts that they consider important, using the same metrics, so that enhancements can be assessed in terms of their ability to provide the most "bang for the buck."

REFERENCES

Abbott-Shim, M., and A. Sibley. Assessment Profile for Early Childhood Programs. Atlanta, GA: Quality Assist, Inc., 1987.

Abidin, R.R. *The Parenting Stress Index Professional Manual.* Odessa, FL: Psychological Assessment Resources, 1995.

Achenbach, Thomas M., and Leslie A. Rescorla. *Manual for the ASEBA Preschool Forms and Profiles.* Burlington, VT: University of Vermont Department of Psychiatry, 2000.

August, D., M. Calderon, and M. Carlo. "Transfer of Skills from Spanish to English: A Study of Young Learners." Washington, DC: Center for Applied Linguistics, 2002.

Beatty, Alix. *Mathematical and Scientific Development in Early Childhood: A Workshop Summary*. Mathematical Sciences Education Board and Board on Science Education, National Research Council. Washington, DC: National Academies Press, 2005.

Bloom, Howard. "Randomizing Groups to Evaluate Place-Based Programs." New York, NY: MDRC, 2004.

Boller, Kimberly, Sharon McGroder, Shefali Pai-Samant, Christine Ross, and Diane Paulsell. "Measuring Outcomes Targeted by Head Start Quality Enhancement Initiatives." Volume 1: Final Concept Paper. Princeton, NJ: Mathematica Policy Research, Inc., September 2004.

Bronfenbrenner, U. "The Ecology of Human Development: Experiments by Nature and Design." Cambridge, MA: Harvard University Press, 1979.

Bronfenbrenner, U. "Ecology of the Family as a Context for Human Development: Research Perspectives. *Developmental Psychology,* vol. 22, 1986, pp. 723-742.

Burts, D.C., C.H. Hart, R. Charlesworth, and L. Kirk. "A Comparison of Frequencies of Stress Behaviors Observed in Kindergarten Children in Classrooms with Developmentally Appropriate versus Developmentally Inappropriate Practices." *Early Childhood Research Quarterly,* vol. 5, 1990, pp. 407-423.

Caldwell, Bettye M., and Robert H. Bradley. "Home Observation for Measurement of the Environment: Administration Manual, Revised Edition." Unpublished manuscript. Little Rock: University of Arkansas at Little Rock, 1984.

Carrow-Woolfolk, Elizabeth. *Oral and Written Language Scales.* Circle Pines, MN: AGS Publishing, 1995

Casey, Beth, Joanne E. Kersh, and Jessica Mercer Young. "Storytelling Sagas: An Effective Medium for Teaching Early Childhood Mathematics." *Early Childhood Research Quarterly,* vol. 19, 2004, pp. 167-172.

Center on the Social and Emotional Foundations for Early Learning. Inventory of Practices for Promoting Children's Social and Emotional Competence. *Young Children: Beyond the Journal*, July 2003. Downloaded from http://www.journal.naeyc.org/btj/200307/materials_support.asp, April 30,2005.

Clements, Douglas H., J. Sarama, and A.M. Dibiase (eds.). *Engaging Young Children in Mathematics: Standards for Early Childhood Mathematics Education.* Mahwah, NJ: Lawrence Erlbaum Associates, 2004.

Cochran, William. *Sampling Techniques.* New York: John Wiley and Sons, 1977.

Cohen, J. *Statistical Power Analysis for Behavioral Sciences.* Hillside, NJ: Lawrence Erlbaum, 1988.

Conduct Problems Prevention Research Group. "Initial Impact of the Fast Track Prevention Trial for Conduct Problems II: Classroom Effects." *Journal of Consulting and Clinical Psychology*, vol. 67, 1999, pp. 648-657.

Cordes, S., and R. Gelman. "The Young Numerical Mind: When Does It Count?" In *Handbook of Mathematical Cognition,* edited by Jamie I.D. Campbell. New York: Psychology Press, 2004.

Domitrovich, C., and Mark Greenberg. "Promoting Social and Emotional Competence in Head Start Children and Their Families." Presentation to the National Head Start Training and Technical Assistance Meeting, December 2001.

Duncan, G.J., and J. Brooks-Gunn (eds.). "Consequences of Growing Up Poor." New York: Russell Sage, 1997.

Duncan, S.E., and E.A. De Avila. *PreLAS 2000.* Monterey, CA: CTB/McGraw-Hill, 1998.

Dunn, L.M., and L.M. Dunn. *Peabody Picture and Vocabulary Test, Third Edition. Examiner's Manual and Norms Booklet.* Circle Pines, MN: American Guidance Service, 1997.

Dunn, L.M., E.R. Padilla, D.E. Lugo, and L.M. Dunn. *Test de Vocabulario en Imagenes Peabody.* Circle Pines, MN: American Guidance Service, 1986.

Espinosa, Linda. "The Promise of Diversity: Educating Young Children from Diverse Backgrounds." Presentation at the Head Start Research Conference, Washington, DC, June 30, 2004.

FACES Research Team. "Behavior Problems Scale." In *Head Start FACES: Longitudinal Findings on Program Performance: Third Progress Report.* Washington, DC: U.S. Department of Health and Human Services, Administration on Children, Youth and Families, Head Start Bureau, 2001, pp. 129-133.

Fagan, J. "African American and Puerto Rican American Parenting Styles, Parental Involvement, and Head Start Children's Social Competence." *Merrill-Palmer Quarterly,* vol. 46, 2000, pp. 592-612.

Fantuzzo, J., Sutton-Smith, B., Atkins, M., Myers, R., Stevenson, H., Coolahan, K., Weiss, A., and Manz, P. "Community-Based Resilient Peer Treatment of Withdrawn Maltreated Preschool Children." *Journal of Consulting and Clinical Psychology,* vol. 64, 1996, pp. 1377-1386.

Frey, K., M.K. Hirschstein, and B. A. Guzzo. "Second Step: Preventing Aggression by Promoting Social Competence." *Journal of Emotional and Behavioral Disorders*, vol. 8, no. 2, 2000, pp. 102-113.

General Accounting Office. "Head Start: Increased Percentage of Teachers Nationwide Have Required Degrees, but Better Information on Classroom Teachers' Qualifications Needed." Washington, DC: GAO, October 2003.

Ginsburg, Herbert P., and Susan L. Golbeck. "Thoughts on the Future of Research on Mathematics and Science Learning and Education." *Early Childhood Research Quarterly,* vol. 19, 2004, pp. 190-200.

Gonzales, Phillip. "Becoming Bilingual : First and Second Language Acquistion." Head Start Information and Publication Center. February 2005. http://www.headstartinfo.org/English_lang_learners_tkit/Bilingual.htm

Greenes, Carole, Herbert P. Ginsburg, and Robert Balfanz. "Big Math for Little Kids." *Early Childhood Research Quarterly,* vol. 19, 2004) pp. 159-166.

Gresham, F.M., and S.N. Elliott. *The Social Skills Rating System.* Circle Pines, MN: American Guidance Service, 1990.

Grossman, D.C., H.J. Neckerman, T.D. Koepsell, P. Liu, K.N. Asher, K. Beland, K. Frey, and F.P. Rivara. "Effectiveness of a Violence Prevention Curriculum Among Children in Elementary School: A Randomized Controlled Trial." *Journal of the American Medical Association,* vol. 277, 1997, pp. 1605-1611.

Hampton, V. "Validation of the Penn Interactive Peer Play Scale (PIPPS) for Urban Kindergartern Children." Unpublished doctoral dissertation, University of Pennsylvania, 1999.

Harms, T., Clifford, R.M., and Cryer, D. *Early Childhood Environment Rating Scale.* Revised edition. New York: Teachers College Press, 1998.

Hart, Betty, and Todd R. Risley. *Meaningful Differences in the Everyday Experience of Young American Children*. Baltimore, MD:Brookes Publishing Co., 1995.

Hart, Katherine, and Rachel Schumacher. "Moving Forward: Head Start Children, Families, and Programs in 2003." *CLASP Policy Brief, Head Start Series,* Brief no. 5. Washington, DC: Center for Law and Social Policy, June 2004.

Hart, C.H., M.D. DeWolf, P. Wozniak, and D.C. Burts. "Maternal and Paternal Disciplinary Styles: Relations with Preschoolers' Playground Behavioral Orientations and Peer Status." *Child Development,* vol. 63, 1992, pp. 879-892.

Heaviside, Sheila, and Elizabeth Farris. "Public School Kindergarten Teachers' Views on Children's Readiness for School." Washington, DC: U.S. Department of Education, National Center for Education Statistics, 1993.

Hedges, Larry. "Effect Sizes in Multi-Site Designs Using Assignment by Cluster." *Working Paper.* Chicago, IL: University of Chicago, 2004.

Hedges, Larry. "Correcting Significance Tests for Clustering." *Working Paper.* Chicago, IL: University of Chicago, 2004.

High/Scope Educational Research Foundation. "Preschool Program Quality Assessment." Ypsilanti, MI: High/Scope, 2003.

Iglesias, Aquiles. "Cultural and Linguistic Sensitivity in Test Construction." Presentation at the Head Start Research Conference Meetings, Washington, DC, June 29, 2004.

Karoly, L.A., M.R. Kilburn, J.H. Bigelow, J.P. Caulkins, J.S. Cannon, and J.R. Chiesa. "Assessing Costs and Benefits of Early Childhood Intervention Programs: Overview and Application to the Starting Early Starting Smart Program." Seattle, WA, and Santa Monica, CA: Casey Family Programs and RAND, 2001.

Kish, Leslie. *Survey Sampling.* New York: John Wiley and Sons, 1965.

Kochanoff, Anita. "Psychometric Properties of Standardized Assessments Available in Spanish." Presentation at the Head Start Research Conference Meetings, Washington, DC, June 2004.

Kochanoff, Anita, Kathy Hirsh-Pasek, Nora Newcombe, and Marsha Weinraub. "Using Science to Inform Preschool Assessment: A Summary Report of the Temple University Forum on Preschool Assessment." Philadelphia, PA: Temple University, 2003.

Kraemer, Helena Chmura. "Why a Randomized Controlled Trial (RCT) to Evaluate the Impact of Head Start? And Which One?" Paper prepared for the Advisory Committee on Head Start Research and Evaluation, June 1999. Available at [www.acf.dhhs.gov/programs/hsb/research/hsreac/jun1999.Kramer.htm]. Accessed July 14, 2003.

Kuperschmidt, J., and Donna Bryant. "The Effectiveness of a Classroom and Parenting Intervention Program for Aggressive Preschoolers in Head Start." Paper presented to Society for Research in Child Development, Minneapolis, MN, 2001.

Kupersmidt, J.B., Donna Bryant, and M. Willoughby. "Prevalence of Aggressive Behaviors Among Preschoolers in Head Start and Community Child Care Programs." *Behavioral Disorders,* vol. 26, no. 1, 2000, pp. 42-52.

Ladd, G.W., and C.C. Coleman. "Children's Classroom Peer Relationships and Early School Attitudes: Concurrent and Longitudinal Associations." *Early Education and Development,* vol. 8, 1997, pp. 51-66.

Lambert, Richard G., Martha Abbott-Shim, and Cindy Oxford-Wright. "Policy and Program Management Inventory." 2004. Downloaded May 16, 2005 from education.uncc.edu/qrc/teacher%20version.pdf.

LaParo, Karen, Robert Pianta, and Megan Stuhlman. "Classroom Assessment Scoring System (CLASS): Findings from the Pre-K Year." *The Elementary School Journal*, vol. 104, pp. 409-426.

Lara-Cinisomo, Sandraluz, Anne R. Pebley, Mary E. Valana, and Elizabeth Maggio. "Are L.A.'s Children Ready for School?" Report no. MG-145-FFLA. Santa Monica, CA: RAND, 2004.

LaFreniere, P.J., and J.E. Dumas. "Social Competence and Behavior Evaluation in Children Ages 3 to 6 Years: The Short Form (SCBE-30)." *Psychological Assessment,* vol. 8, 1996, pp. 369-377.

Lipsey, M.W., and D.B. Wilson. "The Efficacy of Psychological, Educational, and Behavioral Treatment." *American Psychologist,* vol. 48, no. 12, pp. 1181-1209, 1993.

Lopez, M., R. Cook, M. Puma, R. Gonzales, C. Heid, and G. Adams. "The Head Start Impact Study." Presentation at Head Start's Fifth National Research Conference, Washington, DC, June 2002. Available at [www.acf.dhhs.gov/programs/core/ongoing_research/hs/impact_study/hsrc_impact_study.pdf]. Accessed December 11, 2003

Love, J.M. "Instrumentation for State Readiness Assessment: Issues in Measuring Children's Early Development and Learning." In *Assessing the State of State Assessments: Perspectives on Assessing Young Children,* edited by C. Scott-Little, S.L. Kagan, and R.M. Clifford. Greensboro, NC: The Regional Educational Laboratory at SERVE, 2003.

McDermott, P.A., L.F. Green, J.M. Francis, and D.H. Stott. *Preschool Learning Behaviors Scale*. Philadelphia, PA: Edumetric and Clinical Science, 2000.

McMahon, Susan D., Washburn, J., Felix, E.D., Yakin J., and Childrey, G. "Violence Prevention: Program Effects on Urban Preschool and Kindergarten Children." *Applied and Preventive Psychology*, 9 (2000), 271-281.

McWayne, Christine M., John W. Fantuzzo, and Paul A. McDermott. "Preschool Competency in Context: An Investigation of the Unique Contribution of Child Competencies to Early Academic Success." *Developmental Psychology,* vol. 40, no. 4, pp. 635-645.

National Institute for Early Education Research. "Investing in Head Start Teachers." *Preschool Policy Matters,* Issue 4. New Brunswick: NIEER, August 2003.

National Research Council. *Mathematical and Scientific Development in Early Childhood*. Washington, DC: National Academies Press, 2005.

Newcomer, P.L., and D.D. Hammill. *Examiner's Manual: Test of Language Development—Primary—Third Edition*. Austin, TX: Pro-Ed, 1997.

Nord, C.W., and J. West. "Fathers' and Mothers' Involvement in Their Children's Schools by Family Type and Resident Status." Washington, DC: U.S. Department of Education, National Center for Education Statistics, May 2001.

Nord, C.W., D. Brimhall, and J. West. "Fathers' Involvement in Their Children's Schools." Washington, DC: U.S. Department of Education, National Center for Education Statistics, 1997.

Parke, R.D., J. Cassidy, V.M. Burks, J.L. Carlson, and L.A. Boyum. "Familial Contributions to Peer Competence Among Young Children: The Role of Interactive and Affective Processes." In *Family-Peer Relationships: Models of Linkage,* edited by R.D. Parke and G.W. Ladd. Hillsdale, NJ: Lawrence Erlbaum, 1992.

Paulsell, Diane, Kimberly Boller, Christine Ross, Angie KewalRamani, and Shefali Pai-Samant. "Head Start Quality Enhancements: Assessing Implementation, Fidelity, and Interim Outcomes." Princeton, NJ: Mathematica Policy Research, Inc., June 2004.

Paulsell, Diane, Ellen E. Kisker, John M. Love, and Helen R. Raikes. "Understanding Implementation in Early Head Start Programs: Implications for Policy and Practice." *Infant Mental Health Journal*, vol. 23, no. 102, 2002, pp. 14-35.

Raudenbush, Stephen. "Statistical Analysis and Optimal Design for Cluster Randomized Trials." *Psychological Methods,* vol. 2, no. 2, pp.173-185, 1997.

Raudenbush, Stephen, J. Spybrook, X. Liu, and R. Congdon. "Optimal Design for Longitudinal and Multilevel Research: Documentation for the Optimal Design Software." Ann Arbor, MI: University of Michigan, July 2004.

Raver, C. Cybele. "Emotions Matter: Making the Case for the Role of Young Children's Emotional Development for Early School Readiness." *SRCD Social Policy Report*, vol. 16, no. 3, 2002.

Raver, C. Cybele, and Jane Knitzer. "Ready to Enter: What Research Tells Policymakers About Strategies to Promote Social and Emotional School Readiness Among Three- and Four-Year-Old Children." *Promoting the Emotional Well-Being of Children and Families: Policy Paper No. 3*. New York: National Center for Children in Poverty, 2002.

Rivlin, Alice M., and P. Michael Timpane (eds.). *Planned Variation in Education: Should We Give Up or Try Harder?* Washington, DC: The Brookings Institution, 1975.

Ross, Christine, Sheena McConnell, Shefali Pai-Samant, Peter Schochet, John Hall, and Cassandra Rowand. "Evaluating Quality Enhancement Initiatives in Head Start: A Concept Paper." Princeton, NJ: Mathematica Policy Research, Inc., April 2004.

St. Pierre, R. G., J. P. Swartz, B. Gamse, S. Murray, D. Deck, and P. Nickel. "National Evaluation of the Even Start Family Literacy Program." Cambridge, MA: Abt Associates Inc., 1995.

St. Pierre, Robert, Anne Ricciuti, Fumiyo Tao, Cindy Creps, Janet Swartz, Wang Lee, Amanda Parsad, and Tracy Rimdzius. "Third National Even Start Evaluation: Program Impacts and Implications for Improvement." Washington, DC: U.S. Department of Education, 2003.

Sarama, Julie, and Douglas H. Clements. "Building Blocks for Early Childhood Mathematics." *Early Childhood Research Quarterly,* vol. 19, 2004, pp. 181-189.

Scanlon, D., L. Gelzheiser, D. Fanuele, J. Sweeney, and L. Newcomer. "CLASSIC: Classroom Language Arts Systematic Sampling and Instructional Coding." Albany: The University at Albany, State University of New York, 2003.

Scanlon, D.M., and F.R. Vellutino. "A Comparison of the Instructional Backgrounds and Cognitive Profiles of Poor, Average, and Good Readers Who Were Initially Identified as at Risk for Reading Failure." *Scientific Studies of Reading,* vol. 1, no. 3, 1997, pp. 191-215.

Scanlon, D.M., and F.R. Vellutino. "Prerequisite Skills, Early Instruction, and Success in First Grade Reading: Selected Results from a Longitudinal Study." *Mental Retardation and Developmental Disabilities,* vol. 2, 1996, pp. 54-63.

Schochet, Peter Z., Susanne James-Burdumy, and Ellen Kisker. "Social and Character Development (SACD) Research Program: Analysis Plan for the Multisite Impact Analysis." Princteon, NJ: Mathematica Policy Research, Inc., February 2004.

Schumacher, Rachel, and Tanya Rakpraja. "A Snapshot of Head Start Children, Families, Teachers, and Programs: 1997 and 2001." Head Start Series, Policy Brief No. 1. Center for Law and Social Policy, March 2003.

Serna, L., E. Nielsen, K. Lambros, and S. Forness. "Primary Prevention with Children at Risk for Emotional or Behavioral Disorders: Data on a Universal Intervention for Head Start Classrooms." *Behavioral Disorders,* vol. 26, no. 1, 2000, pp. 70-84.

Shadish, William R., Thomas D. Cook, and Donald T. Campbell. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* New York: Houghton Mifflin Company, 2002.

Shonkoff, Jack, and Deborah Phillips. *From Neurons to Neighborhoods: The Science of Early Childhood Development.* Washington, DC: National Academy Press, 2001.

Smith, Judith R., Jeanne Brooks-Gunn, and Pamela K. Klebanov. "Consequences of Living in Poverty for Young Children's Cognitive and Verbal Ability and Early School Achievement." In *Consequences of Growing Up Poor,* edited by G. Duncan and J. Brooks-Gunn. New York: Russell Sage Foundation, 1997.

Smith, M.W., David K. Dickinson, A. Sangeorge, and A. Anasatopoulos. Toolkit for Assessing Early Literacy in Classrooms. Baltimore, MD: Brookes Publishing, 2002.

Sophian, Catherine. "Mathematics for the Future: Developing a Head Start Curriculum to Support Mathematics Learning." *Early Childhood Research Quarterly,* vol. 19, 2004, pp. 59-81.

Springer J.Fred, Elizabeth Sale, Michele Basen, and Peter Pecora. "Starting Early Starting Smart Final Report: Summary of Findings." Washington, DC: Casey Family Programs and the US Department of Health and Human Services, Substance Abuse and Mental Health Services Administration, 2003.

Starkey, Prentice, Alice Klein, Douglas Clements, and Julie Sarama. "Child Math Assessment—Abbreviated." 2002.

Starkey, Prentice, Alice Klein, and Ann Wakeley. "Enhancing Young Children's Mathematical Knowledge Through a Pre-Kindergarten Mathematics Intervention." *Early Childhood Research Quarterly,* vol. 19, 2004, pp. 99-120.

Tabors, Patton O., Mariela M. Paez, and Lisa M. Lopez. "Dual Language Abilities of Bilingual Four-Year-Olds: Initial Findings from the Early Childhood Study of Language and Literacy Development of Spanish-Speaking Children." *NABE Journal of Research and Practice*, winter 2003.

Tamis-LeMonda, Catherine S., Jacqueline D. Shannon, Natasha J. Cabrera, and Michael E. Lamb. "Fathers and Mothers at Play with Their 2- and 3-Year-Olds: Contributions to Language and Cognitive Development." *Child Development,* vol. 75, no. 6, December 2004, pp. 1806-1820.

Tao, Fumiyo, Beth Gamse, and Hope Tarr. "National Evaluation of the Even Start Family Literacy Program: 1994-1997. Washington, DC: U.S. Department of Education, Planning and Evaluation, 1998.

Tramontana, Michael G., Stephen R. Hooper, and S. Claire Selzer. "Research on the Preschool Prediction of Later Academic Achievement: A Review." *Developmental Review,* vol. 8, 1988, pp. 89-146.

Tudge, Jonathan R.H. and Fabienne Doucet. "Early Mathematical Experiences: Observing Young Black and White Children's Everyday Activities." *Early Childhood Research Quarterly*, vol. 19,. No. 1, 2004 pp. 21-39.

U.S. Department of Health and Human Services, Administration on Children, Youth and Families. "Charting Our Progress: Development of the Head Start Program Performance Measures." Washington, DC: DHHS, 1995.

U.S. Department of Health and Human Services. "Creating a Twenty-first Century Head Start: Report of the Advisory Committee on Head Start Quality and Expansion." Washington, DC: DHHS, 1993.

U.S. Department of Health and Human Services, Administration for Children and Families. English Language Learners Focus Group Report: Identifying Strategies to Support English Language Learners in Head Start and Early Head Start Programs. Washington, DC: U.S. DHHS, April 2002.

U.S. Department of Health and Human Services. "Evaluating Head Start: A Recommended Framework for Studying the Impact of the Head Start Program." Report of the Advisory Committee on Head Start Research and Evaluation. Washington, DC, October 1999.

U.S. Department of Health and Human Services, Administration on Children, Youth, and Families. "Head Start FACES: Longitudinal Findings on Program Performance." Third Progress Report. Washington, DC: DHHS, 2001.

U.S. Department of Health and Human Services, Administration on Children, Youth, and Families. "Head Start FACES 2000: A Whole-Child Perspective on Program Performance." Fourth Progress Report . Washington, DC: DHHS, 2003a.

U.S. Department of Health and Human Services. The Head Start Leaders Guide to Positive Child Outcomes: Strategies to Support Positive Child Outcomes. Washington, DC: U.S. DHHS, 2003b.

U.S. Department of Health and Human Services. *Head Start Program Fact Sheet, Fiscal Year 2003*. Washington, DC: U.S. DHHS, April 2004a. Downloaded from http://www.acf.hhs.gov/programs/hsb/research/2004.htm, March 1, 2005.

U.S. Department of Health and Human Services, Administration for Children and Families. "Head Start Program: Final Rule." *Federal Register,* vol. 61, no. 215, p. 57185-57227, November 5, 1996.

U.S. Department of Health and Human Services. "Interim Report for Quality Research Consortium Data Coordination Center: Cross-Sectional Analyses." Washington, DC: U.S. DHHS, 2004b.

U.S. Department of Health and Human Services, Administration on Children, Youth, and Families. "Head Start Research and Evaluation: A Blueprint for the Future. Recommendations of the Advisory Panel for the Head Start Evaluation Design Project." Washington, DC: U.S. DHHS, 1990.

U.S. Department of Health and Human Services, Administration on Children, Youth, and Families. Using Child Outcomes in Program Self-Assessment. Information Memorandum ACYF-IM-HS-00-18. Washington, DC: U.S. DHHS, August 10, 2000. Downloaded from http://www.headstartinfo.org/publications/im00/im00_18.htm June 6, 2005.

U.S. Department of Health and Human Services, Administration on Children, Youth, and Families. Initial Guidance on New Legislative Provisions on Performance Standards, Performance Measures, Program Self-Assessment and Program Monitoring. Information Memorandum ACYF-IM-HS-00-03. Washington, DC: U.S. DHHS, January 31, 2000. Downloaded from http://www.headstartinfo.org/publications/im00/im00_03.htm June 6, 2005.

Vellutino, F.R., and D.M. Scanlon. "Emergent Literacy Skills, Early Instruction, and Individual Differences as Determinants of Difficulties in Learning to Read: The Case for Early Intervention." In *Handbook of Early Literacy Research,* edited by Susan B. Neuman and David K. Dickinson. New York: Guilford Press, 2001.

Walker, H.M., K. Kavanagh, B. Stiller, A. Golly, H.H. Severson, and E. Feil. (1998). "First Step to Success: An Early Intervention Approach for Preventing School Antisocial Behavior." *Journal of Emotional and Behavioral Disorders,* vol. 6, 1998, pp. 66-80.

Webster-Stratton, C. "Preventing Conduct Problems in Head Start Children: Strengthening Parenting Competencies." *Journal of Consulting and Clinical Psychology,* vol. 6, 1998, pp. 715-730.

Webster-Stratton, C, M.J. Reid, and M. Hammond. "Preventing Conduct Problems, Promoting Social Competence: A Parent and Teacher Training Partnership in Head Start. *Journal of Child Clinical Psychology,* vol. 30, 2001, pp. 283-302.

West, Jerry, Kristin Denton, and Elvie Germino-Hausken. *America's Kindergartners: Findings from the Early Childhood Longitudinal Study, Kindergarten Class of 1998-99, Fall 1998*. NCES 2000-070, Washington, DC: U.S. Department of Educaton, 2000.

Whitehurst, Grover J. and Christopher J. Lonigan. "Child Development and Emergent Literacy." *Child Development*, vol. 69, no. 3, June 1998, pp. 848-872.

Whitehurst, Grover J., D.S. Arnold, J.N. Epstein, A.L. Angell, M. Smith, and J. Fischel. "A Picture Book Reading Intervention in Day Care and Home for Children from Low-Income Families." *Developmental Psychology,* vol. 30, 1994, pp. 679-689.

Woodcock, R.W., K. McGrew, and N. Mather. *Woodcock-Johnson-III Tests of Achievement*. Itasca, IL: Riverside Publishing, 2001.

Zaslow, M.J., K.A. Moore, D. Morrison, and M.J. Coiro. "The Family Support Act and Children: Potential Pathways of Influence." *Children and Youth Services Review,* vol. 17, 1995, pp. 231-249.

Zimmerman, Irla Lee, Violette G. Steiner, and Roberta Evatt Pond. *Preschool Language Scale, Fourth Edition (PLS-4), English Edition*. San Antonio: Harcourt Assessment, Inc., 2002.

Zimmerman, Irla Lee, Violette G. Steiner, and Roberta Evatt Pond. *Preschool Language Scale, Fourth Edition (PLS-4), Spanish Edition*. San Antonio: Harcourt Assessment, Inc., 2002.

APPENDIX A

THE HEAD START CHILD OUTCOMES FRAMEWORK[1]

LANGUAGE DEVELOPMENT

Listening and Understanding

- Demonstrates increasing ability to attend to and understand conversations, stories, songs, and poems
- Shows progress in understanding and following simple and multiple-step directions
- Understands an increasingly complex and varied vocabulary* [2]
- For non-English speaking children, progresses in listening to and understanding English*

Speaking and Communicating

- Develops increasing abilities to understand and use language to communicate information, experiences, ideas, feelings, opinions, needs, questions; and for other varied purposes*
- Progresses in abilities to initiate and respond appropriately in conversation and discussions with peers and adults
- Uses an increasingly complex and varied spoken vocabulary*

---

[1] From *The Head Start Bulletin*, Issue No. 76 (2003). Available at [www.headstartinfo.org/publications/hsbulletin76/hsb76_09.htm]

[2] Asterisks indicate the four domain elements and nine indicators that are legislatively mandated.

- Progresses in clarity of pronunciation and towards speaking in sentences of increasing length and grammatical complexity
- For non-English speaking children, progresses in speaking English*

LITERACY

Phonological Awareness

- Shows increasing ability to discriminate and identify sounds in spoken language
- Shows growing awareness of beginning and ending sounds of words
- Progresses in recognizing matching sounds and rhymes in familiar words, games, songs, stories, and poems
- Shows growing ability to hear and discriminate separate syllables in words
- Associates sounds with written words, such as awareness that different words begin with the same sound*

Book Knowledge and Appreciation*

- Shows growing interest and involvement in listening to and discussing a variety of fiction and nonfiction books and poetry
- Shows growing interest in reading-related activities, such as asking to have a favorite book read; choosing to look at books; drawing pictures based on stories; asking to take books home; going to the library; and engaging in pretend-reading with other children
- Demonstrates progress in abilities to retell and dictate stories from books and experiences, to act out stories in dramatic play, and to predict what will happen next in a story
- Progresses in learning how to handle and care for books; knowing to view one page at a time in sequence from front to back; and understanding that a book has a title, author, and illustrator

Print Awareness and Concepts

- Shows increasing awareness of print in classroom, home, and community settings
- Develops growing understanding of the different functions of forms of print such as signs, letters, newspapers, lists, messages, and menus
- Demonstrates increasing awareness of concepts of print, such as that reading in English moves from top to bottom and from left to right, that speech can be written down, and that print conveys a message

- Shows progress in recognizing the association between spoken and written words by following print as it is read aloud
- Recognizes a word as a unit of print, or awareness that letters are grouped to form words, and that words are separated by spaces*

Early Writing

- Develops understanding that writing is a way of communicating for a variety of purposes
- Begins to represent stories and experiences through pictures, dictation, and in play
- Experiments with a growing variety of writing tools and materials, such as pencils, crayons, and computers
- Progresses from using scribbles, shapes, or pictures to represent ideas, to using letter-like symbols, to copying or writing familiar words such as their own name

Alphabet Knowledge

- Shows progress in associating the names of letters with their shapes and sounds
- Increases in ability to notice the beginning letters in familiar words
- Identifies at least 10 letters of the alphabet, especially those in their own name*
- Knows that letters of the alphabet are a special category of visual graphics that can be individually named*

MATHEMATICS

Number and Operations*

- Demonstrates increasing interest and awareness of numbers and counting as a means for solving problems and determining quantity
- Begins to associate number concepts, vocabulary, quantities, and written numerals in meaningful ways
- Develops increasing ability to count in sequence to 10 and beyond
- Begins to make use of one-to-one correspondence in counting objects and matching groups of objects
- Begins to use language to compare numbers of objects with terms such as more, less, greater than, fewer than, and equal to
- Develops increased abilities to combine, separate, and name "how many" concrete objects

Geometry and Spatial Sense

- Begins to recognize, describe, compare, and name common shapes, their parts, and attributes
- Progresses in ability to put together and take apart shapes
- Begins to be able to determine whether two objects are the same size and shape
- Shows growth in matching, sorting, putting in a series, and regrouping objects according to one or two attributes such as color, shape, or size
- Builds an increasing understanding of directionality, order, and position of objects and of words such as up, down, over, under, top, bottom, inside, outside, in front, and behind.

Patterns and Measurement

- Enhances abilities to recognize, duplicate, and extend simple patterns using a variety of material.
- Shows increasing abilities to match, sort, put in a series, and regroup objects according to one or two attributes such as shape or size
- Begins to make comparisons between several objects based on a single attribute
- Shows progress in using standard and nonstandard measures for length and area of objects

SCIENCE

Scientific Skills and Methods

- Begins to use senses and a variety of tools and simple measuring devices to gather information, investigate materials, and observe processes and relationships
- Develops increased ability to observe and discuss common properties, differences, and comparisons between objects and materials
- Begins to participate in simple investigations to test observations, discuss and draw conclusions, and form generalizations
- Develops growing abilities to collect, describe, and record information through a variety of means including discussion, drawings, maps, and charts
- Begins to describe and discuss predictions, explanations, and generalizations based on past experience

Scientific Knowledge

- Expands knowledge of and abilities to observe, describe, and discuss the natural world, materials, living things, and natural processes
- Expands knowledge of and respect for their bodies and the environment
- Develops growing awareness of ideas and language related to attributes of time and temperature
- Shows increased awareness and beginning understanding of changes in materials and cause-effect relationships

CREATIVE ARTS

Music

- Participates with increasing interest and enjoyment in a variety of music activities including listening, singing, finger plays, games, and performances
- Experiments with a variety of musical instruments

Art

- Gains ability in using different art media and materials in a variety of ways for creative expression and representation
- Progresses in abilities to create drawings, paintings, models, and other art creations that are more detailed, creative, or realistic
- Develops growing abilities to plan, work independently, and demonstrate care and persistence in a variety of art projects
- Begins to understand and share opinions about artistic products and experiences

Movement

- Expresses through movement and dancing what is felt and heard in various tempos and musical styles
- Shows growth in moving in time to different patterns of beat and rhythm in music

Dramatic Play

- Participates in a variety of dramatic play activities that become more extended and complex

- Shows growing creativity and imagination in using materials and in assuming different roles in dramatic play situations

SOCIAL AND EMOTIONAL DEVELOPMENT

Self-Concept

- Begins to develop and express awareness of self in terms of specific abilities, characteristics, and preferences
- Develops growing capacity for independence in a range of activities, routines, and tasks
- Demonstrates growing confidence in a range of abilities and expresses pride in accomplishments

Self-Control

- Shows progress in expressing feelings, needs, and opinions in difficult situations and conflicts without harming themselves, others, or property
- Develops growing understanding of how their actions affect others and begins to accept the consequences of their actions
- Demonstrates increasing capacity to follow rules and routines and use materials purposefully, safely, and respectfully

Cooperation

- Increases abilities to sustain interactions with peers by helping, sharing, and discussion
- Shows increasing abilities to use compromise and discussion in working, playing, and resolving conflicts with peers
- Develops increasing abilities to give and take in interactions, to take turns in games or using materials, and to interact without being overly submissive or directive

Social Relationships

- Demonstrates increasing comfort in talking with and accepting guidance and directions from a range of familiar adults
- Shows progress in developing friendships with peers
- Progresses in responding sympathetically to peers who are in need, upset, hurt, or angry; and in expressing empathy or caring for others

Knowledge of Families and Communities

- Develops ability to identify personal characteristics, including gender and family composition
- Progresses in understanding similarities and respecting differences among people, such as genders, race, special needs, culture, language, and family structures
- Develop growing awareness of jobs and what is required to perform them
- Begins to express and understand concepts and language of geography in the contexts of the classroom, home, and community

APPROACHES TO LEARNING

Initiative and Curiosity

- Chooses to participate in an increasing variety of tasks and activities
- Develops increased ability to make independent choices
- Approaches tasks and activities with increased flexibility, imagination, and inventiveness
- Grows in eagerness to learn about and discuss a growing range of topics, ideas, and tasks

Engagement and Persistence

- Grows in abilities to persist in and complete a variety of tasks, activities, projects, and experiences
- Demonstrates increasing ability to set goals and develop and follow through on plans
- Shows growing capacity to maintain concentration over time on a task, question, set of directions or interactions, despite distractions and interruptions

Reasoning and Problem Solving

- Develops increasing ability to find more than one solution to a question, task, or problem
- Grows in recognizing and solving problems through active exploration, including trial and error, and interactions and discussions with peers and adults
- Develops increasing abilities to classify, compare and contrast objects, events, and experiences

PHYSICAL HEALTH AND DEVELOPMENT

Fine Motor Skills

- Develops growing strength, dexterity, and control needed to use tools such as scissors, paper punch, stapler, and hammer
- Grows in hand-eye coordination in building with blocks, putting together puzzles, reproducing shapes and patterns, stringing beads, and using scissors
- Progresses in abilities to use writing, drawing, and art tools, including pencils, markers, chalk, paint brushes, and various types of technology

Gross Motor Skills

- Shows increasing levels of proficiency, control, and balance in walking, climbing, running, jumping, hopping, skipping, marching, and galloping
- Demonstrates increasing abilities to coordinate movements in throwing, catching, kicking, bouncing balls, and using the slide and swing

Health Status and Practices

- Progresses in physical growth, strength, stamina, and flexibility
- Participates actively in games, outdoor play, and other forms of exercise that enhance physical fitness
- Shows growing independence in hygiene, nutrition, and personal care when eating, dressing, washing hands, brushing teeth, and toileting
- Builds awareness and ability to follow basic health and safety rules such as fire safety, traffic and pedestrian safety, and responding appropriately to potentially harmful objects, substances, and activities

APPENDIX B

SAMPLE SIZE REQUIREMENTS

This appendix discusses the estimation of sample size requirements for various designs for evaluating quality enhancement ideas in Head Start. We present a mathematical formulation to demonstrate the sources of variance under each sample design that would be plausible to use for evaluating the quality enhancement ideas. We then present estimates of sample size requirements for these sample designs.

The minimum detectable effect sizes (MDEs) represent the smallest true impact in effect size units (the percentage of the standard deviation of the outcome measure) that can be detected with a high probability. They pertain to overall impact estimates when data are pooled across all programs, centers, classrooms, and children in the study.

The MDE formula used in the calculations can be expressed as follows:

$$(1)\quad MDE = 2.802 * \sqrt{(1-R^2)Var(impact)} / \sigma,$$

where 2.802 is a constant that applies when using a two-tailed test at the 5 percent significance level, 80 percent power, and infinite degrees of freedom; R^2 is the R-squared value from the regression adjusting impact estimates for baseline demographics and other variables; *Var(impact)* is the variance of the impact estimate (mean treatment and control group difference in the outcome measure); this term should take into account design effects arising from the clustering of the sample and weighting; and σ is the standard deviation of the outcome measure.

The main focus of this appendix is the derivation of the variance of the impact estimates when samples are clustered under each design alternative. The precision of the impact estimates decreases substantially as the sample becomes more clustered. Children in Head Start programs are clustered in classrooms; classrooms are clustered in centers; and centers are clustered in grantees. Thus, for a given sample size of children, precision levels are maximized when the unit of random assignment is at the child level, and they decrease when the unit of random assignment is at the classroom, center, or program level. Consequently, by increasing the number of grantees and centers that must participate in the

study to yield a desired level of precision, clustering can have a large impact on the costs of the evaluation.

The design effect of weighting increases the variance of the impact estimates when the analysis incorporates differential weights that are constructed to adjust for differential participation rate of the enhancement and control groups. The design effect of weighting is a constant that is larger if the sample for either the enhancement group or control group is underrepresented and therefore differentially weighted (perhaps because of differential consent rates); it does not vary with sample size. It generally takes values between 1.1 and 1.5, with the lower values corresponding to samples with high participation rates. By making efforts to ensure high participation rates of centers, classrooms, and children, researchers can minimize the design effect of weighting. The likelihood of minimizing this term (and the possibility of other offsetting factors in the calculation of minimum detectable effects) has led us to calculate sample sizes for this appendix without the design effect of weighting. Nevertheless, it should be kept in mind as site recruitment and data collection efforts are planned because of its independent effect on the precision of the impact estimates.

A. VARIANCE CALCULATIONS FOR GROUP-BASED EXPERIMENTAL DESIGNS

In this section, we focus on the sources of variance under each sample design that could be used to evaluate quality enhancement ideas in Head Start. We show how *Var(impact)* from the MDE formula can be estimated under each sample design. The applicability of each design depends on the intervention under investigation. For example, some interventions must be tested at the center level due to potential spillover effects, whereas others could be tested at the classroom or even the child level.

1. Non-Clustered Design

We begin with the simplest, non-clustered design to illustrate the key components of the variance estimate. In this evaluation design, we identify a group of grantees (and/or centers) to participate in an evaluation of an individual child- or family-focused intervention, such as a mentoring program matching high school students and Head Start children or a special parent education program. In this type of non-clustered design, children within purposively selected (or volunteer) centers would be randomly assigned to a research status directly; no program-, center-, or classroom-level clustering would take place. The variance of the estimated impact on an outcome measure (that is, the difference between the mean outcome of treatment and control group members) must account for *between-child* variance only and can be expressed as follows:

$$(2)\quad Var(impact) = \frac{2\sigma^2}{cs},$$

where c is the number of centers in the sample, s is the average number of children per center in the treatment and control groups (which are assumed to be equal) and σ^2 is the variance of the outcome measure.

We assume equal numbers of treatment and control children, because, for a given total research sample size, a 50:50 split between the two groups yields the most precise estimates. A finite sample correction (equal to one minus the proportion of the population being selected) could be included in the equation; however, for simplicity and to produce conservative estimates, we do not include this term. We follow this approach for the remainder of this appendix.

As discussed in the report, a non-clustered design might not be feasible or applicable for many interventions. Instead, the designs typically involve random assignment of groups (classes, centers, or grantees) to the enhancement or control condition. In these designs, the variance is increased relative to that in equation (2) because the variance calculations must take into account the additional variance that arises because of clustering at the level of random assignment. The explanation is that the outcomes for children within these clusters are correlated, and a different random assignment would yield a different set of centers or classrooms in the treatment and control groups. In addition, the variance calculations must take into account additional variance that may arise if the set of programs, centers, and classrooms in the sample are to be representative of a broader Head Start population relative to whether the impact estimates will be generalized only to the sample included in the evaluation. We discuss these issues in additional detail in the next section.

2. Fixed or Random Effects Designs

A fundamental issue to consider when pooling across classrooms, centers, or grantees is whether site effects should be treated as *fixed* or *random*. This decision is likely to depend on whether the evaluation is taking a Stage 2 approach (smaller-scale evaluation, using volunteer sites) or a Stage 3 approach (field test). For most Stage 2 evaluations, sites (such as centers and grantees) will be volunteers that are *purposively* selected for the study because they are in a specific region, have a set of characteristics desired for the control group, and are willing and able to participate in the evaluation. In general, in these instances, the variance calculations for pooled impact estimates should not account for *between-site* variance terms. The sample was not randomly selected, so the pooled impact estimates can be generalized only to the study sites, rather than to a broader population of sites. Stated differently, the study can produce impact estimates that are internally valid, but not necessarily externally valid. For Stage 2 designs, where the goal is to determine whether an intervention *can* work, it is sufficient to demonstrate impacts within a purposively chosen sample (that is, a "best-case" scenario for the efficacy of the quality enhancement idea).

In contrast, Stage 3 evaluations are designed so that the results can generalize to all Head Start centers and children (or to a well-defined subset, such as a region). In these evaluations, grantees and centers will be *randomly selected* from all Head Start grantees and centers (or from a well-defined subpopulation of programs). Study results can be generalized more broadly in these random-effects designs than in the fixed-effects designs. However, this generalization involves a cost in terms of precision levels: the variance formulas must be inflated to account for between-site effects. Stated differently, site effects must be treated in the variance formulas as *random,* not as fixed. Intuitively, in repeated sampling, a different set of sites would be selected for the evaluation, which could influence

the impact findings. Hence, the variance expressions must account for the extent to which mean child outcomes vary across sites. In the remainder of this section, we present formulas for both fixed- and random-effects designs. These formulas can be derived from standard sampling theory for clustered designs (see Cochran 1977 and Kish 1965) or from random effects models assuming normally distributed error terms.

3. Variance Calculations for Head Start Quality Enhancement Designs

For evaluations of Head Start quality enhancements, we consider experimental designs in which any of the following units are randomly assigned to a research status:

- Children
- Classrooms
- Centers
- Grantees

We also consider designs in which these units are randomly *selected* from a larger pool of units, either before or after random assignment. For example, if centers are randomly assigned to the enhancement or a control group, we consider the case in which grantees are randomly selected from a larger set of grantees (the random-effects case), as well as when they are selected purposively (the fixed-effects case). Each level of random assignment and random selection introduces additional layers of clustering into the variance formulas.

a. Variance Estimates for Stage 3 Designs

Table B.1 summarizes the designs that we consider for Stage 3. It also provides an example of how the design might be used for a Stage 3 evaluation, and displays equation numbers for the variance formulas for each design.

For Stage 3 designs, we would randomly select grantees for the evaluation (and would include the appropriate variance terms for the grantee-level clustering). We would use either all centers or randomly select centers (and would do the same for classrooms within centers). No grantees or centers would be purposively selected; the goal would be to generalize the results across all Head Start programs, centers, and children. We present the designs from the least to the most clustered.

Random Assignment of Children. A design that could be used for some evaluations of Head Start quality enhancements is to randomly assign children or families to receive either quality enhancement services or regular Head Start services. For example, an intensive parent education program might be offered to randomly selected families within a Head Start center. Families in the control group would receive the normal Head Start services, whereas those in the intervention group would receive the parent education program. This type of design is appropriate for interventions that are administered on an individual child or family basis (such as a child mentoring program involving a match between selected children and high school students), and for which potential spillover effects are small.

Table B.1. Stage 3 Designs

Purposive Selection	Random (Representative) Selection	Sources of Clustering	Results Can Generalize to:	Equation Number for Variance Formula	Example from Report	Comments
Child-Level Random Assignment						
None	Grantee Center	Grantee Center	All Head Start children	3	Parent education intervention Match children with high school mentors	Intervention must be at the child level (not classroom-wide); few quality enhancements meet this criterion.
Classroom- or Teacher-Level Random Assignment						
None	Grantee Centers a. Use all classrooms b. Randomly select classrooms	Grantee Center Classrooms	All Head Start classes	4	Alternative mathematics curricula	Randomly assign teachers after class lists are set *or* randomly assign children to classrooms after teachers are randomly assigned. In a Stage 3 design, classroom random assignment would be very difficult to monitor.
Center-Level Random Assignment						
None	Grantee a. Use all centers b. Randomly select centers	Grantee Center	All Head Start centers	5, 6	Alternative mathematics curricula Approaches to addressing children's behavioral problems T/TA to use program data for quality improvement	Centers are randomly assigned, so clustering is at the center level; it does not affect power to use all centers or randomly select centers.
Grantee-Level Random Assignment						
None	Grantee Use all centers	Grantee	All Head Start grantees	8	T/TA to use program data for quality improvement	This design would be best if grantees do not include large numbers of centers
None	Grantee Randomly select centers	Grantee Center	All Head Start grantees	7	T/TA to use program data for quality improvement	This design would permit sampling centers so that each grantee includes a similar number of centers for analysis

T/TA = training and technical assistance.

At Stage 3, the participating grantees and centers would be randomly selected to represent all Head Start programs. However, because the intervention focuses on children, the power of the design could be strengthened by selecting the children without regard to classroom. This step would eliminate classroom-level clustering from the variance calculations. For this design, the variance calculations would be approximated by:

$$(3)\ Var(pooled\ impact) = \frac{2\sigma^2\rho_1(1-c_1)}{g} + \frac{2\sigma^2\rho_2(1-c_2)}{cg} + \frac{2\sigma^2(1-\rho_1-\rho_2)}{cgs},$$

where g is the total number of grantees in the sample; c is the average number of centers per grantee; s is the number of treatment or control children per center; ρ_1 is the grantee-level intraclass correlation in the outcome measure; ρ_2 is the center-level intraclass correlation; σ^2 is the variance of the outcome measure; c_1 is the correlation in mean outcomes between treatment and control group children within grantees; and c_2 is the correlation between treatment and control group children's mean outcomes within centers in grantees. The correlation in mean outcomes between treatment and control groups will be *smaller* if impacts are either large or of different sizes (and directions) for subgroups of children within treatment and control groups. In equation (3), the first term essentially represents the extent to which impacts differ across grantees, and the second represents the extent to which impacts differ across centers within grantees.

Random Assignment of Classrooms. Another design is to randomly assign classrooms (or teachers) *within* randomly selected centers to the intervention or control groups. For example, teachers could be randomly assigned to either try a new early mathematics curriculum or continue their usual practices in the Head Start classroom. This type of design is appropriate for interventions that are administered at the classroom level, and for which potential spillover effects are deemed to be small. Note, however, that because a very large number of classrooms and centers would be involved in a Stage 3 evaluation, it would be difficult to visit classrooms frequently enough to ensure that spillover does not occur. An enhancement involving classroom random assignment at Stage 3 would have to be intrinsically difficult to move from one classroom to another.

For a Stage 3 design, the grantees and centers would be sampled to represent all Head Start programs. Under this design, grantee-level, center-level, and classroom-level clustering are present and the variance expression can be approximated as follows:

$$(4)\ Var(impact) = \frac{2\sigma^2\rho_1(1-c_1^*)}{g} + \frac{2\sigma^2\rho_2(1-c_2^*)}{cg} + \frac{2\sigma^2\rho_3}{cgl} + \frac{2\sigma^2(1-\rho_1-\rho_2-\rho_3)}{cgls},$$

where g is the total number of grantees in the sample; c is the average number of centers per grantee; l is the average number of treatment or control classrooms per center; and s is the average number of children per classroom; ρ_1 is the grantee-level intraclass correlation in the outcome measure; ρ_2 is the center-level intraclass correlation; ρ_3 is the classroom-level intraclass correlation; σ^2 is the variance of the outcome measure; c_1^* is the correlation in mean outcomes between treatment and control group classrooms within grantees; and c_2^* is the

correlation in mean outcomes between treatment and control group classrooms within centers within grantees.

Random Assignment of Centers. In some designs, *centers* within grantees would be randomly selected to the intervention and control groups. This design would be chosen if the quality enhancement should be implemented center-wide because classroom spillover effects are likely, or because the quality enhancement will be more effective if everyone in the center is engaged in implementation. For example, teachers implementing strategies to promote positive social-emotional behavior in the classroom might be inclined to share those strategies with teachers in other classrooms who are having difficulties with children's behavior. Similarly, individual teachers might implement the behavioral strategies more successfully if they are able to discuss their experiences and methods with other teachers in the center.

The research design would call for random sampling of grantees and, possibly, of centers before random assignment. Because the intervention is fairly intensive to implement, the evaluation would be more cost-effective if, in every center that is randomly assigned, every classroom were included in the evaluation. In this case, clustering would occur only at the grantee and center levels, and the variance expression can be approximated as follows:

$$(5)\quad Var(impact) = \frac{2\sigma^2\rho_1(1-c_1^{**})}{g} + \frac{2\sigma^2\rho_2}{cg} + \frac{2\sigma^2(1-\rho_1-\rho_2)}{cgs},$$

where g is the total number of grantees in the sample; c is the average number of treatment or control centers per grantee; s is the average number of children per center; ρ_1 is the grantee-level intraclass correlation in the outcome measure; ρ_2 is the center-level intraclass correlation; σ^2 is the variance of the outcome measure; and c_1^{**} is the correlation in mean outcomes between treatment and control group centers within grantees.

An alternative design involving random assignment of centers might test alternative mathematics curricula, which, at Stage 3, could be implemented by distributing manuals, videotaped lessons for teachers, group activities to be conducted among all teachers in the center as part of professional development activities, and assistance from the Head Start Training and Technical Assistance network. For this evaluation, the design could call for random selection of a sample of grantees, random assignment of centers, and data collection in a randomly selected group of classrooms in each center. This design would involve clustering at the grantee, center, and classroom levels, and the variance expression can be approximated as follows:

$$(6)\quad Var(impact) = \frac{2\sigma^2\rho_1(1-c_1^{**})}{g} + \frac{2\sigma^2\rho_2}{cg} + \frac{2\sigma^2\rho_3}{cgl} + \frac{2\sigma^2(1-\rho_1-\rho_2-\rho_3)}{cgls},$$

where g is the total number of grantees in the sample; c is the average number of treatment or control centers per grantee; l is the average number of classrooms per center; s is the

average number of children per classroom; ρ_1 is the grantee-level intraclass correlation in the outcome measure; ρ_2 is the center-level intraclass correlation; ρ_3 is the classroom-level intraclass correlation; σ^2 is the variance of the outcome measure; and c_1^{**} is the correlation in mean outcomes between treatment and control group centers within grantees.

Random Assignment of Grantees. Some interventions are appropriately implemented at the grantee level. For example, a change in management strategies, such as technical assistance to programs to help the programs to use information from a variety of program data sources in order to assess and identify areas for quality improvement, would be most appropriately implemented among all grantees. Similarly, a management course for directors or a change in the way that education coordinators work with teachers would have grantee-wide effects.

For an evaluation of this type of quality enhancement, we would obtain a representative sample of grantees and would randomly assign the sample members to implement the quality enhancement or not. Although many centers potentially would be affected by the enhancement, the evaluation could focus on a random sample of centers and children within centers. Clustering would occur at the grantee level and, if centers are randomly selected for data collection, at the center level as well. For a design involving random assignment of grantees and a random sample of centers within grantees, the variance expression can be approximated as follows:

$$(7)\quad Var(impact) = \frac{2\sigma^2\rho_1}{g} + \frac{2\sigma^2\rho_2}{cg} + \frac{2\sigma^2(1-\rho_1-\rho_2)}{cgs},$$

where g is the average number of treatment or control grantees in the sample; c is the average number of centers per grantee; s is the average number of children per center; σ^2 is the variance of the outcome measure; ρ_1 is the grantee-level intraclass correlation in the outcome measure; and ρ_2 is the center-level intraclass correlation.

For a design involving random assignment of grantees with all centers and classrooms included in the evaluation, the variance expression can be approximated as follows:

$$(8)\quad Var(impact) = \frac{2\sigma^2\rho_1}{g} + \frac{2\sigma^2(1-\rho_1)}{gs},$$

where all terms are defined as above.

b. Variance Estimates for Stage 2 Designs

The variance calculations for Stage 2 designs are simpler than those for Stage 3 because they have to account for clustering only at the level of random assignment and, possibly, at another stage, if the sample is drawn randomly at that level. Table B.2 summarizes the various designs that we consider for Stage 2, provides an example of how the design might be used for a Stage 2 evaluation, and displays equation numbers for the variance formulas for each design.

Table B.2. Stage 2 Designs

Purposive Selection	Random (Representative) Selection	Sources of Clustering	Results Can Generalize to:	Equation Number for Variance Formula	Example from Report	Comments
Child-Level Random Assignment						
Grantee	None	None	All children in the selected grantees	9	Parent education intervention Mentoring by high school students	Limit evaluation to smaller grantees.
Grantee Center	None	None	All children in the selected centers	9	Parent education intervention Mentoring by high school students	Limit evaluation to a small number of centers within selected grantees.
Classroom- or Teacher-Level Random Assignment						
Grantee	Centers	Centers Classrooms	Children/classes in the selected grantees	11	Special curriculum Approach to working with children with behavioral issues	Random assignment at the classroom level introduces classroom-level clustering, so the power is not affected by the choice of (1) using all classrooms, or (2) using a randomly selected set of classrooms.
Grantee Center	None	Classrooms	Children/classes in the selected centers	10	Special curriculum Approach to working with children with behavioral issues	This design differs from the preceding one only in that centers are purposively selected, rather than randomly selected, thereby increasing power.
Center-Level Random Assignment						
Grantee	None	Centers	All children in the selected grantees	12	Alternative mathematics curriculum T/TA to use program data for quality improvements	Random assignment at the center level introduces center-level clustering, so power is not affected by the choice of (1) using all centers, or (2) using a randomly selected set of centers. For power, it would be better to select students for the sample without regard to classroom.
Grantee Center	None	Centers	All children in the selected centers	12	Alternative mathematics curriculum T/TA to use program data for quality improvements	The only reason to choose this design over the preceding one is to obtain a sample of volunteer centers. The power would not differ.
Grantee Center	Classrooms	Centers Classrooms	All children in the selected classrooms	13	Alternative mathematics curriculum T/TA to use program data for quality improvements	This design would be used if classrooms within centers are selected for observation. The power would be much less than under the previous design.

Table B.2 (continued)

Purposive Selection	Random (Representative) Selection	Sources of Clustering	Results Can Generalize to:	Equation Number for Variance Formula	Example from Report	Comments
Grantee-Level Random Assignment						
None	Grantee Center	Grantee Center	All Head Start children	15	T/TA to use program data for quality improvements	Children would be sampled from within centers. This design is unlikely to be used for Stage 2 because its sample size requirements are high. For most designs, it is not practical to deploy a quality enhancement within a full grantee, and to then sample some centers and/or children to measure outcomes.
None	Grantee	Grantee	All Head Start children	14	T/TA to use program data for quality improvements	This design has the same issues as the preceding one, although power would be somewhat better without center-level clustering.

T/TA = training and technical assistance.

In each of the Stage 2 designs listed, grantees are purposively selected, which is consistent with conducting these evaluations with programs that agree to participate. In some cases, all centers would be included in the evaluation; in others, centers might be purposively selected (volunteer) or might be randomly selected. Including all centers in the evaluation might be feasible if all of them agree to participate when the grantee agrees to participate; this scenario is more likely to occur when grantees include a small number of centers. Alternatively, centers that do not want to participate could decline, as their choice affects only the ability to generalize results to the entire grantee population, or only to centers included in the evaluation. Random selection of centers should occur only if centers are randomly assigned, as random selection at the center level with random assignment at a different level would unnecessarily add to the sample size requirements for a Stage 2 evaluation.

Similar principles can be identified for selection of classrooms within centers. If random assignment is conducted at the classroom level, clustering at the classroom level must be included in variance calculations, and, therefore, power will not be affected by the choice of whether or not to sample classrooms. However, if random assignment occurs at the center or grantee level, one should avoid adding a level of clustering by randomly selecting classrooms. Instead, the best choice from the perspective of maximizing power is to include all classrooms (within the designated grantees and centers) in the evaluation, or by selecting children for the study without regard to classroom.

Random Assignment of Children. For some Stage 2 designs, children (or families) could be randomly assigned to a quality enhancement that focuses on individuals. A mentoring program for children and an intensive parent education program are examples of this type of quality enhancement.

In these designs, clustering can be avoided by purposively selecting grantees and centers, and then selecting children for the study from a center-wide list (without regard to classroom). The variance for the impact estimates across centers can be approximated as follows:

$$(9)\quad Var(impact) = \frac{2\sigma^2}{cs},$$

where c is the total number of centers in the sample, σ^2 is the variance of the outcome measure, and s is the average number of treatment or control children per center.

Random Assignment of Classrooms Within Centers. Some Stage 2 designs could involve classroom-level random assignment. For example, an evaluation might test a computer-based language curriculum in randomly assigned classrooms. This type of design is appropriate for interventions that are administered at the classroom level and for which potential spillover effects are deemed to be small.

In these designs, clustering is at the classroom level, but additional clustering can be avoided by purposively selecting grantees and centers for the study. Intuitively, if sampling

were repeated, a different random allocation of classrooms would be selected to the treatment and control groups. Hence, the variance expressions must account for the extent to which mean child outcomes vary across classrooms. If grantees and centers agree to participate in the study, then center and grantee effects should be treated as fixed. Accordingly, the variance formula for these impact estimates can be expressed as follows:

$$(10)\quad Var(impact) = \frac{2\sigma^2\rho_3}{cl} + \frac{2\sigma^2(1-\rho_3)}{cls},$$

where c is the total number of centers in the sample, l is the average number of treatment (control) classrooms per center, s is the average number of children per classroom, σ^2 is the variance of the outcome measure, and ρ_3 is the intra-classroom variance as a proportion of the total variance.

If centers are randomly selected to participate in the evaluation and classrooms are randomly assigned to the quality enhancement or control groups, then clustering would occur at the center and classroom levels. Under this design, the variance expression would be approximated by:

$$(11)\quad Var(impact) = \frac{2\sigma^2\rho_2(1-c_2^*)}{c} + \frac{2\sigma^2\rho_3}{cl} + \frac{2\sigma^2(1-\rho_2-\rho_3)}{cls},$$

where c is the average number of centers in the sample; l is the average number of treatment or control classrooms per center; s is the average number of children per classroom; ρ_2 is the center-level intraclass correlation; ρ_3 is the classroom-level intraclass correlation; σ^2 is the variance of the outcome measure; and c_2^* is the correlation in mean outcomes between treatment and control group classrooms within centers. Because the additional center-level clustering term would add to the variance of the impact estimates, this design would require a larger sample than the previous design.

Random Assignment of Centers. Other Stage 2 designs might involve random assignment of centers to different quality enhancements. For example, an approach to addressing children's behavioral problems by working with the children and families with the support of a behavioral specialist might be most efficiently implemented center-wide.

When centers are randomly assigned as part of a Stage 2 design, grantees would be voluntary participants (purposively selected). Although centers likely would be volunteers as well, random assignment at the center level would mean that the power would not be affected if centers were randomly selected for the evaluation instead. Whether or not classrooms are randomly selected for the evaluation would affect power, so we discuss the alternatives here.

Under one design option, classrooms would not be sampled within the treatment and control group centers. For this option, either *all* relevant classrooms in the selected centers are included in the research sample or children are sampled directly to the research sample without regard to their classrooms. In practice, the researcher would be choosing between

inclusion of all consenting children from each center in the research sample or the random selection of children for the research sample from the pool of children for whom consent had been given. For both options, clustering occurs at the center level, but not at the classroom level. Intuitively, if sampling were repeated, a different random allocation of centers would be selected to the treatment and control groups, but not a different set of classrooms within centers. Consequently, the variance of an impact estimate can be expressed as follows:

$$(12)\quad Var(impact) = \frac{2\sigma^2\rho_2}{c} + \frac{2\sigma^2(1-\rho_2)}{cs},$$

where c is the average number of treatment or control centers; s is the average number of children per center; σ^2 is the variance of the outcome measure; and ρ_2 is the center-level intraclass correlation in the outcome measure.

For a center-based experimental evaluation, another design option that conserves project resources is to sample classrooms within the study centers. In this case, clustering occurs at both the center and classroom levels. In the presence of both center- and classroom-level clustering, the variance formula can be expressed as follows:

$$(13)\quad Var(impact) = \frac{2\sigma^2\rho_2}{c} + \frac{2\sigma^2\rho_3}{cl} + \frac{2\sigma^2(1-\rho_2-\rho_3)}{cls},$$

where c is the average number of treatment or control centers per grantee; l is the average number of classrooms per center; s is the average number of children per classroom; σ^2 is the variance of the outcome measure; ρ_2 is the center-level intraclass correlation in the outcome measure; and ρ_3 is the classroom-level intraclass correlation. The addition of a center-level clustering term to the variance calculation increases the sample size requirements of this design relative to the previous one.

Random Assignment of Grantees. For some evaluation designs, the quality enhancement would be implemented at the grantee level. For example, a change in approach to managing the Head Start program would have grantee-wide effects. Nevertheless, the sample size requirements for the evaluation of this option are likely to be comparable to those of a Stage 3 evaluation.

$$(14)\quad Var(impact) = \frac{2\sigma^2\rho_1}{g} + \frac{2\sigma^2(1-\rho_1)}{gs},$$

where g is the average number of treatment or control grantees in the sample; s is the average number of children per center; σ^2 is the variance of the outcome measure; and ρ_1 is the grantee-level intraclass correlation in the outcome measure.

If grantees are randomly assigned and the evaluation includes only a random sample of centers within each of the grantees, the variance formula can be expressed as follows:

$$(15)\quad Var(impact) = \frac{2\sigma^2\rho_1}{g} + \frac{2\sigma^2\rho_2}{cg} + \frac{2\sigma^2(1-\rho_1-\rho_2)}{cgs},$$

where g is the average number of treatment or control grantees in the sample; c is the average number of centers per grantee; s is the average number of children per center; σ^2 is the variance of the outcome measure; ρ_1 is the grantee-level intraclass correlation in the outcome measure; and ρ_2 is the center-level intraclass correlation.

4. Estimating Correlations

Before we can calculate the sample sizes required for the various evaluation designs, we must obtain estimates of the key parameters in the variance formulas we have presented in this appendix. In particular, we must have estimates for the following correlations that enter into the variance formulas:

- ρ_1 = the extent to which mean outcomes differ across grantees (that is, the intraclass correlation at the grantee level)
- ρ_2 = the extent to which mean outcomes differ across centers within grantees (that is, the intraclass correlation at the center level)
- ρ_3 = the extent to which mean outcomes differ across classrooms within centers (that is, the intraclass correlation at the classroom level)
- c_1 = the correlation between the mean outcomes of treatment and control group children within grantees
- c_2 = the correlation between the mean outcomes of treatment and control group children within centers in grantees
- c_1^* = the correlation between the mean outcomes of treatment and control group classrooms within grantees
- c_2^* = the correlation between the mean outcomes of treatment and control group classrooms within centers in grantees
- c_1^{**} = the correlation between the mean outcomes of treatment and control group centers within grantees

We discuss our estimates of these parameters in the following section.

a. Intraclass Correlations

To obtain estimates for the intraclass correlations, we used data from the Head Start Family and Child Experiences Survey (FACES), 2000 cohort. The Head Start FACES data are a representative sample of children and classrooms selected by multi-stage random sampling of grantees, centers, and classrooms. In many cases, the samples of children per classroom and of classrooms per center generally are small, but they nevertheless are representative of Head Start classrooms. To obtain as large a sample as possible, we

included all children in the sample (ages 3 to 5 years) who had valid scores on the Peabody Picture Vocabulary Test (a measure of children's receptive vocabulary) for the spring (end-of-Head Start year) assessments. Data from the Head Start National Reporting System would be ideal for estimating these correlations, but they are not readily available for this analysis.

Our analysis indicates that ρ_1, the grantee-level intraclass correlation, is 0.247; ρ_2, the center-level intraclass correlation, is 0.056, and ρ_3, the classroom-level intraclass correlation, is 0.073. These estimates indicate that child assessment scores are more divergent across grantees than they are across centers within a single grantee or across classrooms within a center. Essentially, there is no sorting by ability across Head Start classrooms or across centers within a grantee. Instead, all classrooms and centers include children with the same variation in assessment scores as exists across the grantee as a whole. However, from one grantee to another, there are greater differences in assessment scores. Thus, the populations served by each Head Start grantee seem to vary, whereas, within grantees, the centers and classrooms look similar. Stratified random sampling would generate lower grantee-level intraclass correlations within strata, which would increase the power of designs that involve grantee-level clustering.[1]

The intraclass correlation at the classroom level (0.073) is slightly higher than at the center level (0.056), suggesting that teacher effects—the influence on scores of individual strong teachers—are stronger than are center effects. The same pattern of correlations at the classroom (teacher) and school levels are found in studies of older children.

b. Correlation Between Treatment and Control Group Means

The correlations between treatment and control group means ideally should be estimated from evaluation data, and having more than one evaluation on which to base these estimates would be very useful. For example, data from the Head Start Impact Study, in which children were randomly assigned to Head Start programs or to a control group, many of whom received alternative center-based child care, could help to estimate values for c_1 and c_2.[2] Data from the Preschool Curriculum Evaluation Research (PCER) evaluation, in which most sites randomly assigned classrooms to an enhanced curriculum or to regular services, would provide a useful source of information for estimating c_1 and c_2. We are not aware of any evaluations that have randomly assigned early childhood centers to an intervention or control group.

[1] The increase in power due to stratification of the sample will be offset in part by a reduction in degrees of freedom that depends on the number of strata.

[2] The Head Start Impact Study is not a perfect source of information for these parameters because it did not involve a comparison of children randomly assigned to enhanced Head Start services and regular Head Start services. However, the child-level random assignment to different types of preschool program services should be informative about values for c_1 and c_2.

We do not currently have access to data from the Head Start Impact Study or from PCER to help to generate estimates of these parameters. However, we have estimated a value for c_1 using data from the Early Head Start evaluation, in which families with a pregnant woman or an infant in 17 sites were assigned to Early Head Start or a control group that did not receive Early Head Start services. The estimate for c_1, using data from the three-year-old followup, is 0.5. Additional refinement of these assumptions will be possible as data become available from evaluations using classroom- or center-level random assignment.

B. ESTIMATED SAMPLE SIZES UNDER ALTERNATIVE STAGE 2 AND STAGE 3 DESIGNS

Tables B.3–B.7 present estimated sample sizes for the Stage 2 and Stage 3 designs that we have discussed in Section A. The following assumptions underlie all of the sample size estimates:

- ***A two-tailed test at 80 percent power and a 5 percent significance level.*** We adopt a two-tailed test, rather than a one-tailed test, because we want to be able to measure any inadvertently negative effects of the enhancements, as well as the positive impacts.
- ***Children are divided equally between treatment and control groups.*** Unbalanced assignment reduces the precision of the estimates.
- ***Interview nonresponse is 10 percent.*** Ninety percent of the initial sample is available for analysis at the spring followup because of the level of interview nonresponse.
- ***One enhancement is compared with regular Head Start services.*** If additional enhancements are to be compared with one another or with a control group, the sample sizes would have to increase accordingly. The tables show the number of units (grantees, centers, or classrooms) in the treatment group. These numbers must be doubled if one enhancement is compared with a control group, tripled if two enhancements are evaluated against a control group, and so on.
- ***Intraclass correlations (ρ_1, ρ_2 and ρ_3) are based on calculations from FACES 2000 data.*** We assume that the grantee-level intraclass correlation (ρ_1) is 0.123, which is lower than estimated from FACES data because we would use stratified random sampling, rather than sampling grantees from the full universe of Head Start programs. We use the FACES estimates for the center-level intraclass correlation ($\rho_2 = 0.056$) and the classroom-level intraclass correlation ($\rho_3 = 0.073$). In future work, more FACES samples could be used to estimate these correlations, although the FACES sample is small for estimating some of these correlations. Samples drawn from the National Reporting System data would be ideal, as a sufficient sample of centers within grantees classrooms within centers, and children within classrooms could be obtained.

- ***Correlation between treatment and control group outcomes.*** For child-level correlations, we assume a value of .30 for the correlation between treatment and control group children within grantees, and a value of .50 for the correlation between the mean outcomes of treatment and control group children within centers in grantees. For classroom-level correlations, we assume a value of .30 for the correlation between mean outcomes of treatment and control group classrooms within grantees, and a value of .15 for the correlation between the mean outcomes of treatment and control group classrooms within centers in grantees. For the center-level correlation, we assume a value of .10 for the correlation between mean outcomes of treatment and control group centers within grantees.

- ***We assume an average of four centers per grantee, two classes per center, 15 children per center, and 10 children per class.*** These assumptions are consistent with information about center-based Head Start programs from the Head Start Program Information Report (PIR) that indicates the average number of centers per grantee is 8; the median number of classrooms per center is 3; the median number of children per classroom is 18; and the median number of children per center is approximately 50.[3]

In addition, we present the estimates for alterative assumptions that affect the power of the sample to detect impacts:

- ***Alternative levels of minimum detectable effect sizes.*** To address the issue that smaller effect sizes are plausible for improvements in Head Start practice, yet smaller effect sizes require larger samples, we show the minimum sample sizes necessary to detect effect sizes of 0.1, 0.2, 0.25, and 0.33.

- ***The regression R^2 value.*** Regression-adjusted impact estimates have greater precision because the regression adjustment eliminates some of the variability in the intervention and control group mean outcomes. If impacts are estimated without any regression-adjustment, then the R^2 is zero, and the estimated impact is the simple difference in means. If we have some baseline information on family demographics (for example, the mother's education level), then the R^2 typically is about 0.2. If we have baseline information on the same (or similar) outcome variables as those measured at followup, then the R^2 is likely to be approximately 0.5. Therefore, we use the three values for R^2 to represent plausible alternative levels of information available for the impact analysis.

[3] When classrooms are randomly assigned and the center has only 2 or 3 classrooms, one classroom is assigned to the enhancement group and one to the control group. Under this design, there are not enough degrees of freedom to estimate classroom effects (since there is only one classroom in each research group). To estimate impacts, classrooms will need to be pooled across centers and centers will need to be treated as if they were randomly selected.

Table B.3. Required Sample Sizes of Grantees, Centers, and Children to Detect Target Minimum Detectable Effect Sizes When Grantees Are Randomly Assigned, Stage 3

	Total Number of Grantees	Enhancement Grantees	Total Number of Centers	Total Number of Children (15 per center)	Initial Sample of Children
MDE = 0.1					
R^2 = 0	474	237	1,896	28,440	31,600
R^2 = 0.2	378	189	1,512	22,680	25,200
R^2 = 0.5	236	118	944	14,160	15,733
MDE = 0.2					
R^2 = 0	118	59	472	7,080	7,867
R^2 = 0.2	94	47	376	5,640	6,267
R^2 = 0.5	60	30	240	3,600	4,000
MDE = 0.25					
R^2 = 0	76	38	304	4,560	5,067
R^2 = 0.2	60	30	240	3,600	4,000

Source: Authors' calculations.

Note: The sample size calculations assume a two-tailed test of statistical significance at 80 percent power and a 5 percent significance level; grantees are equally divided among treatment and control groups; a 90 percent response rate to the follow-up interview; intraclass correlations and correlations between treatment and control groups as described in the text; four centers per grantee; and 15 children randomly selected from each center for the research sample. The formula used to calculate sample sizes is:

$$MDE = 2.802 * \sqrt{(1 - R^2) Var(impact)} / \sigma,$$

where:

$$Var(impact) = \frac{2\sigma^2\rho_1}{g} + \frac{2\sigma^2\rho_2}{cg} + \frac{2\sigma^2(1-\rho_1-\rho_2)}{cgs},$$

R^2 is the regression R-squared value; g is the total number of grantees in the sample; c is the average number of centers per grantee; and s is the average number of children per center; ρ_1 is the grantee-level intraclass correlation in the outcome measure; ρ_2 is the center-level intraclass correlation; σ^2 is the variance of the outcome measure.

We have omitted examples that result in estimated sample sizes below 30 grantees, as the group mean estimates would be unstable if the group has fewer than 30 members.

Table B.4. Required Sample Sizes of Centers and Children to Detect Target Minimum Detectable Effect Sizes When Centers Are Randomly Assigned, Stage 3

	Total Number of Centers	Enhancement Centers	Total Number of Children (15 per center)	Initial Sample of Children
MDE = 0.1				
R^2 = 0	1,043	522	15,646	17,384
R^2 = 0.2	834	417	12,517	13,907
R^2 = 0.5	522	261	7,823	8,692
MDE = 0.2				
R^2 = 0	261	130	3,911	4,346
R^2 = 0.2	209	104	3,129	3,477
R^2 = 0.5	130	65	1,956	2,173
MDE = 0.25				
R^2 = 0	167	83	2,503	2,781
R^2 = 0.2	134	67	2,003	2,225
R^2 = 0.5	83	42	1,252	1,391
MDE = 0.33				
R^2 = 0	96	48	1,437	1,596
R^2 = 0.2	77	38	1,149	1,277

Source: Authors' calculations.

Note: The sample size calculations assume a two-tailed test of statistical significance at 80 percent power and a 5 percent significance level; children are equally divided among treatment and control groups; a 90 percent response rate to the follow-up interview; intraclass correlations and correlations between treatment and control groups as described in the text; four centers per grantee, with two assigned to the enhancement group and two to the control group; and 15 children randomly selected from each center for the research sample. The formula used to calculate sample sizes is:

$$MDE = 2.802 * \sqrt{(1-R^2)Var(impact)} / \sigma,$$

where:

$$Var(impact) = \frac{2\sigma^2\rho_1(1-c_1{}^{**})}{g} + \frac{2\sigma^2\rho_2}{cg} + \frac{2\sigma^2(1-\rho_1-\rho_2)}{cgs},$$

R^2 is the regression R-squared value; g is the total number of grantees in the sample; c is the average number of treatment or control centers per grantee; s is the average number of children per center; ρ_1 is the grantee-level intraclass correlation in the outcome measure; ρ_2 is the center-level intraclass correlation; σ^2 is the variance of the outcome measure; and c_1^{**} is the correlation in mean outcomes between treatment and control group centers within grantees.

We have omitted examples that result in estimated sample sizes below 30 centers, as the group mean estimates would be unstable if the group has fewer than 30 members.

Table B.5. Required Sample Sizes of Centers, Classrooms, and Children to Detect Target Minimum Detectable Effect Sizes When Classrooms Are Randomly Assigned, Stage 3

	Total Number of Centers	Total Number of Classrooms	Total Number of Children (10 per class)	Initial Sample of Children
MDE = 0.1				
R^2 = 0	848	1,695	16,952	18,836
R^2 = 0.2	678	1,356	13,562	15,069
R^2 = 0.5	424	848	8,476	9,418
MDE = 0.2				
R^2 = 0	212	424	4,238	4,709
R^2 = 0.2	170	339	3,390	3,767
R^2 = 0.5	106	212	2,119	2,354
MDE = 0.25				
R^2 = 0	136	271	2,712	3,014
R^2 = 0.2	108	217	2,170	2,411
R^2 = 0.5	68	136	1,356	1,507
MDE = 0.33				
R^2 = 0	78	156	1,557	1,730
R^2 = 0.2	62	125	1,245	1,384
R^2 = 0.5	39	78	778	865

Source: Authors' calculations.

Note: The sample size calculations assume a two-tailed test of statistical significance at 80 percent power and a 5 percent significance level; children are equally divided among treatment and control groups; a 90 percent response rate to the follow-up interview; intraclass correlations and correlations between treatment and control groups as described in the text; four centers per grantee; two classrooms per center, with one assigned to the enhancement group and one to the control group; and 10 children randomly selected from each classroom for the research sample. The formula used to calculate sample sizes is:

$$MDE = 2.802 * \sqrt{(1-R^2)Var(impact)} / \sigma,$$

where:

$$Var(impact) = \frac{2\sigma^2\rho_1(1-c_1{}^*)}{g} + \frac{2\sigma^2\rho_2(1-c_2{}^*)}{cg} + \frac{2\sigma^2\rho_3}{cgl} + \frac{2\sigma^2(1-\rho_1-\rho_2-\rho_3)}{cgls},$$

R^2 is the regression R-squared value; g is the total number of grantees in the sample; c is the average number of centers per grantee; l is the average number of treatment or control classrooms per center; and s is the average number of children per classroom; ρ_1 is the grantee-level intraclass correlation in the outcome measure; ρ_2 is the center-level intraclass correlation; ρ_3 is the classroom-level intraclass correlation; σ^2 is the variance of the outcome measure; c_1^* is the correlation in mean outcomes between treatment and control group classrooms within grantees; and c_2^* is the correlation in mean outcomes between treatment and control group classrooms within centers within grantees.

Table B.6. Required Sample Sizes of Centers and Children to Detect Target Minimum Detectable Effect Sizes When Centers Are Randomly Assigned, Stage 2

	Total Number of Centers	Enhancement Centers	Total Number of Children (15 per center)	Initial Sample of Children
MDE = 0.1				
R^2 = 0	374	187	5,603	6,225
R^2 = 0.2	299	149	4,482	4,980
R^2 = 0.5	187	93	2,801	3,113
MDE = 0.2				
R^2 = 0	93	47	1,401	1,556
R^2 = 0.2	75	37	1,121	1,245
MDE = 0.25				
R^2 = 0	60	30	896	996

Source: Authors' calculations.

Note: The sample size calculations assume a two-tailed test of statistical significance at 80 percent power and a 5 percent significance level; children are equally divided among treatment and control groups; a 90 percent response rate to the follow-up interview; intraclass correlations and correlations between treatment and control groups as described in the text; four centers per grantee, with two assigned to the enhancement group and two to the control group; and 15 children randomly selected from each center for the research sample. The formula used to calculate sample sizes is:

$$MDE = 2.802 * \sqrt{(1 - R^2) Var(impact)} / \sigma,$$

where:

$$Var(impact) = \frac{2\sigma^2 \rho_2}{c} + \frac{2\sigma^2 (1 - \rho_2)}{cs},$$

R^2 is the regression R-squared value; c is the average number of treatment or control centers; s is the average number of children per center; σ^2 is the variance of the outcome measure, and ρ_2 is the center-level intraclass correlation in the outcome measure.

We have omitted examples that result in estimated sample sizes below 30 centers, as the group mean estimates would be unstable if the group has fewer than 30 members.

Table B.7. Required Sample Sizes of Centers, Classes, and Children to Detect Target Minimum Detectable Effect Sizes When Classrooms Are Randomly Assigned, Stage 2

	Total Number of Centers	Enhancement Classrooms	Total Number of Children (10 per class)	Initial Sample of Children
MDE = 0.1				
R^2 = 0	260	260	5,204	5,782
R^2 = 0.2	208	208	4,163	4,626
R^2 = 0.5	130	130	2,602	2,891
MDE = 0.2				
R^2 = 0	65	65	1,301	1,445
R^2 = 0.2	52	52	1,041	1,156
R^2 = 0.5	33	33	650	723
MDE = 0.25				
R^2 = 0	42	42	833	925
R^2 = 0.2	33	33	666	740

Source: Authors' calculations.

Note: The sample size calculations assume a two-tailed test of statistical significance at 80 percent power and a 5 percent significance level; children are equally divided among treatment and control groups; a 90 percent response rate to the follow-up interview; intraclass correlations and correlations between treatment and control groups as described in the text; four centers per grantee; two classrooms per center, with one assigned to the enhancement group and one to the control group; and 10 children randomly selected from each classroom for the research sample. The formula used to calculate sample sizes is:

$$MDE = 2.802 * \sqrt{(1 - R^2) Var(impact)} / \sigma,$$

where:

$$Var(impact) = \frac{2\sigma^2 \rho_3}{cl} + \frac{2\sigma^2 (1 - \rho_3)}{cls},$$

R^2 is the regression R-squared value; c is the total number of centers in the sample, l is the average number of treatment (control) classrooms per center, s is the average number of children per classroom, σ^2 is the variance of the outcome measure, and ρ_3 is the between-classroom variance as a proportion of the total variance.

We have omitted examples that result in estimated sample sizes below 30 classrooms, as the group mean estimates would be unstable if the group has fewer than 30 members.