Evaluation of Child Care Subsidy Strategies

Findings from an Experimental Test of Three Language/ Literacy Interventions in Child Care Centers in Miami-Dade County

Final Report

January 2009

Prepared for
Ivelisse Martinez-Beck
U.S. Dept. of Health and
Human Services
Administration for Children and Families
Office of Planning, Research and Evaluation
370 L'Enfant Plaza Promenade, SW
Washington, DC 20447

Prepared by
Jean I. Layzer
Carolyn J. Layzer
Barbara D. Goodson
Cristofer Price

Abt Associates Inc.
55 Wheeler Street
Cambridge, MA 02138

Contents

Introduction

This report presents findings from Project Upgrade, one of four experiments conducted as part of the Evaluation of Child Care Subsidy Strategies. Recognizing the need for information that would help states and communities allocate their child care subsidy funds as effectively as possible, the Child Care Bureau and the Office for Planning, Research and Evaluation (OPRE) of the Administration for Children and Families within the US Department of Health and Human Services launched this major study in 2001. The study is being conducted by Abt Associates Inc, with its research partners MDRC and the National Center for Children in Poverty of Columbia University.

The evaluation is a multi-site, multi-year effort to determine whether and how different child care subsidy policies and procedures and quality improvement efforts help low-income parents obtain and hold onto jobs and improve outcomes for children. Study staff worked with states and communities across the country to identify significant issues and develop hypotheses about the use of child care subsidy funds that could be rigorously tested in a series of experiments. A guiding principle of the study was that state (or community) interests and preferences should play a large role in the choice of research topics and strategies.

The funds that flow to states through the Child Care and Development Fund (CCDF), administered at the federal level by the Child Care Bureau, have two purposes. The major portion of the funds provides subsidies for child care for children of low-income working parents whose eligibility is determined by states within broad federal guidelines. A small percentage of the federal funds (4%) is set aside, with state matching funds added, to improve the quality of child care for all children. It was the expressed intention of the Child Care Bureau that the study generate a set of experiments that examined aspects of the use of both types of funds.

While some states expressed interest in testing some alternative policies governing the use of direct service dollars, many more were concerned about the effectiveness of their current use of funds intended to improve child care quality. Ultimately, study staff working closely with state and local staff, implemented four experiments, two that are testing alternative subsidy policies and two that test approaches to the use of quality set-aside funds. Project Upgrade in Miami-Dade County falls into the latter group of experiments.

Summary of Design and Findings

Project Upgrade was a two-year experimental test of the effectiveness of three different language and literacy interventions, implemented in child care centers in Miami-Dade County that served children from low-income families. One hundred and sixty-four centers were randomly assigned to one of three research-based curricula or to a control group that continued with its existing program. The curricula, while grounded in a common set of research findings, differed in intensity, pedagogic strategies and use of technology. In each center, one classroom that served four-year-old children was selected for the study. Teachers and aides assigned to the three treatment groups received initial and follow-up training as well as ongoing mentoring over a period of approximately 18 months, from Fall 2003 to Spring 2005. All classrooms in the study, whether treatment or control, received an initial package of literacy materials (paper, crayons, books, tape recorders, books on tape etc.). To reduce staff turnover, teachers in all four groups who remained in centers received $500 in July, at the end of each year of the study.

The hypotheses tested by the study stipulated two kinds of outcomes: teacher behavior and interactions with children, and aspects of the classroom environment that support children's language and literacy development, measured through direct observation; and children's language and pre-literacy skills, measured by their performance on a standardized assessment. Study staff conducted classroom observations in Fall 2003, Spring 2004 and Spring 2005. Four-year-old children in the study classrooms were assessed in Spring 2005.

Key findings are summarized below and in Exhibit 1. Here, and in the body of the paper, impacts are described in terms of effect sizes. Effect sizes are standardized measures of the magnitude (size) of treatment effects. For each outcome measure, the effect size is equal to the estimated impact of the treatment, divided by the control group standard deviation (a measure of the variation in scores within the group). The standardization makes possible a comparison of the size of treatment effects across studies and, within limits, across outcome measures.[1] For example, if the effect sizes of a treatment on outcome measures A and B are 0.50, and 0.25, respectively, then the size of the treatment impact on A is considered to be twice the size of the impact on B.

Findings

- The initial observations, conducted before the interventions, showed that, across all groups, teachers engaged in few of the behaviors and interactions that have been shown to support children's development of language and literacy skills.
- Within six months of training, in Spring 2004, all three language/literacy interventions produced significant impacts on teacher behaviors and interactions with children that supported their language and literacy development; by Spring 2005, these impacts were generally more pronounced, and there were significant impacts on the number of classroom activities that involved literacy, and on literacy resources in the classroom.

[1] Comparisons across studies must be approached cautiously. Even if the same outcome measure is used, the comparison assumes that the two study samples have similar standard deviations. Comparison of effect sizes for very different outcome measures may be misleading.

- The interventions had significant positive impacts on teacher behavior. These impacts were generally stronger for teachers whose primary language was Spanish than for their English-speaking counterparts.
- Two of the three interventions, ***Ready, Set, Leap*** and ***Breakthrough to Literacy***, had significant impacts on all four measures of emergent literacy outcomes for children: definitional vocabulary; phonological awareness; knowledge and understanding and the overall index of early literacy. The impact of the two effective interventions was much greater for children in classrooms with Spanish-speaking teachers than for children in classrooms with English-speaking teachers.
- The two interventions that had impacts on child outcomes brought children close to or above the national norms on three of the four outcomes. On the fourth, although children in the two treatment groups had significantly higher scores, they still lagged considerably behind the national norms. The impacts represent between four and nine months of developmental growth, depending on the outcome.
- The interventions resulted in a substantial increase in the time spent on language and literacy activities, both teacher-directed and child-initiated. This did not eliminate other important developmental activities. Rather, time spent on each of the other activities was reduced slightly.
- There was a small but significant relationship between teachers' educational attainment and some aspects of their behavior with children ***before*** the interventions. The training and on-going mentoring provided as an integral part of the interventions eliminated this effect. That is, as a result of the training and mentoring, less-educated teachers looked remarkably similar to their better-educated counterparts in the extent to which they provided activities that supported literacy. Teachers' educational qualifications did not modify the ***impacts*** of the interventions on child outcomes.

Exhibit 1

Key Impact Findings

Domain/Construct (measure)	All Teachers Effect size	Spanish-dominant Teachers Effect size	English-dominant Teachers Effect size
Teacher behavior (OMLIT, 2005)			
Support for Oral Language	.61***	.63**	.55*
Support for Phonological Awareness	.49**	.43*	.52*
Support for Print Knowledge	.74***	.90**	.54*
Support for Print Motivation	.43**	.59*	ns
Classroom literacy environment (OMLIT, 2005)			
Literacy Resources	.28*	ns	ns
Literacy Activities	.80***	.80***	.77**
	All children **Effect Size**	**Children in Classrooms with Spanish-dominant Teachers** **Effect Size**	**Children in Classrooms with English-dominant Teachers** **Effect size**
Child language and emergent literacy (TOPEL, Spring 2005)[2]			
Definitional Vocabulary	.30***	.39**	ns
Phonological Awareness	.39 ***	.55 ***	ns
Print Knowledge	.63***	.86 ***	.41**
Early Literacy Index	.53 ***	.72 ***	.36**

*** = p<.001, ** = p<.01, * = p<.05

[2] Outcomes shown are combined outcomes for the two interventions that showed significant impacts. Results for the two treatments were combined since they were very similar and to provide additional statistical power. Outcomes for the individual curricula are shown separately later in the paper and in the attached tables.

Chapter One: Policy and Research Context for the Study

In April 2002, President Bush introduced the *Good Start, Grow Smart* initiative, which includes a Federal-State partnership to create linkages between the Child Care and Development Fund (CCDF), the vehicle through which child care subsidy funds are allocated to states, and state and private efforts to promote early learning. The initiative reflected the understanding that, while many children from low-income families participate in Head Start or a state-funded prekindergarten program intended to enhance their readiness for school, this goal may not have received similar attention in child care programs that support the work-related needs of low-income parents.

In Florida, the Agency for Workforce Innovation (AWI)'s Office of Early Learning administers CCDF and state funds for child care and quality enhancement, and requires annual assessment of children who receive subsidies as well as annual plans for the use of quality dollars to effect improvements in children's school readiness. As Abt staff explored potential experiments with AWI staff, they were referred to Miami-Dade County (as well as several other counties) where the mandated child assessments had revealed large gaps in the language skills of subsidized four-year-olds. In Miami-Dade County, the Early Learning Coalition (ELC)[3] acts as the county's fiscal agent for CCDF subsidy and quality improvement funds. In response to the President's initiative and AWI's requirements, the ELC embarked on an effort to improve the school readiness of low-income children. In the first phase of this effort (Spring 2003), the ELC commissioned developmental assessments of all four-year-old children who were receiving subsidies.[4] In a subsequent phase, the coalition's intent was to put in place system-wide curriculum interventions that focused on the developmental gaps identified by the assessments.

The first round of assessments of four-year-olds, using a broad-based diagnostic tool, the Learning Accomplishment Profile-Diagnostic Assessment (LAP-D), indicated a serious lag in one of the areas tested -- children's language development. For that reason, the ELC's stakeholder advisory committee recommended that program interventions focus on language development and early literacy. Working closely with staff at the ELC, the central agencies that administered child care subsidies and Florida International University, staff from Abt Associates and MDRC developed a plan for an experimental test of three language and literacy curricula in child care centers serving low-income children in Miami-Dade County. The coalition agreed to commit CCDF quality improvement funds to pay for the curricula and the associated training. In addition, quality funds were allocated to hire literacy mentors who would provide ongoing support for teachers who were implementing the curricula. In return, the coalition hoped that the study would provide strong evidence about the effectiveness of the interventions that would guide the system-wide implementation of one or more curricula.

Miami-Dade County is Florida's largest and most populous county, and is the eighth largest county in the United States, with a population of almost 2.4 million. It has experienced continuous and rapid population growth since the early part of the last century. Two-thirds of population growth is attributable

[3] Before 2005, the agency was named the School Readiness Coalition. In 2005, it was renamed the Early Learning Coalition of Miami-Dade and Monroe Counties.

[4] The assessment of **subsidized** four-year-olds in 2003 was state-mandated. In subsequent years, the Coalition mandated that all four-year-olds in centers that served subsidized children be assessed with the Learning Accomplishment Profile – Diagnostic Assessment (LAP-D).

to migration, most of it from Cuba and other Caribbean and Central American countries. In 2001, over half the county's residents were born outside the United States. The county is ethnically and linguistically diverse: Hispanics constitute a majority (57%), non-Hispanic Whites are 24% and non-Hispanic Blacks are about 19% of Miami-Dade County's population. Many segments of the population are highly mobile, although much of the movement is within the county.

The child care system in the county poses challenges to the implementation of high-quality early childhood education. Florida's licensing requirements are not stringent, turnover of teachers and staff is high, in large part because of low wages, and many classroom staff have low levels of educational achievement. The high levels of mobility among low-income families make stable child care arrangements difficult. These challenges, while they may differ in degree, are also found in many large US cities. A successful intervention in Miami-Dade County could provide guidance for many communities beyond its borders.

Research Context

This experiment focuses specifically on the development of language and emergent literacy skills. This focus reflects the ELC's concern about serious delays in language development among low-income four-year-olds in the county. It was also influenced by the increasing emphasis in the last decade on the importance of early language and literacy development for later reading success, which itself is seen as the foundation for learning. Research on child development and emergent literacy has identified four key domains that are strong predictors of subsequent literacy development: oral language development, phonological sensitivity (sensitivity to the sounds of language, including phonemes), print knowledge (including concepts of print and alphabet knowledge), and print motivation (Dickinson & Tabors, 2001; Lonigan, Burgess, and Anthony, 2000; Whitehurst & Lonigan, 1998; 2001).

Also over the last decade, there has been growing recognition of the important role early childhood care and education programs can play in promoting these skills in children, especially at-risk children. The National Association for the Education of Young Children (NAEYC) has reversed its earlier position on direct literacy instruction in response to three decades of research that provides evidence about the importance of early support for children's language growth, engagement with print materials, and literacy-related activities (National Research Council, 1999; Neuman, Copple, & Bredekamp, 2000; Neuman & Roskos, 1998).

At the same time, these and other research efforts have identified practices through which early childhood educators can support these outcomes for children.

Oral Language. Research has identified a number of effective practices for supporting children's language development. These include:

- Two to three read-alouds daily in a full-day program, to broaden children's knowledge and vocabulary, and to build listening and comprehension skills;
- Extended, cognitively challenging conversations between children and adults;
- Adult use of open-ended questions;
- Adult scaffolding of children's language, including questions and prompts to extend children's language, listening and giving children time to respond, expanding children's ideas, providing feedback to encourage, interpret, and evaluate children's responses;

- Opportunities for children to use language to communicate thoughts and ideas, using complex sentences and vocabulary;
- Rich language by adults; and
- Pretend play and pretend talk among children.

Children should be exposed to daily reading of high-quality books, preferably in small groups or one-on-one with an adult. The books read aloud should help children learn new vocabulary and concepts. In addition, the adult reader also needs to provide support for children's learning by using dialogic reading techniques – asking questions, especially open-ended questions that promote higher-order thinking, supporting children's comprehension of the text through questions and by drawing attention to illustrations, and helping children understand the meaning of new vocabulary.

Print Knowledge. Instruction should ideally include actively directing children's attention to specific letters and words in the classroom environment as well as some direct instruction in letter names and sounds. These practices are intended to help children develop knowledge of letter names, skill in recognizing letter shapes and distinguishing letters from one another, and awareness of the functions and conventions of print (e.g., directionality, spacing, punctuation).

Phonological Awareness. One of the critical functions of instruction in the classroom is to help children develop the ability to hear and manipulate sounds in words. Particular forms of language stimulation appear to help children develop this phonological sensitivity. These include language games (rhymes, songs, poems) and books in which phonemic patterns (such as rhyme and alliteration) are present. Beyond simple exposure to these literary forms, the teacher's efforts to focus and encourage children's phonological sensitivity has been linked to increased phonological awareness.

Print-Rich Environment. In addition to rich language stimulation, children should be exposed to written language in the classroom used for a variety of purposes. Such an environment provides opportunities for interactions that foster all of the key areas of early literacy development.

Chapter Two: Conceptual Framework and Study Research Questions

Efforts to enhance child care providers' skills are an important part of most states' agendas for improving the quality of children's experience in child care. This experimental test of three focused curricula was intended to answer important questions about the possibility of training child care staff, many of whom have limited education beyond high school, to deliver such curricula with fidelity, the level of support needed to accomplish this, and the impact of the interventions on children's language development and emergent literacy. For the experiment, staff who teach four-year-old children in centers that were randomly assigned to one of the three language/literacy interventions received initial and refresher training in the curriculum they were assigned. To support them as they worked to use the curriculum in their classrooms, specially-trained mentors visited them every two weeks over an 18-month period to observe them and provide appropriate feedback and support.

The hypotheses that underlie the experiment are that: at this level of training and support, teacher knowledge and attitudes will change: changes in knowledge and attitudes will be reflected, in specific ways, in behavior and interactions with children and in the classroom environment that they create; and changes in behavior and interactions with children, combined with changes in the classroom environment, will result in positive impacts on children's language and emergent literacy skills. We hypothesized that, over time, most teachers would be able to implement the curricula with fidelity, though the time needed would probably differ for individual teachers and for the three curricula. Successful implementation of the curricula would bring about positive change in the type and amount of teacher language and literacy interactions with children, change the classroom environment and increase the amount and type of children's activities and interactions related to literacy. If staff changed their behavior and the learning environment as the curricula require, children's language and literacy skills would improve as a direct consequence.

The study's major research questions flowed from these hypotheses and examined three areas of impact: impacts on teacher behavior and the classroom environment (intermediate outcomes); and impacts on children's language development and early literacy skills. In addition, the study examined the differential effectiveness of the three curricula on all three sets of outcomes, and for teachers and children whose first language was not English. The major questions addressed by the study were:

- Does training in and ongoing support for preschool language/literacy curricula have positive impacts on the type and amount of staff language and literacy interactions with children?
- Does training in and ongoing support for preschool language/literacy curricula have positive impacts on those aspects of the classroom environment that foster early literacy?
- Does training in and ongoing support for preschool language/literacy curricula have positive impacts on children's language development and emergent literacy skills?
- Do the interventions have different effects on teacher and child outcomes?
- Do the interventions have differential effects on teachers whose primary language is not English?
- Do the interventions have differential effects on children whose home language is not English?

- Does the focus on intentional teaching of language and literacy change the pattern of activities in the classroom? and
- To what extent does the teacher's educational background influence the impact of the interventions?

To address these questions, a rigorous experimental test of three strong language/literacy interventions was designed. The next chapter describes the experimental design and its actual implementation.

Chapter Three: Study Design and Implementation

This chapter sets forth the design of the study, describes the process of recruitment and random assignment of centers and examines the extent to which random assignment was carried out successfully. The chapter ends with a description of the classrooms and teachers participating in the study.

Overview of the Design

The study designed to address the research questions set forth in the previous chapter called for random assignment of 162 child care centers serving subsidized children to one of three language/literacy interventions or to an “as is” control group. One four-year-old classroom in each center was selected to participate in the study. Classroom staff in the “treatment” groups were trained to implement the curriculum to which they were assigned. The initial training was supported by ongoing mentoring and follow-up and refresher training sessions. The experiment was conducted over a two-year period with the same centers, classrooms and staff.

In an effort to encourage staff stability, teachers in all the study classrooms received a bonus payment for each year of the study that they remained at the center. If classroom staff left during the year, replacement staff were trained in the appropriate curriculum. To ensure a “prepared environment” in which literacy curricula could be implemented, all study classrooms (treatment and control) received a basic package of literacy materials (books, a tape player, jumbo pencils, pre-writing paper, crayons, whiteboard). In addition, control centers received a package of materials for their infant-toddler classrooms or a set of outdoor play materials.

Observations were conducted in study classrooms in the Fall of 2003, the Spring of 2004, and the Spring of 2005. Outcomes for children were assessed in late Spring 2005. In addition, the ELC shared with the study background information on teachers collected in Fall 2003, and data from the ELC-sponsored assessments of children, using the LAP-D, from Fall 2003, Spring 2004 and Spring 2005.

Implementation of the interventions was studied through observation of training sessions, visits to classrooms, interviews and group meetings with mentors, and analysis of the implementation rating scales completed by mentors for each of the interventions.

Sample Design and Statistical Power

The experiment required a sample size of 162 centers (four-year-old classrooms) to be randomly assigned—36 to each of the three curricula and 54 to the control group (Exhibit 3-1). An unbalanced design was chosen because of budget considerations that constrained the number of curricula to be tested and the number of centers that could be included in the treatment groups.

The sample of classroom staff included 162 teachers. Many classrooms were expected to have aides but the number was unknown at the point of study design and so aides were not included in the design. The sample of 1944 children was based on the assumption that, of the children in each classroom at the end of the 2004-2005 program year, approximately 12 would have been in the classroom for two or more months and parents would have given permission for assessment. Given the instability of classroom enrollment, there was no anticipation of a need to sample children.

Exhibit 3-1

Expected Number of Centers, Classroom Staff, and Children, by Assignment

	Treatment 1	Treatment 2	Treatment 3	Control Group	Total
Centers	36	36	36	54	162
Teachers	36	36	36	54	162
Children	432 (12 per classroom)	432	432	648	1944

Exhibit 3-2 shows the minimum detectable effect (MDEs) sizes for Project Upgrade. The study was designed to have 80 percent power to detect MDEs of around 0.20 for impacts on child outcomes. Since there were, by design, fewer teachers than children in the study, it was expected that the impact on teacher behaviors would have to be larger, in the range of 0.38 to 0.61, in order to have 80 percent power to detect impacts.

The rows in the exhibit show three experimental comparisons: a) a comparison of one of the treatment strands with the control group; b) a comparison of two treatment groups with each other; and c) a comparison of the average outcome for the combined treatment groups with the average outcome for the control group.

Exhibit 3-2

Projected Minimum Detectable Effects[5] for the Evaluation of Project Upgrade

	Unit of Analysis	
Comparison	Teachers	Children
One treatment strand compared with the control group	(a) 0.52	(d) 0.22
One treatment group compared with another treatment group	(b) 0.61	(e) 0.26
Combined treatment strands compared with control	(c) 0.38	(f) 0.17

Eligibility for Participation

Child care centers in Miami-Dade County were eligible to participate in the study if they served some children whose care was subsidized, as well as other children from low-income families. The centers had

[5] Calculations of minimum detectable effects (MDEs) assumed two-sided hypothesis tests with alpha level p<0.05, 80 percent power, and sample sizes shown in Exhibit 3.1. For impacts on teacher behaviors, MDE calculations assumed that model terms for randomization blocks and baseline observational covariates would account for 15 percent of total variance. For impacts on child outcomes, MDE calculations assumed a class-level intra-class correlation equal to 0.10, and that model terms for blocks and class-level mean child assessment scores at baseline would account for 25 percent of class-level variance. To arrive at these estimates we used a set of measures likely to be used for the study.

to have at least one classroom with at least five four-year-olds enrolled at the time of recruitment. They could not be already testing or implementing a literacy curriculum. All children in the selected classrooms were eligible to participate.

Recruitment and Random Assignment

In Miami-Dade County, two central agencies, Family Central and Child Development Services held subcontracts with the Early Learning Coalition to administer CCDF child care subsidies in the county (and provide training and technical assistance to providers, using CCDF quality improvement funds). Although both agencies provided subsidies for children in family child care, the majority of subsidies went to children in licensed centers in the county. Between them, the two agencies made subsidy payments to approximately 900 centers.

The two central agencies and the ELC took the lead in recruiting the sample of centers for the experiment. The ELC began by sending out information about the study to all the centers that served subsidized children in the county. This was a very wide net to cast, because between one-quarter and one-third of the centers provide before- and after-school care only. However, because in the county centers are licensed to serve a specific number of children, regardless of their age, there were no available data that would allow the ELC or the central agencies to automatically sort for those centers that served only school-age children and eliminate them from the mailing. Given the very brief amount of time available for recruiting, the decision was made not to embark on a time-consuming hand sorting of centers, and to ask about the presence of four-year-olds as one of the elements of the fact sheet that interested centers were asked to fill out and return.

ELC staff and staff from the two central subsidy agencies then made follow-up telephone calls and screening calls to determine interest and investigate eligibility. Abt, MDRC and ELC staff held a series of informational meetings for center directors and staff to answer questions and explain the random assignment process.

Fact sheets on all centers that expressed interest in participating were then sent to Abt staff to determine eligibility. Centers were determined to be ineligible for the following reasons:

- The center served only school-age children;
- The center served one or two subsidized children but primarily served more affluent families;
- The center had too few four-year-olds[6]; or
- The center was currently testing or using a literacy curriculum.

While many centers that served school-age children only or were otherwise ineligible undoubtedly understood that they were ineligible and did not respond for that reason, there were other reasons why a center might not respond. Many centers in Miami are either small, for-profit businesses or faith-based entities. While both groups are well-represented in our sample, many faith-based centers use A BEKA

[6] Because of the timing of recruitment, the requirement that centers have at least one classroom with at least five four-year-old children eliminated some centers that almost certainly enrolled that many children by mid-to late September. Centers in Miami-Dade County tend to be small; more than half of the licensed centers serve fewer than 60 children. In some areas, such as Hialeah, many of these small centers are clustered in close geographic proximity, compete vigorously for children and are permanently under-enrolled.

Book[7], a Christian school curriculum designed for 4-14-year-olds, and are committed to it. Some small business owners may not have wanted to participate in a government-sponsored study. When agency staff followed up with telephone calls, a small number of center directors (less than 20) informed them that, although the center was eligible to participate, they were not interested in doing so.

Design and Implementation of Random Assignment

The design for random assignment called for a single classroom to be selected and centers to be grouped by agency affiliation and teacher's dominant language (i.e., the language she preferred to be trained in).

For centers with more than one four-year-old classroom serving subsidized children, one classroom was chosen for the experiment. If one classroom had more subsidized children than the other(s), that classroom was selected. If two or more classrooms had the same number of subsidized children, then the one with the most children was chosen. If classrooms were equally large and had the same number of subsidized children, then one classroom was chosen randomly.

Assigning priority to larger classrooms was done for two reasons. First, selection of larger classrooms would make it easier to detect significant impacts on children's language and literacy. Second, if the interventions benefited children, then more children would be helped. For the same reason, if the recruitment and eligibility process yielded more centers than were needed for the study (with some held in reserve to serve as replacements for centers whose directors refused to participate before hearing their assignment), random assignment would be done first with centers with larger classrooms.

The recruitment and eligibility determination processes yielded a total of 300 eligible centers. Study staff conducted a series of meetings with directors from these centers to ensure that they understood and agreed with the random assignment process and to explain what participation in the study would entail. We met with all the directors or owners from these centers, and in that process eliminated more of them. Sometimes a director was initially interested but really wanted one of the three curricula or didn't want to risk being part of the control group. A few centers sent staff to meet with us because they were without a director; in these cases we were concerned that a new director would not honor the agreement to participate.

At this point, the 200 centers that remained were randomly assigned. The centers were first sorted by agency affiliation and teacher's language preference, creating four groups. Within each group, centers were sorted by the size of the classroom (i.e., the number of four-year-old children) into groups of 12. Within each group of 12 centers, three were placed randomly into the control group, two were placed in each treatment group, and three held in reserve. These reserve centers would be used as a pool of replacements for centers whose directors declined to sign an agreement to participate before learning their assignment.

The 200 centers that had been assigned were invited to a meeting at the beginning of October (180 attended). Directors or owners were asked to bring the teacher from the four-year-old classroom with them to the meeting. Once again, the random assignment process was explained, as well as the study

[7] A Beka Book was developed at Pensacola Christian Academy by Beka Horton, wife of its president and founder. Its goal is to build the content of every textbook and activity on the Bible. For preschoolers, it offers Bible story flashcards, Bible memory picture cards and Bible stories for read-aloud activities.

requirements. Directors were asked to review and sign an agreement to participate (most had reviewed this document at earlier meetings) and teachers were asked to review and sign a similar, though simpler agreement. At this point, the ELC also distributed and collected staff background questionnaires. Some director/owners left the meeting at this point, before hearing their assignment, and a replacement was randomly selected from the reserve group. Ultimately 164 centers signed agreements to participate and received their assignments; there were no refusals after centers learned their assignment, something the process had been carefully designed to avoid.

Over the course of two years, seven centers left the study. Five left because the center was closed or sold to an owner who chose not to participate; only two left because the director decided not to continue with the curriculum to which they were assigned. While, in spite of the incentives offered, teachers did leave and were replaced, our concern was about the attrition of centers, since they were the unit of random assignment. Center attrition was low and distributed quite evenly across the four groups.

Success of Random Assignment

In studies with small samples such as this one, there must always be a concern about the possibility of a carefully-implemented random assignment process that nevertheless produces groups that differ significantly, purely by chance. Three classroom-level measures were used to assess the success of random assignment, that is, the equivalence of the four groups: a staff background questionnaire (collected for other purposes by the ELC), the baseline observation measures and the LAP-D assessments of children administered in Fall 2003 and Fall 2004. There were no significant differences between treatment and control groups. We therefore concluded that, in terms of measurable aspects of the classrooms, random assignment was successfully carried out. Exhibits 3-3, 3-4, 3-5, 3.5a and 3-6 provide a detailed comparison of the baseline characteristics of the four groups.

There is one additional question that can be only partially addressed. The study centers represent approximately one-quarter of the centers in Miami-Dade County that were providing child care to subsidized four-year-olds in 2003. To what extent are they representative of those centers? Because the licensing agency, the ELC and the county's central agency maintain minimal information on center characteristics, there is a limit to our ability to assess their representativeness. However, Florida International University, under a subcontract with the ELC, maintains LAP-D records on all centers in which children were assessed. A comparison of the Project Upgrade average LAP-D scores on the three subscales of interest with the average scores for all centers showed no significant differences on any subscale. Therefore, we conclude that, on the measures of particular interest to the ELC, that generated the need for the study, the Project Upgrade centers were representative of the wider community of centers that served low-income children.

Exhibit 3-3

Difference between Intervention Groups and Control on LAP-D, OMLIT and Arnett Scores at Baseline (Fall 2003)

Construct	Control Mean (SD)		Treatment Mean (SD)		Mean Difference T-C	Effect Size	P-value
LAP-D							
Cognitive Total	30.43	(4.16)	30.86	(4.19)	0.43	0.10	0.53
Language Total	28.84	(4.34)	29.51	(4.26)	0.80	0.16	0.30
Fine Motor Total	38.88	(4.97)	39.68	(4.60)	0.68	0.16	0.34
OMLIT							
Support for Oral Language	53.26	(9.71)	53.82	(9.87)	0.57	0.06	0.72
Support for Print Knowledge	53.27	(2.89)	53.38	(2.83)	0.11	0.01	0.81
Support for Print Motivation	54.41	(9.03)	53.14	(7.30)	-1.27	-0.14	0.33
Literacy Resources	50.14	(5.62)	50.69	(4.64)	0.55	0.05	0.50
Arnett							
Positive Affect	51.20	(8.81)	48.76	(9.31)	-2.44	-0.24	0.10
Not Punitive	47.83	(6.59)	46.49	(7.26)	-1.34	-0.13	0.25
Engaged	49.78	(11.75)	46.64	(13.8)	-3.14	-0.31	0.15
Sample size (centers/classrooms)	***55***		***110***				

Exhibit 3-4

Baseline (Fall, 2003) OMLIT Score, by Treatment Group

Measure	Control Mean (SD)		Treatment 1 RSL Mean (SD)		Treatment 2 BELL Mean (SD)		Treatment 3 (BTL) Mean (SD)	
Support for Oral Language	53.26	(9.71)	51.28	(9.01)	53.66	(10.01)	56.68	(10.07)
Support for Print Knowledge	53.27	(2.89)	53.44	(2.76)	52.89	(2.13)	53.81	(3.45)
Support for Print Motivation	54.41	(9.03)	51.65	(7.24)	53.67	(8.05)	54.20	(6.48)
Literacy Resources	50.14	(5.62)	50.70	(4.97)	49.81	(4.62)	51.54	(4.26)
Sample Size (centers/classrooms)	***n=54***		***n=38***		***n=36***		***n=36***	

There were no significant differences among groups (Oral Language p=0.11; Print Knowledge p=0.57; Print Motivation p=0.38; Literacy Resources p=0.44)

Exhibit 3-5

Baseline (Fall, 2003) Scores on Three LAP-D Subtests, by Treatment Group

Subtest	Control Mean (SD)		Treatment 1 RSL Mean (SD)		Treatment 2 BELL Mean (SD)		Treatment 3 (BTL) Mean (SD)	
Cognitive	30.43	(4.16)	31.56	(4.52)	31.33	(3.95)	29.71	(3.91)
Fine Motor	38.88	(4.97)	39.98	(4.58)	40.33	(4.10)	38.76	(5.03)
Language	28.84	(4.34)	29.76	(4.38)	30.43	(4.17)	28.41	(4.09)
Sample Size (centers/classrooms)	n=53		n=36		n=33		n=35	
Sample Size (children)	580		350		319		350	

There were no significant differences among groups (Cognitive, p=0.19; Fine Motor p=0.35; Language p=0.16)

Exhibit 3-5a

Scores on Three LAP-D Subtests, by Treatment Group (Fall, 2004)

Subtest	Control Mean (SD)		Treatment 1 RSL Mean (SD)		Treatment 2 BELL Mean (SD)		Treatment 3 (BTL) Mean (SD)	
Cognitive	32.92	(5.15)	34.50	(5.44)	34.50	(5.02)	32.05	(5.15)
Fine Motor	42.14	(6.24)	44.00	(5.92)	44.17	(5.42)	41.78	(6.35)
Language	31.41	(5.71)	33.42	(6.69)	33.66	(5.98)	30.55	(4.89)
Sample Size (centers/classrooms)	n=53		n=36		n=33		n=35	
Sample Size (children)	509		320		340		354	

There were no significant differences among groups or between the control and combined treatment group (Cognitive, p=0.47; Fine Motor p=0.26; Language p=0.26)

For 20 centers for which 2004 LAP-D scores were not available, 2003 LAP-D scores were used

Exhibit 3-6

Baseline (Fall, 2003) Characteristics of Teachers and Classrooms, by Treatment Group

	Control	**RSL**	**BELL**	**BTL**
Measure	**%**	**%**	**%**	**%**
Spanish-language preference	49.1	47.2	48.5	45.7
Sample Size (teachers)	n=53	n=36	n=33	n=35
Chi-square test of independence, df=3, p= 0.99				
Education				
High school only	21.6	30.6	30.3	37.1
Some college	13.7	13.9	15.2	11.4
AA or BA	64.7	55.6	54.6	51.4
Sample Size (teachers)	n=51	n=36	n=33	n=35
Chi-square test of independence, df=6, p= 0.84 2 Teachers had missing data.				
Percent of classrooms with all English-speaking children	36	32	33	44
Percent of classrooms with all Spanish-speaking children	46	47	58	42
Percent of classrooms with mixture of language	18	21	9	14
Sample Size (centers/classrooms)	n=53	n=36	n=33	n=35
Chi-square test of independence, df=6, p= 0.74				

Classrooms and Teachers in Fall 2003

Across the 164 classrooms in the study, 54% of the children were predominantly Spanish-speaking, 41% spoke English as their primary language, less than 1% spoke Haitian Creole and the remainder spoke languages other than those. In spite of this linguistic diversity, most classrooms were linguistically homogeneous. In 36% of the classrooms, all the children spoke English as their primary language; in 48% all the children spoke Spanish as their primary language. In those classrooms, Spanish was also the primary language of the teacher and was the dominant language in the classroom. In 16% of classrooms there was a mix of languages. (Exhibit 5-6 shows the distribution by treatment group.) In classrooms with one or more Spanish-speaking children, at least one staff member spoke Spanish, and was able to communicate with monolingual Spanish-speaking children.

Although Florida licensing regulations allow a staff/child ratio of 1:20 for four-year-olds, and have no group size requirements, the observational data suggest better ratios and relatively small group sizes. The average observed ratio was one staff member to 10 children, with an average group size of 15 children.

Three observational measures used in Fall 2003 captured the quality of the literacy environment before literacy materials were distributed and training for the curricula began. In general, they reflect an environment that offered little support for emergent literacy. On a measure of the richness of the print environment, that is the type and quantity of materials that support the development of early literacy skills, the average score across all classrooms was 1.1 out of a possible 3.0. While reading aloud was observed in 59% of the classrooms, most of those had only one read-aloud session and the average time spent in reading aloud was 13 minutes. Most activities involved the group as a whole or large groups of children, and only a small proportion of activities involved anything that might encourage emergent reading or writing.

Classroom Teachers

More than half of the 164 teachers in the study spoke Spanish as their primary language, though only 28% reported speaking only Spanish in the classroom. Just over one-quarter spoke English at home and 11% spoke both languages. A majority spoke English only (42%) or a mix of English and Spanish (26%) in the classroom. More than one-quarter (28%) had no education beyond high school. A small percentage (14%) reported some college education but no degree. More than half (58%) reported having an Associate or BA degree.[8] Of the post-secondary degrees reported, more than 75% were from institutions outside the United States. The distribution of staff characteristics was similar across the four groups (Exhibit 3-6).

[8] This is a higher proportion than we expected to find. Most of the more highly-educated teachers were Spanish-speaking and had obtained their degrees outside the US.

Chapter Four: Study Measures and Data Collection

The study directly employed three types of measures: a self-administered staff questionnaire to provide information on the educational background and experience of teachers in the Upgrade classrooms; a battery of observation measures, the *Observation Measures of Language and Literacy Instruction* (OMLIT, Goodson et al., 2004), that focuses on the language and literacy environment of and interactions within the preschool classroom, but also captures a wide range of other activities,[9] paired with the *Arnett Caregiver Rating Scale* (Arnett, 1989), that rates the caregiver's emotional tone, discipline style, supervision of and interest in children and encouragement of independence; and the *Test of Preschool Emergent Literacy* (TOPEL: Lonigan, Wagner, Torgesen, & Rashotte, 2002), a standardized assessment of the aspects of language development and pre-literacy skills that research has shown to predict later reading success. We discuss the rationale for the selection of the observational and child assessment measures below.

In addition, center- and classroom-level scores on the LAP-D, a broad diagnostic screening measure applied to four-year-olds receiving subsidies for child care, were provided by the ELC for use as covariates in the analysis.

Classroom Environment Measures

The model tracing the pathway of effects of the language and literacy interventions in the Miami experiment shows that impacts on children depend on prior changes in the children's experiences in the child care centers. That is, the interventions must, first, change the center environments as a necessary condition for improving outcomes for the children. Random assignment allows us to attribute treatment-control differences in children's outcomes to the interventions, even without knowing anything about the center environments. However, the impacts on children will be better understood if we know about the extent to which the centers themselves changed. If we failed to find any impacts on children, it would be important to know if the interventions failed to effect significant changes in the centers. Further, in the event that there are child impacts, we wanted to know how these were achieved—i.e., the types of changes that occurred in the classroom. Therefore, the design of the study called for measuring treatment-control differences in the center environments, in addition to measuring differences in child outcomes.

Since the purpose of assessing center environments is to identify differences in treatment and control centers that could be logically linked to effects on children, we wanted to use measures that would be sensitive to changes in those aspects of the center care environments that are hypothesized to be modified

[9] The individual measures in the OMLIT are described in Attachment B.

as a result of the interventions.[10] This requires an initial analysis of the expected differences between classrooms using the intervention curricula and the "business-as-usual" classrooms.

Examination of the goals and activities of the three interventions led us to identify the following aspects of the treatment classrooms as central to the changes that should result from implementing any of the three curricula:

- Focused emergent literacy activities
    - Phonological awareness activities (singing, breaking apart words into syllables, language games about alliteration and rhyming)
    - Print knowledge activities (alphabet knowledge, letter-sound correspondence, grammatical rules)
    - Print awareness activities (focus on uses of print, emphasis on reading aloud)
    - Oral language activities (in-depth discussions, conversations, scaffolded language, open-ended questions, exposure to new vocabulary)
    - Writing activities (dictation, invented spelling, journals)
- Reading aloud using dialogic reading methods
- Small group activities involving caregivers and children (individual children, pairs, small groups)
- Integration of print throughout the day and throughout the classroom
- Authentic print, literacy activities
- Print-rich classroom environments
- • Caregiver engagement with the children in activities outside management/routines.

The OMLIT (Observation Measures of Language and Literacy Instruction) was a new battery developed for the national study of the Even Start Family Literacy Program being conducted by the U.S. Department of Education. The CLIO[11] study was also an experimental test of early childhood language and literacy curricula, and, as with the Miami study, CLIO needed measures of classroom process that would be sensitive to the interventions. The CLIO study also reviewed available measures, including the Early Language and Literacy Observation Tool (ELLCO)[12] and the ECERS-R, and determined that new

[10] To assess the quality of early childhood programs, the most commonly-used measure in the field is the Early Childhood Environment Rating Scale, now revised (ECERS-R). Based on at least two hours of observation, it provides an overall quality score and subscores in 6 domains. Although the ECERS has been used in many studies of early childhood care, it had significant limitations for the Miami study. Although the ECERS has been used in many studies of early childhood care, it had significant limitations for the Miami study. First and foremost, it has very few items that measure the emergent literacy instructional behaviors that were the central focus of the interventions. Although the revised version of the ECERS was an attempt to strengthen the measure in the area of early literacy, we did not believe the measure would be sufficiently focused or detailed to be sensitive to changes in these areas. We considered several other measures that focused more specifically on the literacy environment, but none had training materials or psychometric information available.

[11] The study is named CLIO, for Classroom Literacy Intervention and Outcomes study.

[12] Like the ECERS-R, the version of the ELLCO available at the time of the study focused primarily on environmental supports for literacy (classroom materials, activities) with little attention to teacher-child interactions which we believed to be critical elements of support for language and literacy development.

measures would have to be developed if measuring effects on classroom process was a priority. The Department of Education supported the development of the OMLIT battery, with the charge that the measure would be closely linked to the most up-to-date research on instructional practices shown to predict children's reading and other academic outcomes in school. Given the more than adequate reliability of the OMLIT battery, its clear link to all of the critical classroom outcomes in the study, and its track record in large-scale applied research, we selected the OMLIT for the Miami study. Although we considered administering the ECERS-R together with the OMLIT, for purposes of comparison with other early childhood studies, we judged that the two measures would have to be administered in separate visits to classrooms (i.e., observers could not reliably code both the OMLIT and the ECERS-R simultaneously). The cost of the additional training and doubling the visits to classrooms was determined to be prohibitive, especially in light of the limited usefulness of the ECERS-R for measuring treatment-control differences (versus allowing us to characterize the quality of the child care centers in the Miami sample compared with other samples).

The complete battery of observation measures used includes five instruments from the Observation Measures for Language and Literacy (OMLIT; Goodson, Layzer, Smith, Rimdzius, 2004) battery and the Arnett Caregiver Rating Scale (Arnett, 1989)[13].

The Snapshot of Classroom Activities (OMLIT-Snapshot)

The OMLIT-Snapshot is a description of classroom activities and groupings, integration of literacy in other activities, and language in the classroom. It has two sections. The Environment section describes the number of children and adults present, as well as the type of adult (staff, parents, etc.). The Activities section describes activities that are taking place. For each activity, the observer records the number of children and adults in that activity, whether any adult or child is talking, and whether they are speaking English or another language, and whether literacy materials are used (text, writing, letters, singing).

The Read Aloud Profile (OMLIT-RAP)

The OMLIT-RAP is a description of staff behavior when reading aloud to children (in CLIO, the RAP was completed when an adult was reading to at least *two* children). The RAP records adult behavior during the read-aloud session in four categories: (a) pre-reading (set-up) behavior, (b) behavior while reading the book, (c) post-reading behavior, and (d) the language the adult uses when talking to children during the read aloud. The RAP records characteristics of the adult, the children, and the book itself in three categories: (a) role of the adult involved in the read-aloud (e.g., teacher, aide, etc.), (b) characteristics of the book being read, and (c) number of children involved in the read-aloud. The RAP also includes five quality indicators which summarize particular aspects of the read-aloud: (a) the degree to which the adult introduces and contextualizes new vocabulary to support children's learning, (b) the depth of the discussion related to the story that the adult facilitates with the children before, during, and after the read-aloud, (c) the extent to which the adult uses open-ended questions that invite children to engage in prediction, imagination, and/or rich description, (d) the depth of children's engagement with the read-aloud activity, and (e) the quality of any post-reading book-related activities that the adult organizes (beyond oral discussion).

[13] The Arnett measures the teacher's affective behavior and disciplinary style.

The Classroom Literacy Opportunities Checklist (OMLIT-CLOC)

The OMLIT-CLOC is an inventory of classroom literacy resources. It provides an overall rating of the extent to which a classroom is a literacy-rich environment and delineates eight aspects of the literacy environment: (a) physical layout of the classroom, (b) the text or print environment, (c) books and reading or listening areas, (d) writing resources, (e) literacy-related materials and toys, (f) cultural diversity in literacy materials, (g) literacy integrated in classroom areas or learning centers, and (h) the richness and integration of a curriculum theme.

The Classroom Literacy Instruction Profile (OMLIT-CLIP)

The OMLIT-CLIP involves a two-stage coding protocol in which the observer first determines if any classroom staff member is involved in a literacy activity and, if so, the observer codes seven characteristics of the literacy activity: the type of activity, the literacy knowledge being afforded to the children, the adult's level and type of participation in the activity, any text support, languages spoken by staff and children, and the number of children involved. If the literacy activity involves adult-child discussion, the quality of this discussion is rated on three characteristics—the cognitive challenge in the discussion, the extensiveness of the discussion, the level of abstraction of the discussion.

The Quality of Language and Literacy Instruction (OMLIT-QUILL)

The OMLIT-QUILL is an overall evaluation of the quantity and quality of the instructional practices that build children's print awareness and oral language skills, expose children to a rich and varied vocabulary, and build children's phonological awareness. These practices are predictors of better reading outcomes for children once they are in school; this is particularly true of those at risk for reading difficulties (Dickinson and Tabors, 2001; Lonigan, Burgess, and Anthony, 2000; NICHD, 2000; Snow, Burns, and Griffin, 1998; Whitehurst and Lonigan, 1998). In addition, the *QUILL* evaluates instructional practices with English language learners.

The development of the OMLIT took nearly two years, and included reliability studies and multiple rounds of piloting in child care centers. In the CLIO study, the OMLIT was administered in the field over three years, with trained observers using the measure in more than 200 classrooms in each year, and calculation of inter-observer agreement for each group of observers. A description of the methods used to assess reliability and the results of the tests can be found in Appendix A.

Measures of Child Outcomes

The goal of the Miami-Dade experiment was to improve the language development of the children in the centers, since the first round of county-wide testing had shown that the children receiving child care subsidies scored, on average, at the 30th percentile on the language subscale of the LAP-D. At the same time, children in the Miami-Dade public schools were performing poorly in the high-stakes testing conducted statewide in 3rd grade. Therefore, the ELC was interested in testing curricula designed specifically to improve language and early literacy skills in preschool that might lead to improved performance when the children reached 3rd grade.

The ELC, planned to continue its own testing of subsidized and other low-income children using the LAP-D.[14] The LAP-D, which is administered by staff from the county agency that provides resource and referral services and administers subsidies, requires more than an hour of testing per child. In light of this ongoing county-wide testing program, the ELC was cautious about conducting additional testing of children for the purposes of the experiment. Therefore, the following guidelines had to be met in selecting child outcome measures:

- The assessment should impose as little additional burden as possible on the children (and classrooms), with the goal of less than 30 minutes of assessment per child; and
- The outcome battery should be sensitive to the content of the curricula, to increase the chances of detecting impacts;
- The outcome battery should use standardized, norm-referenced measures; and
- The outcome battery should assess skills identified in the research to have longer-term significance for children's academic success.

The study team reviewed the available child assessment measures, as well as consulting with national experts in language development, and also reviewed the measures being used in other national early childhood studies, including the national study of the Even Start Family Literacy Program, the National Head Start Impact study, the Head Start National Reporting System, the PCER (Preschool Curriculum Evaluation Research) studies, and the national evaluation of Early Reading First. Across these studies, one measurement battery was being consistently used to assess children's emergent literacy skills, the TOPEL (Test of Preschool Emergent Literacy),[15] which tests three major domains: Phonological Awareness, Print Knowledge, and Definitional Vocabulary. We therefore recommended to the ELC and to ACF that the additional child assessments for the experiment should use the TOPEL, since it met all of the study criteria, as well as the ELC guidelines, and the recommendation was accepted by both groups.

Data Collection

The experiment was conducted over a two-year period. Centers were recruited and randomly assigned between August and October 2003. Observers were recruited and trained in September 2003, and retrained in Spring 2004 and Spring 2005. Baseline observations were conducted before training in the interventions took place, from October to late November 2003. Classrooms were observed in late Spring 2004, after approximately six months of implementation of the curricula and again in late Spring 2005,

[14] The LAP-D was intended to be used as a diagnostic screening test to identify children who were at or lagging behind normal development in 4 major domains: cognitive, language, fine motor, and gross motor.

[15] At the time that this battery was adopted in whole or part in all of these national studies, it had a different name (the Pre-CTOPPP for the Preschool Comprehensive Test of Phonological and Print Processing) and was in the process of being normed by Pro-Ed, an international publisher of standardized tests and assessments based in Austin, Texas. The norming data was expected to be available by late 2005, according to the test authors, although the raw scores would be appropriate for analytic purposes. As promised, the norming data were released in spring 2006, in time for the experiment to use standardized scores for analysis and to characterize the developmental status of the sample children in comparison with a national sample of children of similar age. (It should be noted that all of the other studies will conduct their analyses using the raw TOPEL scores, since the norming data were not available in time for their analyses. Future analyses of data from these studies may be able to convert the raw scores to standardized scores.)

after 18 months of implementation. Child assessors were recruited and trained in Spring 2005. Outcomes for four-year-olds were measured in late Spring 2005, after between two and ten months of potential exposure to the interventions.[16] Child assessments were conducted for all children in the study classrooms whose parents gave permission for them to be assessed and who had been in the classroom for at least two months.

The ELC distributed and collected staff questionnaires at recruitment meetings and at the meeting when centers were told their assignments. The ELC's Upgrade Project Manager collected similar information on replacement teachers in most cases.

Hiring and Training Observers

Staff to conduct the classroom observations were recruited and hired in Miami-Dade County from several academic institutions, including Florida International University (FIU) and Miami-Dade Community College. Because the observation battery was extensive and required an understanding of early childhood education, we were particularly interested in individuals who had experience working in an early child care setting but also had some academic qualification (or were working on a degree) in early childhood education. In addition, because the observation measures focused on teacher language, and many of the teachers observed used Spanish in the classroom, we required that a majority of observers be bilingual. Those with prior experience actually conducting observations in child care settings were given priority. The field manager employed by FIU was responsible for recruiting potential candidates; Abt staff interviewed candidates and selected observers.

Before each data collection period, Abt staff held six-day training sessions for observers in Miami. After the initial training session, observers who had stayed with the study were required to attend only a two-day refresher training before the subsequent data collections. All new observers were required to attend a six-day training session.

The training sessions included information about the study and a background session on children's language and literacy development and classroom practices that support that development. Each instrument in the battery was introduced and the items described. Trainees coded vignettes for each instrument and their accuracy was assessed against a criterion. Over a period of two days, observers were sent to child care centers not participating in the study to observe and practice coding each of the measures. An Abt staff trainer accompanied each group of observers and coded the measures simultaneously. Discrepancies and coding difficulties were discussed in group sessions at the end of each of the two live observation periods. Finally, each observer was paired with an Abt trainer to conduct a live observation. Trainees who failed to achieve satisfactory levels of agreement with the trainer were not hired. Inter-rater agreement was 80% or higher, depending on the measure.

Hiring and Training Child Assessors

Because children were assessed for the study at one time-point only, in Spring 2005, child assessors were recruited and hired in the two months preceding the data collection. As with the observers, child assessors were recruited from FIU and Miami-Dade Community College, both of which enroll a large number of

[16] The study did not measure the exposure of individual children to the interventions; we simply set a lower bound on exposure by excluding from assessment children who had entered the classroom less than two months prior to the assessment.

former early childhood educators who are seeking a graduate degree. Again, criteria for hiring included a background in ECE and bilinguality, with experience in child assessment given priority. The field manager was responsible for recruiting and screening applicants; Abt staff interviewed and selected assessors.

Child assessors were trained on the three TOPEL subtests over a three-day period, including actual administration of the TOPEL on non-study children. During the training, assessors were trained to a standard of 95% agreement on coding of standardized test protocols and use of appropriate probes. Each trainee practiced test administration using practice scripts (designed to test administration rules) while trainers observed and provided feedback.

Trainees then practised administering the test with children in volunteer sites while trainers observed and coded children's responses while monitoring administration to ensure that standardized procedures were followed. Each trainee's record booklets were then compared with trainers' simultaneously coded booklets to check for variance immediately following each administration, and feedback was provided as necessary. Trainees continued practice administration until no variances occurred. Prior to working with study children, each trainee was "tested-out" by administering the complete battery to a trainer, who followed a script designed to test administration rules.

Finally, initial data collection was conducted under the immediate supervision of trainers; that is, each assessor was observed (by a trainer) while conducting assessments with actual study children. Deviation from standardized procedures or variance in recorded scoring were both grounds for termination. Thus, all assessors who were invited to continue data collection had demonstrated mastery of standard administration.

Data Collection and Quality Assurance

At each of the three observation time-points, observers spent a morning in the classroom, arriving as the children began to arrive and leaving at lunch-time or just after. Observers worked closely with center staff to ensure that the time scheduled for the observation did not conflict with other center activities such as field trips, photo sessions (before graduation in the spring) or health screening.

The TOPEL was administered once, in Spring 2005, to about 1600 children in the study classrooms. These children represent the second cohort of children who received the enhanced language and literacy curricula. All child assessments were conducted individually, in the child's classroom. The assessments took place over a seven-week period and children in the treatment and control groups were assessed at approximately the same time. The child assessors were bilingual in Spanish and English and provided instructions in Spanish for children as needed. All children were tested in English.

To manage the data collection on-site and to provide continuous quality assurance, an on-site data collection manager was hired. She remained with the study throughout the two-year period and was responsible for meeting weekly with observers and assessors during the data collection periods, reviewing observation protocols and test booklets and conducting regular quality assurance visits with observers. In addition, Abt trainers held weekly conference calls with observers and child assessors and traveled to Miami halfway through each of the observation data collection waves to meet with observers and to conduct paired observations with them. Because the child assessments occurred only once, the quality assurance procedures used were slightly different. Abt trainers stayed on-site during the entire data collection period, meeting with assessors weekly or more frequently and observing child assessment sessions throughout the period.

Implementation of the interventions was assessed in a variety of ways: trainers for each developer used measures tailored to the individual curriculum; mentors for each curriculum, were asked to rate the level of curriculum implementation in the classrooms for which they were they were responsible on a scale developed for the study and applied across curricula; and senior study staff met monthly with developers, trainers and mentors to discuss implementation issues.

Chapter Five: The Interventions and Their Implementation

This chapter describes the elements of the three language and literacy interventions and their implementation over a period of 18 months in child care centers in Miami-Dade County. In this experiment, as in many other educational experiments, an explanation of the elements of the interventions and how they were actually implemented is important because the impacts described later in the report are outcomes of the specific models described here. Prior to this experiment, none of the three curricula had employed the combination of sequenced training and ongoing mentoring that is described in this chapter.

The Interventions

The three interventions comprised three linked components: a curriculum based on the most recent research on predictors of later reading success; three group training sessions for classroom staff at intervals across the 18-month period, with concepts and materials introduced in a planned sequence; and bi-monthly on-site mentoring by trained coaches.

The structure of the interventions was identical: for each curriculum, the developer provided an on-site trainer/coach (whose salary was paid for as part of the ELC purchase of the curriculum) to supervise two mentors (employees of the two central agencies, whose salaries were paid under the agencies' contracts with the ELC) each of whom was responsible for mentoring and support in 18 classrooms.

Element 1: The Three Curricula

The three curricula tested were selected by the Early Learning Coalition after a systematic and comprehensive review of language/literacy curricula had been conducted by Abt Associates' staff[17]. To be considered for the study, a curriculum had to meet the following criteria:

- Provides support for children's language **and** early literacy;
- Provides support for all four of the elements of language and early literacy that research has shown to be predictive of later reading success: oral language; phonological processing; print knowledge; and print motivation;
- Is appropriate for and has been used with children whose first language is not English and with low-income populations;
- Is supportive of children's home culture and language;
- Is appropriate for both three- and four-year-olds (since the ELC was interested in introducing a curriculum in three-year-old, as well as four-year-old classrooms);
- Has some preliminary evidence of effectiveness; and
- Has a training plan that would allow the curriculum to be implemented by child care center staff.

[17] The review was funded by the US Department of Education as part of a different contract.

Six of the curricula reviewed appeared to meet the criteria and were recommended to the ELC. Staff at the ELC reviewed the curriculum descriptions and interviewed three developers. One curriculum was rejected after the interviews because the developer believed it was not suitable for staff who were either monolingual in Spanish or would be more effectively trained in Spanish. Finally, the ELC chose to test two nationally-known curricula, *Ready, Set, Leap!* and *Breakthrough to Literacy.* The ELC also selected *Building Early Language and Literacy (BELL)*, a curriculum not on the Abt list which was developed by a local academic and at that time implemented in public preschools in the county.

The three curricula selected differed in instructional approach, materials provided, intensity and cost, but all three focused on the development of early literacy skills and knowledge. All three included take-home components (books and materials to be used by families with children at home), and tools that teachers could use to assess children's progress in the curriculum.

Ready, Set, Leap! (RSL; LeapFrog SchoolHouse Inc.) is a comprehensive curriculum with activities throughout the day that include math and science. A multi-sensory curriculum, it builds oral language, phonological awareness, print awareness and a language-rich environment through whole-group, small group and individualized instruction, including the use of three different interactive technology tools. Teachers organize their activities with children around multiple thematically-grouped trade books over a one- or two-week period (before moving to another set of books). The interactive technology uses separate, unrelated materials. The Leap Pad™ is an electronic story book system that uses controlled-language stories to focus on discrete concepts (e.g, color, number, letter shapes and names, word meaning, directionality of print). The Leap Desk™ is a simplified keyboard with key cards and is used primarily to help children learn to identify upper and lower case letters by name, to begin to learn letter-sound correspondence and, ultimately, to practice spelling (i.e., encoding). The Leap Mat™ is used for the same purposes. All three are meant to be used each day by each child.

Building Early Language and Literacy (BELL; unpublished) is an add-on pre-kindergarten literacy component designed to promote children's general language proficiency, phonological awareness, shared reading skills, and print awareness. It entails two daily 15-minute whole-group lessons. One daily? lesson typically involves oral reading (the teacher reads aloud from a Big Book[18] or a poster and the children repeat the rhyme or story) and a discussion of the story. Another daily lesson focuses on phonemic awareness and knowledge of letters and typically involves attending to the sounds in words (forming compound words, detecting rhymes, alliteration, etc.) and activities targeting print knowledge (directionality, identifying letters and known words in text). Materials provided for small group activities in centers after the whole group lesson include letter cards and books and listening center materials (books on tape). Teachers are encouraged to create their own materials.

Breakthrough to Literacy (BTL; Wright Group/McGraw-Hill) is a comprehensive curriculum with activities throughout the day that include math and science. The language and literacy curriculum is built around a series of weekly books with a focus on reading aloud and active discussion about the book. At the heart of the curriculum are five "daily essential practices": listening to and discussing stories; reading; writing; individualized software instruction; and talking, reading and writing at home. The teacher's whole group instruction is thematically focused on a Book of the Week (BOW). Each day, the teacher reads the BOW aloud, develops some kind of graphic organizer using the target vocabulary (linked to the BOW), and sets up small group activities that use the target vocabulary. Small group activities include writing, matching vocabulary/picture cards, and counting/early math. Each child is expected to spend 8-

[18] A Big Book, for those unfamiliar with them, is a large version of a children's book that allows the teacher to point to words or details of an illustration while reading to a group of children.

12 minutes a day on the computer. Computer software provides individualized self-paced literacy activities for children, also organized around the weekly book, that focus on phonological and print knowledge, with additional books for use once the child has completed the BOW activities.

As noted above, all three curricula included materials to be sent home weekly with each child: RSL provided interactive books to be read at home; BELL provided small book versions of the Big Books read and discussed that week; and BTL provided copies of the week's book to be read aloud at home. All three provided some materials in Spanish for children with the aim of motivating reading, regardless of the language. All three met the Florida Preschool Language and Literacy Learning Standards; it should be noted that two of the three, RSL and BTL, also met the state standards for a comprehensive curriculum, since they included math and science concepts

Exhibit 5-1 provides more specific examples of how the three curricula addressed each of the four critical aspects of language and emergent literacy.

Element 2: Training

The interventions involved two kinds of training; first, mentors who would visit classrooms and support curriculum implementation had to be trained in the curriculum and in the developer's approach to mentoring; and then the classroom staff who would implement the curriculum had to be trained. In many ways, the approach to training and supporting mentors mirrored the approach to training classroom staff. In neither case was it assumed that a one-time training would be adequate by itself to ensure adherence to the desired approach. Training for mentors was provided through a two-day training session, with one day devoted to the curriculum itself and the second to the mentoring process. The schedule was similar across the three interventions. In addition, the on-site coach/trainer provided ongoing consultation as well as weekly in-person meetings of the curriculum team.

Although the three curricula differed from each other in a variety of ways, classroom staff in all three groups received comparable amounts of training. Each curriculum developer provided three in-service training sessions, each lasting two days, off-site, for all teachers and aides who were involved in implementing the curricula, as well as interested directors. The first and second sessions were intended to cover all elements of the curriculum; however, rather than dealing with all aspects of the curriculum at the initial training (Fall, Winter 2003), the plan was to introduce some more difficult concepts and activities such as child assessment at the second training session (January, February 2004). The third training session (August 2004) was intended as a refresher training that would build on the experience of several months of implementation (1 day) but also provide training for replacement staff (2 days).

Element 3: Mentoring

As we noted earlier, none of the curricula chosen had employed ongoing mentoring in other sites that used their curriculum. Nor did the specification of the level and intensity of mentoring necessarily reflect only what the developers thought necessary. For this study, the assignment of two mentors to each curriculum, each of whom would be responsible for 18 classrooms, reflected two different considerations: the developers' recognition that, for child care center staff, with widely differing levels of education and training, some ongoing support would be essential; and the ELC's budget constraints. The decision to assign 18 classrooms to each mentor assumed bi-monthly visits to each classroom, but there was no prior assumption made about the length of the visits.

Implementation of the Interventions

The research questions about implementation set forth in Chapter 3 required several different data collection strategies. These included:

- Meetings with curriculum developers and ELC staff;
- Observation of training sessions;
- Document review (resumes, training agendas, manuals, measures of curriculum implementation and fidelity[19] applied by mentors and their coaches);
- Observation of mentoring visits;
- Monthly meetings with all Upgrade staff and partners;
- Informal interviews and discussions with individual mentor staff; and
- Implementation rating across curricula[20].

The discussion that follows reflects input from all of these sources.

Preparing for Implementation

Our discussions with the ELC assumed a two-year implementation of the selected curricula. To accomplish this, the decision to embark on the experiment needed to be made in the Spring of 2003. In reality, the decision was made at the end of July 2003. The preceding chapter described the recruiting of centers for the study that was accomplished in the two-month period following that decision. However, before classroom staff could be trained to deliver the curriculum, there were many steps to be taken, including the hiring of on-site coach/trainers by two of the three developers[21]; hiring, assignment, and training of mentors; the collection of baseline observation data in the study centers; and the collection of LAP-D child assessments by the ELC. Once these and other essential activities were completed, classroom staff could be trained in November/December 2003. This reduced the implementation period to 18 months (mid-December 2003 to mid-June 2004).

In addition to the steps noted above, the curriculum developers identified two concerns after visiting centers that had been recruited to the study. The first concern was that many centers were poorly equipped to put in place a strong literacy curriculum. They lacked many of the basic materials that we would expect to see in an early childhood setting: e.g., art and writing materials; whiteboards; tape players and audiocassettes; and a variety of children's books, among other things. The decision was made by the ELC, in consultation with members of the study team, to provide a package of basic literacy materials to

[19] Each developer provided their coaches and mentors with a fidelity measure tailored to the specific elements of the curriculum and the sequence in which they were to be introduced and implemented. These ratings provided the basis for the mentoring session and for discussions between coaches and mentors about how to address gaps in implementation and to eliminate barriers that might prevent the teacher from implementing some aspects of the curriculum .

[20] Abt staff attempted to create a five-point scale measure of implementation that might be used across the three curricula, but it proved impossible to develop a common set of definitions, in large part because the curricula differed in the timing and sequence of some key concepts, so this measure was not used. Instead, we relied on the fidelity measures developed for each of the three curricula.

[21] BTL already had a trainer on staff who moved to Miami for the study.

all the four-year-old classrooms in the study, including those in the control group. These had to be selected, ordered and delivered before teachers were trained.

The developers' second concern arose from the large proportion of teachers (47%) who had indicated that they would prefer to be trained in Spanish. (Not all of these were monolingual; almost half of teachers whose first language was Spanish reported that they spoke a mixture of English and Spanish in the classroom.) Although all three developers had expressed their willingness and readiness to train Spanish-speaking teachers, it was clear that without a training plan specifically tailored to meet the needs of this population, the interventions were likely to be unsuccessful. The development of these plans was helped by the fact that all three of the on-site coach/trainers were fluent in both English and Spanish, and Spanish was the first language for all three of them.

The final step in preparation for implementation was to hire, assign, and train the mentors who would be responsible for providing ongoing support for classroom staff. As noted earlier in the report, by design, approximately equal numbers of centers participating in the study were assigned to each of the two central agencies, and random assignment to treatment and control groups was done within the two groups, so that each agency had 18 centers in each of the intervention groups (and 27 centers in the control group). As part of their overall contract with the ELC to support quality improvement, negotiated in Summer 2003, each of the agencies was responsible for the salaries of three of the six mentors (two per intervention) who were hired. The interviewing and hiring of the mentors, however, was conducted by the ELC Project Manager assigned as a liaison between the ELC, the developers, the Abt study team and the centers participating in the study.

Required qualifications for the mentor position included an educational background in early childhood education (bachelor's degree or higher) and some experience in an early childhood care and education setting. In addition, at least half of the mentors needed to be bilingual in Spanish and English. Two of the mentors hired had master's degrees in education and had taught kindergarten, first and second grades in the Miami-Dade Public Schools. One mentor was working on her master's degree at FIU while working in the school system. The three remaining mentors had undergraduate degrees in ECE and two had experience working with one of the central agencies and their centers on child assessment and quality improvement. Three of the six were bilingual. Once hired, the mentors were divided into two groups, bilingual/monolingual English and then randomly assigned to one of the curricula, so that each curriculum had one bilingual mentor who could work with teachers who were monolingual in Spanish.

The mentor training, conducted by each developer and their on-site coach/trainer in late Fall 2003, was a two-day training that combined an introduction to the elements of each curriculum with a focus on the approach to mentoring that each developer prescribed.[22] There were many commonalities among the three interventions in their approach to mentoring. All three emphasized that the mentor would observe classroom staff, model activities and strategies, and continuously elicit feedback from teachers on aspects of the curriculum that were working better than others and on areas in which they needed more help. While the BELL mentor training envisaged a consistent mentoring strategy across classroom staff with different backgrounds, both RSL and, especially, BTL emphasized the need for flexibility to ensure that they met the needs of classroom staff with a wide range of education and experience. Trainers urged mentors to spend time initially observing teachers in the classroom and to use their observations to

[22] The mentors participated in the training sessions for classroom staff held later in the year, solidifying their understanding of the curriculum they were to support.

suggest and model activities that were feasible for an individual teacher and fit into whatever program or schedule she was currently using. The more individualized approach stressed by BTL recognized that not all classroom staff would be equally comfortable with the computer software or equally able to grasp more complex concepts, at least initially.

Training Classroom Staff in the Curricula

Each curriculum developer conducted three major training sessions for classroom staff. The first of these was a two-day training session for classroom staff and directors in late November/early December of 2003. The second set of training sessions was held in late January/early February of the following year. This schedule allowed the trainers to introduce the curriculum in a sequence, with more complex material covered in the second session, after classroom staff had had a chance to absorb and begin to implement the initial material and associated activities. The third training session, held in August of 2004, was intended as a refresher training for staff remaining in their classrooms as well as training for newly-hired classroom staff.

The training sessions represented a substantial effort on the part of developers, with national staff at BTL and RSL training sessions and the original authors at the BTL and BELL sessions. In addition, training sessions were set up so that at least part of the content was delivered in Spanish for teachers who had indicated a preference for being trained in Spanish. Developers used different strategies to accomplish this: one provided a translator, who sat at a table with the Spanish-dominant teachers; another had the site coordinator/trainer do a parallel translation of the whole training; the third had no whole-group training in Spanish, but had the bilingual site coordinator/trainer facilitate one of the training rotation stations. In addition, some teacher materials were provided in Spanish: one developer (B.E.L.L) translated all the training materials, the other two each provided translations of at least one type of teacher resource material (lesson extensions in one case (RSL), and the teacher's guide in the other (BTL)).

As with the training for mentors, very little of the training was didactic, although trainers for each curriculum began the initial session with an overview of the rationale underlying each element of the curriculum, including the research supporting it. Beyond that introduction, the sessions were highly interactive with many hands-on activities for trainees. In the judgment of study staff who observed the training sessions, all three were equally effective in engaging participants.

The neat training schedule conceals the reality that training was ongoing throughout the 18-month implementation period. The decision to train both directors and aides in addition to teachers was an astute one, because it ensured some continuity as teachers left their jobs. However, replacement teachers had to be trained and, to achieve this in a timely fashion, on-site trainers and mentors had to schedule individual training sessions. In addition, not all classroom staff were able to attend the scheduled training sessions, so that additional group training sessions had to be scheduled close to the planned session. No provision was made by the ELC to pay centers for substitute staff while classroom staff attended training sessions so that, while attendance was high at the initial session, some directors were reluctant or unable to release staff for subsequent training sessions. It seems likely that a stipend to cover the cost of substitute staff would have reduced the need for additional group training. However, the challenge of training new staff is probably irreducible.

The Mentoring Process

Each curriculum was assigned two mentors, paid for by the ELC, and supervised by on-site coach/trainers employed by the developers. Each mentor was responsible for approximately 18 classrooms, which she visited twice a month, on average. After the initial round of visits, the mentors, in consultation with their trainers, designed more flexible schedules that reflected the individual needs of classroom staff. Some classrooms required visits weekly or even more frequently, while others in which staff were moving quickly in implementing the curriculum required less frequent visits. The site coordinators also conducted mentoring visits, especially to new teachers or to teachers who were experiencing difficulty implementing the curriculum. The visits were similar across curriculum models, with each mentor visiting one or two classrooms in a morning, one or two classrooms in the afternoon, and completing paperwork at the end of the day. Each team developed a systematic way of recording and rating implementation progress and providing instructional feedback to teachers. The forms used by the coaches reflect the developers' ideas about key components of the curriculum and effective strategies to communicate them. They were used to identify specific areas for teachers to work on, such as conducting more activities in small groups, spending less time in whole-group activities, using graphic organizers to build vocabulary from the book of the week, and strategies for classroom management to help children focus.

In the Spring and Fall of 2004, Abt staff accompanied both mentors for each of the interventions on mentoring visits and interviewed the mentors after the visits. The mentors and their coaches were allowed to select the classrooms to be visited and we proposed that they select, in each case, classrooms where implementation of the curriculum was proceeding as planned. These observations were not meant to judge how well the curriculum was being implemented, since the mentors and trainers were much better judges of that, but to be able to describe, for each intervention, what a relatively-well implemented version would look like and what role the mentors played.

Ready, Set, Leap! In a visit to one classroom in early Fall 2004, children were engaged in writing activities in small groups. Some children were seated at tables writing first their own name and then a friend's name. Others were using LeapFrog lighted writing surfaces to trace letters and words. In this classroom, the teacher conducted the class in Spanish for the most part, but struggled to conduct the letter-naming activity in English. The class was large – 22 children, — and no aide was present. Although this was only ten days into the program year, the four-year-olds, who all spoke Spanish as their first language, could identify all 26 upper-and lower-case letters of the alphabet in English, and did so with great pride. The mentor explained that RSL viewed this competence as an easy one to be mastered early in the program year so that children could move onto more challenging activities.

In another class observed, two children worked together on letter-sound identification at the Leap Desk, which was one of several activity centers in the classroom. In this classroom, all the RSL technology was grouped in the one center and the teacher rotated children in small groups through the center. After the children had spent some time in the activity centers, the teacher gathered them for circle time and engaged them in a discussion of the picture book about Emperor penguins that she had earlier read aloud to them. In another classroom, no children were using the technology during the observation, but the teacher was working with a group of 10 children on letter-sound recognition, while the aide worked with the other half of the class on phonics (word families: -at and –un). Both groups were using teacher-made materials rather than materials provided by the developer.

BELL. In one classroom visited, the two teachers divided the class into two groups of ten children and conducted simultaneous sessions so that, while one group was engaged in shared reading, the other group was engaged in a phonological awareness/letter knowledge activity. The groups switched teachers and activities in the afternoon. In the shared reading group, the children listened to a story on tape while the teacher turned the pages of the book. At intervals, she paused the tape for comprehension checks, asking the children questions about the story. At the end, the teacher repeated the story, with the children saying the lines of the story out loud. In the other group, the teacher was playing "thumbs up/thumbs down" – children put their thumbs up when the teacher said a pair of rhyming words and down when the pair did not rhyme. They then explored an alliteration chart (poem) that focused on words that begin with the letter L. The teacher used this opportunity to review print concepts, asking questions such as "where do I start reading next, after I finish this line?" and "who can show me where a sentence begins?" The teacher also had children hold plastic snapping beads and push them together as they formed compound words, pulling them apart as they broke the words into their two separate components.

In another class, working on the same activities in a different order, the mentor had identified two classroom management strategies for the teacher to work on:

- When the children stand up or move around and interrupt an activity such as read-aloud, stop the reading and wait until they are all listening;
- Collect the "break-it/make it" manipulatives when the activity is over so that the children can focus on their next activity.

After the "make it/break it" activity (children held the plastic balls and broke them apart as they broke a word like "sunflower" into "sun' and "flower" and pushed them together to make "haystack" from "hay" and "stack"), the teacher collected the plastic manipulatives and began working with the same alliteration chart focused on the letter L. The children were very engaged in this activity, so much so that several stood up and moved close to the chart, blocking it from view. The teacher ignored them and continued with the lesson. So she had used one of the mentor's recommendations but not the other.

Breakthrough to Literacy. In one classroom visited, there were several activity groups operating. In one, children were engaged in dramatic play. In the main part of the classroom, a small group was writing about one piece of the Book of the Week (*This Old Man/He Played Paddywhack*), using a picture prompt from the book. Two children were copying a line of text from the book, one was writing about one aspect of the text ('he gave the dog a bone", one had added more detail to the picture and described the new scene ("it was a beautiful day"). At another table, children were molding play-doh into the shapes of items in the story (bones, a knee, a shoe, a thumb); at another table, children were counting out paper bones to match the number written on a card. At the sand table, one boy was digging up numbered bones, trying to find all 20. On the rug, a girl was putting story cards in sequence. Two children were working at computers on different activities related to the BOW. The song "This Old Man" was playing on a CD player. The teacher sat with the writers while the aide circulated, talking to the play-doh group, the math group and checking on the computer users. After a few minutes, one of the children completed his time on the computer, moved his name-card and notified the child whose name-card was next in line. As the writers finished their work, they moved to a different activity center.

In this classroom, the mentor had been working with the teacher for most of the year on shifting from whole-group to small-group and individual activities. During the whole of the observation, children spent the time in five or six small groups. However, the teacher still had some misgivings and wasn't convinced that the small groups were effective. In this case, the teacher was following the mentor's recommendation but still getting used to the feel of a different teaching approach.

Moving to Full Implementation

To assess the extent to which the interventions were fully implemented, and at what point that was true, study staff had to rely on assessments made by mentors and coach/trainers. While, in the second year of the study, an attempt was made to impose a rating scale across all three curricula, this effort was completely unsuccessful, because each curriculum team had its own metric by which it measured implementation. Each intervention team developed a systematic way of recording and rating implementation progress (and providing instructional feedback to teachers). These forms, used by the mentors, reflect each developer's ideas about key components of the curriculum and strategies or instructional approaches. Two models used five-point scale ratings; one used a 35-point rating which was arithmetically converted to a five-point scale.

Data from each model's implementation rating scale were analyzed separately. While curriculum developers differed in their criteria for a "fully implemented" curriculum, the scales provided an estimate of the degree of implementation achieved by centers in each group. By the end of the first study year, six to seven months after the initial training sessions, key elements of all three curricula were being implemented in most classrooms. Two-thirds of BELL classrooms were fully implementing the curriculum by June 2004, and all but four classrooms were judged to have reached a minimally satisfactory level of implementation by that time. More than half (57%) of Ready, Set, Leap classrooms had reached full implementation by June 2004 and, as with BELL, only four classrooms had not reached at least a minimally satisfactory level of implementation. Breakthrough to Literacy was slower to achieve full implementation; only one-third were judged to have reached full implementation by June 2004 and eight were judged not to be even minimally satisfactory in terms of implementation. The reason for the difference in the rate of implementation is not clear, but it is likely that the finer-grained implementation measure used by BTL left fewer judgments to be made by mentors and may have been a more stringent measure of implementation than the five-point scale used by the other two.

In any case, by the end of the second study year, all three curricula were well-implemented in the majority of centers. Across the three interventions, a similar small number of centers (3 to 4) in each group were still not implementing the curricula at a satisfactory level. In some cases, this reflected a teacher, newly hired late in Year 2, who was not sufficiently familiar with the curriculum; in others, there was resistance on the part of the director or the teacher or both. Mentors were not allowed to stop visiting these resistant teachers, but often the on-site coordinator assumed responsibility for regular visits to attempt to change practices.

Characteristics of Successful Implementers

In interviews, mentors from all three models reported independently the same features of successful implementers:

- **A positive attitude towards instructional change.** Mentors reported that perhaps the most important aspect of successful centers was willingness on the part of both directors and classroom staff to refocus on specific literacy goals and adjust their space and routines to support those goals;
- **Effective management of time.** All of the mentors noted that the ability to organize lesson plans and manage time to accommodate all the planned activities were hallmarks of classrooms that could successfully implement a curriculum;
- **Well-organized classroom space and effective classroom management.** Once classroom materials have been set up and are available to children, the challenge is to cycle children through the different areas so that they are exposed to a variety of materials and activities over the course of a day and a week. High-implementing classrooms had effective strategies for handling transitions and helping children to take turns. For example, in one classroom, children placed cards with their names and a representative icon into a labeled pocket in the area where they chose to play. In another, children chose a song to play when it was time to move to another center. In many of the classrooms, at least initially, the mentors helped develop classroom schedules and provided lesson plans as a guide, intending to wean teachers from them as they became more competent in managing the classroom. In the end, however, most continued to provide lesson plans as way of maintaining a clear focus.

  Trainers and coaches for all three interventions noted that, in a non-experimental situation, they would have assessed the teacher's classroom management and, for those who lacked those skills, would have delayed training on the curriculum until the teacher had received some basic training in managing time, space and children's activities in the classroom.
- **Healthy working relationships among director, staff and parents.** Implementation was smoother and more complete when directors trusted teachers to act as instructional leaders and make changes where necessary, with guidance from mentors. Directors' involvement in classroom activities varied greatly; some taught regularly and visited the classroom often. Others were more concerned with running the center as a business and focused on those aspects of administration, visiting the classroom infrequently.
- **Frequent one-on-one interactions between teachers and children.** In classrooms where the teacher was already practiced in making meaningful contacts with individual children, full implementation was more likely. In one classroom, the teacher used greeting time, as children gathered on the playground in the morning, to have brief conversations with each child. She

treated these conversations as informal conferences to reinforce classroom routines and activities, help children plan their activities and encourage them individually. Conversations described actions, experiences and events and the teacher listened and responded to children's comments and suggestions. Another teacher worked hard to maintain encouragement and one-on-one engagement with children at all ability levels, reinforcing the literacy goal positively whether answers given orally were correct or incorrect. In another classroom, some children were able to write names and addresses with or without adult help, while others made letter-like marks on the page. The teacher provided positive feedback to each child, regardless of ability level, reinforcing literacy goals as she did so.

Barriers to Implementation

The mentors reported similar barriers to implementation across the three models:

- **Resistance to instructional change.** Some teachers were resistant to using the lesson plans offered by the mentors but had no alternative strategies. Some were reluctant to send books home with the children for fear that the books would be damaged or lost. For the same reason, early in the study, some teachers kept the curriculum materials locked away, because they worried that they would be blamed if books, materials or equipment were torn or damaged.
- **Lack of trust and cooperation between teachers and administrative staff.** Some directors were protective of their role as teacher supervisors and felt threatened by the mentor's role as instructional guide. In these cases, the teachers often found themselves caught between the two, and mentors and trainers needed to meet repeatedly with directors to build trust and understanding.
- **Teachers' difficulty in making the transition from Spanish-language to English-language instruction.** Many monolingual-Spanish teachers implemented the curricula very successfully, some using Spanish materials during group instruction and encouraging children to use books, materials and technology that supported English-language learning in individual and small-group activities. However, these teachers were limited in their ability to support English oral language in classroom discussions. In one classroom with a monolingual Spanish teacher, Spanish-speaking children were observed talking to each other in English but could only communicate with the teacher in Spanish.
- **Teacher turnover.** Teacher turnover is high in centers in Miami-Dade County – by some estimates as high as 45 percent per year. The problem was only slightly ameliorated by the retention stipends offered to all teachers. Over the two years of the study, teacher turnover was 28% in RSL classrooms, 42 % in BTL classrooms and 44% in B.E.L.L. classrooms (in control classrooms, two-year turnover was 49%). Most of the teacher turnover in B.E.L.L. classrooms occurred in the first year; in Year 2 of the study, turnover was only 5%. For the other two curricula and for the control group, turnover rates were roughly the same for each of the two years. As noted earlier, the developers made appropriate provision for training replacement teachers. Because aides, and in many instances center directors, had been trained on the curricula, they were able to provide guidance for new teachers and ensure some consistency during the transition. However, the need for on-going training (as opposed to mentoring) was greater than developers anticipated and made considerable demands on the time of the on-site coordinators.

Exhibit 5-1

Strategies Used to Address Four Critical Aspects of Language/Literacy Development

Skill/Knowledge Area	**Curriculum**		
	RSL	**BELL**	**BTL**
Oral Language	• Read-aloud with **trade books** (rich natural language) o Post-reading discussions target: comprehension & vocabulary development activities o Frequency: Daily during whole-group/circle time; no set time limit • Independent reading of short stories with controlled (reduced) language on Leap Pad (**electronic storybooks**) • Theme important—many suggested follow-up activities based on theme related to target book (i.e., whatever book teacher reads aloud to whole group that week).	• Read-aloud with Wright Group books (**controlled (reduced) language books**) o Post-reading discussions target: comprehension & motivation o Frequency: 2-4 times/week, in 15-minute large (6+ children) group sessions • Theme important—many suggested follow-up activities based on theme related to target book (i.e., whatever book teacher reads aloud to whole group that week).	• Read-aloud with Wright Group books (**controlled (reduced) language books**)—Book of the Week (one book read aloud every day in week); teacher could supplement with other books o Post-reading discussions target: vocabulary development (primary) and comprehension o Frequency: Daily 10-15-minute sessions in large (6+ children) group • Independent reading/virtual shared reading with same Book of the Week and other (Wright Group) books on computer (electronic storybooks) • Theme important—many suggested follow-up activities based on theme related to target book (i.e., whatever book teacher reads aloud to whole group that week (Book of the Week)).
Print Knowledge	• Leap Desk: this tool enables the child to engage in self-correcting (that is, the tool provides feedback or the correct answer) activities aimed at learning letter shape (see, touch, trace, fit like puzzle); letter name (hear), letter sound (hear) • Leap Mat: this tool has the same self-correcting aspect and features letter name, letter sound, and beginning spelling • Leap Pad: electronic books on which children can control the pace of reading with a stylus	• Laminated charts with poems and short controlled language stories (e.g., in which target letters or sounds are over-represented): Direct instruction in letter identification (e.g, find the letter 'm'), directionality of print (e.g., left to right, top to bottom), concept of word & sentence, punctuation (e.g., first word in sentence starts with uppercase letter, sentence ends with period), o Frequency: 1-3 times/week in 15-minute large (6+ children) group sessions	• Computer (Individualized Software Instruction, ISI): letter identification (letter name), letter sound, concept of word, directionality of print o Frequency: daily @ 8-11 min./child • ISI also includes electronic storybook in which program tracks (highlights) print, highlights words, etc. Small group activities (letter ID, story cards) related to the Book of the Week, including using the small Take-me-home books (small newsprint versions of the Book of the Week that children could take home with them (identifying target letters or sight words in the book of the

Exhibit 5-1

Strategies Used to Address Four Critical Aspects of Language/Literacy Development

Skill/Knowledge Area	**Curriculum**		
	RSL	**BELL**	**BTL**
			week, identifying words (e.g., underline each word, to show the understanding of concept of word)
Phonological Sensitivity	• Leap Pad: electronic books featuring beginning and ending sounds of words and rhyming. • Teachers encouraged to include songs and rhymes in whole group activities	• Manipulatives and games involving breaking compound words into component words, breaking sounds in words apart and blending them together; Direct instruction o Frequency: 1-3 times/week in 15-minute large (6+ children) group sessions • Songs and rhymes included in 15-minute sessions and recommended for follow-up small group activities.	• ISI: beginning and ending sounds, rhyming and alliteration, breaking sounds of words apart and blending together; interactive software (program corrects child and provides more practice or provides more challenging tasks) o Frequency: daily @ 8-11 min./child
Print Motivation	• Read-aloud (with trade books); discussion: focus on motivation—relating books to children's experience or introducing toys for center time that are related to the theme • Leap Pad: electronic books featuring short stories, basic concepts (color, shape, position, number, etc.). • Developer considers that trade books inherently increase motivation.	• Read-aloud (Wright Group books)+ follow-up activities (2-4 times/week in 15-minute sessions) • Developer uses Wright Group books for convenience; no philosophical commitment or reason not to use other books, but units are based on WG books	• Read-aloud (Wright Group books); follow-up activities that seem fun and/or relate to children's experience • Developer considers work with books that children are able to read independently (even if "reading" rather than actually decoding) is inherently motivating.

Chapter 6: Analysis Strategy

Introduction

The analysis strategy for Project Upgrade was motivated and guided by the conceptual framework and research questions presented in Chapter 2. This chapter begins with a discussion of the size of the analytic samples used to estimate program impacts on child outcomes and teacher behavior and literacy environment. Subsequent sections describe how the outcome measures were created, the analytical models for estimating impacts on teacher behavior, the literacy environment, and child outcomes, the analytical approaches to subgroup analysis, and finally, the analytical models used for non-experimental analyses.

Analytic Samples

Analyses of impacts on teacher behavior and the literacy environment were based on data collected in the time-frame spanning the Fall of 2003 through Spring of 2005. Baseline data were collected in the Fall of 2003, prior to implementation of the experimental treatments. Data collected in Spring of 2004 and Spring of 2005 represent about six months and eighteen months of implementation, respectively. The measurements used to estimate impacts on child outcomes were collected in the Spring of 2005.

Baseline data on teacher behavior and the literacy environment were collected on 164 classes nested within 20 randomization blocks (described in Chapter 3). Within randomization blocks, centers were randomly assigned to each of the three treatment groups, or to the control group. Data were obtained from one class per center. Consequently, the numbers of classes and centers are always identical and the terms 'class' and 'center' are used interchangeably throughout this discussion. Over the two years of the study several centers were lost to attrition, resulting in analysis samples composed of 161 classes with data from year 2004, and 157 classes with measurements from year 2005. Exhibit 6-1 summarizes the number of classes in the analytic samples in each treatment group for each data collection year.

Exhibit 6-1

Sizes of Analysis Samples of Classes

Treatment Group	2003 (Baseline)	2004 (1 Year Post Implementation)	2005 (2 Years Post Implementation)
RSL	38	37	36
Bell	36	34	33
BtL	36	36	35
Control	54	54	53
Total:	164	161	157

Impacts on child outcomes were estimated using data from 1,535 children nested in 154 classes. These impact estimates correspond to children who were tested using the English language version of the child assessment instrument. Exhibit 6-2 shows information on the number of children per treatment group included in the analytic data set. In 2005, there were three classes in which classroom observations were made using the OMLIT instruments, but for which no child outcome measures were obtained. Enrollment in the study classrooms was lower in Spring 2005 than it had been in earlier years[23], so all children present in the classroom who had been in the classroom for at least two months, and whose parents had given permission for the child's assessment, were tested.

Exhibit 6-2

Size of 2005 Child Outcome Analysis Sample

Treatment Group	Number of Children per Treatment Group	Number of Centers per Treatment Group	Mean (Min., Max) Number of Children per Center per Treatment Group
RSL	320	36	9 (3,13)
Bell	346	33	10 (1,16)
BtL	355	35	10 (4,16)
Control	514	50	10 (4,18)
Total:	1535	154	10 (1,18)

Creation of Analysis Variables

Teacher Behavior and Literacy Environment Outcome Measures

To assess whether, the three interventions were successful in changing the teaching activities and literacy environment in the intervention classrooms, observations were conducted using a battery of measures (Observation Measures of Language and Literacy Instruction, or OMLIT; Goodson, Layzer, Smith & Rimdzius, 2004). Constructs were derived from the multiple OMLIT measures to correspond to key elements of the classroom that are being manipulated by the interventions. These included constructs for the four key components of emergent literacy, and the two literacy environment domains. A preliminary step in the creation of the four OMLIT teaching practices constructs involved the identification, *on a conceptual basis,* of the set of individual teaching practices from across the OMLIT battery of measures that, on the basis of research, are believed to be linked to children's development in that domain. Similarly, to create the two literacy environment constructs, we identified, *on a conceptual basis,* the set of environmental factors from across the OMLIT battery of measures that are believed to be related to the development of emergent literacy. These constructs are shown in Exhibit 6-3, together with the specific teaching behaviors or environmental supports that comprise each.

[23] The lower enrollment was a result of a temporary freeze in the intake for child care subsidies to avert potential overspending and affected centers in all four study groups.

Exhibit 6-3

Teaching Behaviors and Environmental Supports in OMLIT Constructs

OMLIT Construct	Specific Teaching Behaviors or Environmental Supports
Support for oral language development	Reading aloud: • Time, # books • proportion of read alouds with different supports for comprehension of text (5 types) • Proportion of read alouds with open-ended questions • Quality of open-ended questions, vocabulary supports, post-reading extensions Literacy activities: • Time on oral language activities • Proportion of oral language activities with small groups • Quality of teacher/child discussion • Overall rating of oral language support • Frequency of oral language activities • Quality of oral language activities
Support for print knowledge (letters, letter-sound correspondence, writing, concepts of print)	Reading aloud: • Proportion of read-alouds with discussion of print concepts • Classroom activities • Time in activities with text, letters • Time in activities with writing (copying, emergent) • Proportion text, writing activities in small groups • Proportion activities with print involved Literacy activities: • Time on print knowledge activities • Proportion of print knowledge activities with small groups • Time on emergent writing activities • Time on copying/tracing activities • Proportion of writing activity in small groups • Proportion of print knowledge activities with small groups • Overall rating of print knowledge support • Frequency of writing activities • Quality of writing activities • Frequency of print knowledge activities • Quality of print knowledge activities
Support for phonological awareness	Reading aloud: • Proportion of read alouds with discussion of sounds Literacy activities: • Time on sounds • Proportion of activities on sounds with small groups • Overall rating of quality of support for learning sounds: • Frequency of activities on sounds • Quality of activities on sounds
Support for print motivation	Reading aloud: • Proportion of read alouds with support for print motivation • Number of RAPs • Number of minutes of reading aloud Literacy activities: • Time on activities involving print motivation • Proportion of activities on print motivation with small groups

Exhibit 6-3

Teaching Behaviors and Environmental Supports in OMLIT Constructs

OMLIT Construct	Specific Teaching Behaviors or Environmental Supports
Literacy resources in classroom	• Adequacy of: • environmental print • text materials • writing resources • rich, integrated theme • literacy manipulatives • integration of print in other centers
Literacy activities in classroom	• teacher presents information/reads text • teacher writing • focused oral language activity • child(ren) reading/shared reading • child(ren) writing

The six outcome measures were created from individual items on the OMLIT measurement instrument as follows.

At each of the three data collection points, some of the OMLIT measures were collected once; others, like the SNAP and the RAP were completed several times in the course of the observation. The first step was to aggregate the multiple observations per item per class per year into a single item measure per class per year. The aggregated item score was calculated as the item mean across repeated observations.

The teaching behaviors and measures of classroom environment within each domain were on different scales--some were proportions of time, some were counts. Therefore, to build scales, we converted all of the OMLIT items (aggregated in the previous step) into standard scores with a mean of 0 and a standard deviation of 1. Preliminary outcome constructs were then calculated as the sum of the relevant standardized items. We then examined the internal consistency of the resulting scales using the Cronbach's alpha statistic. Items that diminished the reliability of the scale were omitted from subsequent versions of the construct. The process was repeated until the most reliable subset of items, from the bank of the original items considered for use in the construct, was obtained. The Chronbach's alpha statistics from the final scales are shown in Exhibit 6-4. The constructs with the fewest behaviors had the lowest internal consistency, as would be expected. We also computed Cronbach's alphas for the final constructs (derived from the reliability analyses) in a second OMLIT data set from 199 child care center classrooms in another study (CLIO). As shown in Exhibit 6-4, the Cronbach's alphas in the CLIO sample of classrooms were very similar to those for the current study.

Exhibit 6-4

Reliability of OMLIT Constructs : Internal Consistency and Inter-Rater Reliability

Construct	# Items in Final Scale	Cronbach's Alpha CLIO (n = 199 classrooms)	Cronbach's Alpha Miami[a] (n = 161 child care center classrooms)	Inter-Rater Reliability (n = 33 paired observations of CLIO classrooms)[b]
Oral language	14	.84	.80	.87
Print knowledge	16	.84	.82	.89
Phonological awareness	4	.58	.61	.83
Print motivation	5	.66	.60	.89
Literacy resources in class	7	.75	.73	.80
Literacy activities in class	4	.74	.74	.80

[a] In Miami data, Cronbach's alpha derived from same set of OMLIT variables that are included in the final version of constructs derived from the CLIO data

[b] The reliabilities shown here represent the range of inter-rater reliabilities for the component variables in each construct. The inter-rater agreement on the final OMLIT constructs will be calculated for this exhibit.

A scale score for each of the six outcome constructs was created for each class for each year by summing the relevant standardized items, as previously described. The final step involved re-scaling each of the constructs to a more convenient metric. The rescaling was such that the Year 2004 control group mean and standard deviation for each of the six constructs was 50 and 10, respectively. This rescaling enhanced the interpretability of results as in the following example. The Year 2004 control group mean and standard deviation for the construct *Support for Oral Language* was 50, and 10 respectively. For the treatment group *Ready, Set, Leap!*, the Year 2004 mean for the construct *Support for Oral Language* was 57.2. Thus, this treatment group scored 7 points higher than the control group on this outcome measure, which corresponds to 7.2/10 standard deviation units, or an effect size equal to 0.72. The Year 2005 and the Year 2003 constructs were also standardized relative to the Year 2004 control group means. Thus, for example, the score of 49.8 observed for the control group in the Year 2005 for the construct *Support for Oral Language*, is interpreted as representing a decrease of 0.02 standard deviation units from spring 2004 to spring 2005.

We note that additional items were added to the OMLIT observation instrument between the Year 2003 and Year 2004 data collection cycles. Some of the items that were used in the construction of the 2004 and 2005 construct scales were not available in the 2003 data. Therefore the 2003 scales were created from the available subsets of items that were used in the 2004 and 2005 scales. Thus, even though the Year 2003 constructs were scaled relative to the Year 2004 control group means, differences between the Year 2003 means and the Year 2004 control group means may be due in part to the differences in the items used to create the scales.

The steps for re-scaling the constructs relative to the Year 2004 control group means were as follows. For each data year, each of the six constructs was created as the sum of relevant standardized items. Next, the 2004 control group mean and standard deviation was calculated for each OMLIT construct. Then, each construct for each year was standardized by subtracting the 2004 control group mean and dividing by the 2004 control group standard deviation of the construct. After completion of this step, the 2004 control group mean and standard deviation were zero and one, respectively. Each construct was then rescaled by multiplying by 10, and adding 50. After completion of this step, the 2004 control group mean and standard deviation were 50 and 10, respectively. The resulting scores are such that any mean can be interpreted relative to the 2004 control group mean.

The correlations among the six constructs are shown in Exhibit 6-5.

Exhibit 6-5

Correlation Among 2004 OMLIT Scores

Constructs	**Oral Language**	**Print Knowledge**	**Phonological Awarenes**	**Print Motivation**	**Literacy Resources**	**Literacy Activities**
Oral language		.39***	.30***	.49***	.28***	.51***
Print knowledge			.42***	.22***	.30***	.77***
Phonological awareness				.11	.14	.47***
Print motivation					.05	.37***
Literacy resources in class						.28***
Literacy activities in class						

*** = p<.001, ** = p<.01, * = p<.05, NS = not significant.

Child Outcome Measures

Child outcomes measures were composed of four scale scores from the Test of Preschool Emergent Literacy (TOPEL) assessment instrument. At the time that the current study was designed and data collection was underway, the TOPEL instrument had not yet be finalized and normed. A pre-cursor to the TOPEL (the Pre-CTOPP for the Preschool Comprehensive Test of Phonological and Print Processing) was available for use and was administered to the children in this study. While the study was underway, the test developer, Pro-Ed, was in the process of norming the scales. The norming data was expected to be available by late 2005. As promised, Pro-Ed released the norming data in spring 2006, in time for the experiment to convert the Pre-CTOPP results to the TOPEL standardized scores for both analysis and to characterize the developmental status of the sample children in comparison to a national sample of children of similar age.

The procedure for converting raw TOPEL scores into standardized scores is straightforward given the child's chronological age and "Raw Scores to Percentile Ranks and Standard Scores" conversion tables provided in the *Test of Preschool Early Literacy Examiner's Manual*[24]. However, the conversion of the Pre-CTOPP test results into scores that are as nearly equivalent as possible to the raw TOPEL scores requires some explanation.

All of the test items on the TOPEL assessment instrument were administered as part of the Pre-CTOPP instrument. However, the Pre-CTOPP instrument included items that are not administered in the TOPEL. Furthermore, there were several differences between the two instruments in the order that items were

[24] For information, go to www.proedinc.com

administered. And finally, the following important difference between the two instruments in the administration instructions had to be considered. In the administration of the Pre-CTOPP, all items within each subtest were administered to a child, regardless of the number of items (s)he answered correctly. The administration instructions for the TOPEL are such that, when a child gives incorrect responses to three items in a row, no additional items on the subtest are to be administered, and all remaining items are to be scored as zeros (incorrect).

In order to create TOPEL raw scores from the Pre-CTOPP data, we re-ordered the item responses in our data file to match the order of administration of items in the TOPEL. Items that are not used in the TOPEL were ignored. With the newly re-ordered items, we looked for any instances where three items in a row were incorrect. Whenever that occurred, we set all remaining items in the newly ordered sequence to zero. The raw score for each subtest was then calculated as the sum of the item scores in the subtest, where a correct item takes the value 1, and incorrect takes the value zero. The final step was the conversion of raw scores to standard scores, resulting in the four previously described child-level outcome measures: *Definitional Vocabulary, Phonological Awareness, Print Knowledge*, and *Early Literacy Index*. TOPEL scores are standardized so that the population mean and standard deviation are 100 and 15, respectively.

Measures Used as Covariates or as Descriptors of the Sample

The *Arnett Caregiver Rating Scale* (Arnett, 1989) was completed for the lead teacher in each classroom at baseline (Fall of 2003) and each follow-up data collection point (Spring of 2004 and Spring of 2005). The instrument produces ratings on the caregiver's emotional tone, discipline style, supervision of and interest in children, and encouragement and independence. Scores were produced for three subscales Positive Affect, Not Punitive, and Engaged (opposite of detached), and a total scale created from the three subscales. Using the same process as described previously for the OMLIT scales, the scores on each subscale and the total score were re-scaled so that the 2004 control group had a mean of 50 and a standard deviation of 10. Scores for treatment groups or the control group from other years can be interpreted relative to the 2004 control group mean. The three subscale scores from 2003 were used to test for baseline equivalence among treatment and control groups. The 2003 Arnett total score was used as a covariate in models used to estimate treatment impacts on teacher behavior and literacy environment (OMLIT outcomes).

The *LAP-D* is a broad diagnostic screening measure. It was administered to four-year-olds receiving subsidies by staff from the county agency that provides resource and referral services, and administers subsidies. The *LAP-D* data collected in fall 2003 were provided to the study by the School Readiness Coalition. Child-level scores were used to create baseline class-level mean *LAP-D* cognitive total scores, which were used as covariates in models of the treatment impact on child-level TOPEL outcomes. The 2003 *LAP-D* scores were also used to evaluate the baseline equivalence of treatment and control classrooms.

The *education level* of the lead teacher for each class was obtained from a self-administered staff questionnaire administered by the ELC in Fall 2003. For the purpose of baseline equivalence testing, education level was coded into three exhaustive and mutually exclusive categories: high school only, some college, and Bachelor's degree or Associate's degree. For analyses relating teacher education level to measures of teacher behavior and class environment, a dichotomously coded variable was used that took the value 1 if the teacher had a Bachelor's degree, and took the value 0 otherwise.

Subgroup analyses were conducted on groups of teachers defined as *Spanish-dominant* and *English-dominant*. Prior to randomization, teachers were asked what language they would prefer to be trained in. Their response to the question formed the basis for the *Spanish-dominant* vs *English-dominant* dichotomy.

Covariates used in models of treatment impacts on child-level outcomes (TOPEL measures) included the *child's age* at time of testing, the *sex* of the child, and a measure of the primary language spoken in the child's home. Child's home language was coded into three mutually exclusive and exhaustive categories: English only; Spanish only or mix of English and Spanish; and other

Analysis Methods

Baseline Balance Tests

Baseline balance tests were conducted to answer the question of whether the treatment and control groups were equivalent at baseline on:

- The *Cognitive Total, Language Total*, and *Fine Motor Total* subscales of the *LAP-D*
- The following measures of teacher behavior derived from the OMLIT - *Support for Oral Language*, *Support for Print Knowledge,* and *Literacy Resources*
- The *Arnett* subscale measures of *Positive Affect, Not Punitive*, and *Engaged*
- Proportion of teachers preferring training in Spanish
- Teacher education level

Baseline equivalence of treatment and control groups on *LAP-D*, *OMLIT*, and *Arnett* measures was assessed using two-level hierarchical models where classrooms (level 1) were nested within randomization blocks (level 2). Models were of the form:

Level-1 Model:

$$Y_{(2003)jk} = \beta_{0k} + \beta_{1k}(Trt_{jk}) + r_{jk}$$

Level-2 Model:

$$\beta_{0k} = \gamma_{00} + u_k$$

$$\beta_{1k} = \gamma_{10}$$

$$r_{jk} \sim N(0, \sigma^2)$$

$$u_k \sim N(0, \tau_{00})$$

where

$Y_{(2003)jk}$	=	*LAP-D*, *OMLIT* construct, or *Arnett* measure from year 2003 observation of classroom *j* nested in block *k*.
Trt_{jk}	=	1 if classroom *j* nested in block *k* was in Treatment Groups 1, 2, or 3;
	=	0 if control group.

The parameter estimate $\hat{\gamma}_{10}$ from the model above is the estimated difference between treatment and control groups at baseline. The value of $\hat{\gamma}_{10}$ is entered as results in the summary table (Exhibit 3-3) in the column labeled "Mean Difference T-C". In the results summary table, the values shown in the column labeled "Control Mean (SD)" were calculated as the simple mean and standard deviation of the *LAP-D*, *OMLIT*, or *Arnett* measures of the n=55 classes in the control group. The value of the mean shown in the column labeled "Treatment Mean (SD)" was calculated as the sum of the estimate of the treatment-control difference, $\hat{\gamma}_{10}$, and the control group mean. The treatment group standard deviation shown in the exhibit was calculated as the standard deviation of the *LAP-D*, *OMLIT*, or *Arnett* measures of the n=110 classes in the combined group of the three treatment groups. The effect size was calculated by dividing the treatment-control difference, $\hat{\gamma}_{10}$, by the Year 2004 control group standard deviation. The p-value corresponds to a two-sided test of the null hypothesis that the treatment effect is equal to zero.

Baseline equivalence of treatment and control groups on *LAP-D*, *OMLIT*, or *Arnett* measures was also assessed for subgroups consisting of classes with either Spanish dominant or English dominant teachers. These tests were conducted by subsetting the data to the appropriate subgroups and fitting the model described above to the subset of data. There were no significant differences at baseline for either of the two subgroups on these measures.

Baseline equivalence of the proportion of teachers preferring training in Spanish, and education level were assessed using chi-square tests of independence.

Estimation of Impacts on Teacher Behavior and Instructional Practices

Year 1 (Spring 2004) and Year 2 (Spring 2005) impacts on teacher behavior and instructional practices were estimated to obtain:

- The averaged effect of all three treatment groups contrasted with control
- The estimated effects of each of the three treatments contrasted with control
- Subgroup Analyses: Impacts on classes with Spanish-dominant teachers
    - The averaged effect of all three treatment groups contrasted with control
- Subgroup Analyses: Impacts on classes with English-dominant teachers
    - The averaged effect of all three treatment groups contrasted with control

Exhibit 7-1 summarizes results from models for Year 2004 OMLIT construct outcomes, where all three treatment groups combined were contrasted with the control group. The data were analyzed in two-level hierarchical linear models where classrooms (level-1) were nested in randomization blocks (level-2). The models included a random intercept term for blocks. Treatment impacts (any of the three treatment groups contrasted to control) were estimated in models that controlled for year 2003 baseline OMLIT construct measures,[25] and year 2003 baseline value of the Arnett "positive, not punitive,not detached" construct. The models were specified as shown below.

Level-1 Model:

$$Y_{(2004)jk} = \beta_{0k} + \beta_{1k}(Trt_{jk}) + \beta_{2k}(Y_{(2003)jk}) + \beta_{3k}(Arnett_{(2003)jk}) + r_{jk}$$

[25] This term was omitted from models for phonological awareness and literacy activities because those measures were not available from the 2003 classroom observational data.

Level-2 Model:

$$\beta_{0k} = \gamma_{00} + u_k$$

$$\beta_{1k} = \gamma_{10}$$

$$\beta_{2k} = \gamma_{20}$$

$$\beta_{3k} = \gamma_{30}$$

$$r_{jk} \sim N(0, \sigma^2)$$

$$u_k \sim N(0, \tau_{00})$$

where

$Y_{(2004)jk}$ = OMLIT construct from year 2004 observation of classroom *j* nested in block *k*.

$Y_{(2003)jk}$ = OMLIT construct from year 2003 observation of classroom *j* nested in block *k*.
(This term was omitted from models for *phonological awareness* and *literacy activities* because those measures were not available from the 2003 classroom observational data.)

Trt_{jk} = 1 if classroom *j* nested in block *k* was in Treatment Groups 1, 2, or 3;
= 0 if control group.

$Arnett_{(2003)jk}$ = Arnett "positive, punitive, detached" construct from year 2003 observation of classroom *j* nested in block *k*

The parameter estimate $\hat{\gamma}_{10}$ from the model above is the estimated treatment effect. The value of $\hat{\gamma}_{10}$ is entered Exhibit 7-1 in the column labeled "Mean Difference T-C". In Exhibit 7-1, the values shown in the column labeled "Control Mean (SD)" were calculated as the simple mean and standard deviation of the OMLIT construct values of the n=54 classes in the control group. The value of the mean shown in the column labeled "Treatment Mean (SD)" was calculated as the sum of the treatment effect, $\hat{\gamma}_{10}$, and the control group mean. The treatment group standard deviation was calculated as the standard deviation of OMLIT construct values of the n=107 classes in the combined group of the three treatment groups. The effect size was calculated by dividing the treatment effect, $\hat{\gamma}_{10}$, by the Year 2004 control group standard deviation. The p-value corresponds to a two-sided test of the null hypothesis that the treatment effect is equal to zero.

Year 2 (Spring 2005) OMLIT construct outcomes were analyzed in a similar model, the only difference being that the outcome variables were the 2005 measures, i.e.,

$Y_{(2005)jk}$ = OMLIT construct from year 2005 observation of classroom *j* nested in block *k*.

All other model terms were as specified for the model for the spring 2004 outcomes. The effect sizes for the 2005 outcomes were calculated by dividing the 2005 impact by the Year 2004 control group standard deviation.

Subgroup analyses were conducted by creating two separate subsets of data, one composed of data from classes with Spanish-dominant teachers, the other composed of classes with English-dominant teachers. Impacts for these subgroups were estimated from the same model as specified above, fit to data from a subgroup. The denominator used in the calculation of all effect sizes was the Year 2004 full sample control group standard deviation.

In order to estimate impacts of each of the three treatments, the previously described model was modified to include three treatment dummy variables that contrasted each of the three treatments to control. The models were specified as shown below.

Level-1 Model:

$$Y_{(2004)jk} = \beta_{0k} + \beta_{1k}(Trt1_{jk}) + \beta_{2k}(Trt2_{jk}) + \beta_{3k}(Trt3_{jk}) + \beta_{4k}(Y_{(2003)jk}) + \beta_{5k}(Arnett_{(2003)jk}) + r_{jk}$$

Level-2 Model:

$$\beta_{0k} = \gamma_{00} + u_k$$
$$\beta_{1k} = \gamma_{10}$$
$$\beta_{2k} = \gamma_{20}$$
$$\beta_{3k} = \gamma_{30}$$
$$\beta_{4k} = \gamma_{40}$$
$$\beta_{5k} = \gamma_{50}$$
$$r_{jk} \sim N(0,\sigma^2)$$
$$u_k \sim N(0,\tau_{00})$$

where

$Y_{(2004)jk}$ = OMLIT construct from year 2004 observation of classroom *j* nested in block *k*.

$Y_{(2003)jk}$ = OMLIT construct from year 2003 observation of classroom *j* nested in block *k*. (This term was omitted from models for *phonological awareness* and *literacy activities* because those measures were not available from the 2003 classroom observational data.)

$Trt1_{jk}$ = 1 if classroom *j* nested in block *k* was in Treatment Group 1; = 0 else.

$Trt2_{jk}$ = 1 if classroom *j* nested in block *k* was in Treatment Group 2; = 0 else.

$Trt3_{jk}$ = 1 if classroom *j* nested in block *k* was in Treatment Group 3; = 0 else.

$Arnett_{(2003)jk}$ = Arnett "positive, punitive, detached" construct from year 2003 observation of classroom *j* nested in block *k*

The parameter estimates $\hat{\gamma}_{10}, \hat{\gamma}_{20}, \hat{\gamma}_{30}$ from the model above are the estimated impacts of Treatments 1, 2, and 3, as contrasted to control, respectively.

Estimation of Impacts on Child Outcomes

Year 2 (Spring 2005) impacts on child outcomes were estimated to obtain:

- The averaged effect of all three treatment groups contrasted with control
- The estimated effects of each of the three treatments contrasted with control
- The averaged effect of Treatments 1 and 3 contrasted with control
- Subgroup Analyses: Impacts on child outcomes for children with Spanish-dominant teachers
    - The averaged effect of Treatments 1 and 3 contrasted with control
- Subgroup Analyses: Impacts on child outcomes for children with English-dominant teachers
    - The averaged effect of Treatments 1 and 3 contrasted with control
- Subgroup Analyses: Impacts on child outcomes for children with Spanish or Haitian Creole as their home language

  - The averaged effect of Treatments 1 and 3 contrasted with control
- Subgroup Analyses: Impacts on child outcomes for children with English as their home language
  - The averaged effect of Treatments 1 and 3 contrasted with control

Impacts on Year 2005 child-level outcomes (TOPEL scores) were estimated in three-level hierarchical linear models where students (level-1) were nested in classrooms (level-2), and classes were nested in randomization blocks (level-3). The models included random intercept terms for classes and blocks. Treatment impacts were estimated in models that controlled for child's age, sex, and language spoken at home, and for classroom-level mean *Lap-D Cognitive Total* scores obtained from measurements taken in the Fall of 2004 or the Fall of 2003 (for the small number of classrooms for which the 2004 score was not available).

Models where all three treatment groups combined were contrasted with the control group were of the form specified below.

Level-1 Model:

$$Y_{(2005)ijk} = \pi_{0jk} + \pi_{1jk}(Age_{ijk}) + \pi_{2jk}(SexMale_{ijk}) + \pi_{3jk}(HomeLang1_{ijk}) + \pi_{4jk}(HomeLang2_{ijk}) + e_{ijk}$$

Level-2 Model:

$$\pi_{0jk} = \beta_{00k} + \beta_{01k}(Trt_{jk}) + \beta_{02k}(MeanLapD\_CT_{jk}) + r_{jk}$$

$$\pi_{1jk} = \beta_{10k}$$

$$\pi_{2jk} = \beta_{20k}$$

$$\pi_{3jk} = \beta_{30k}$$

$$\pi_{4jk} = \beta_{40k}$$

Level-3 Model:

$$\beta_{00k} = \gamma_{000} + u_k$$

$$\beta_{01k} = \gamma_{010}$$

$$\beta_{02k} = \gamma_{020}$$

$$\beta_{10k} = \gamma_{100}$$

$$\beta_{20k} = \gamma_{200}$$

$$\beta_{30k} = \gamma_{300}$$

$$\beta_{40k} = \gamma_{400}$$

$$e_{ijk} \sim N(0,\phi^2)$$

$$r_{jk} \sim N(0,\sigma^2)$$

$$u_k \sim N(0,\tau_{00})$$

where

$Y_{(2005)ijk}$ = TOPEL outcome measure from spring of 2005 for student *i*, nested in classroom *j* nested in block *k*.
Age_{ijk} = Age at time of testing of student *i*, nested in classroom *j* nested in block *k*.
$SexMale_{ijk}$ = 1 if student *i*, nested in classroom *j* nested in block *k* is male;

0 if female

$HomeLang1_{ijk}$ = 1 if home language of student *i*, nested in classroom *j* nested in block *k* is English only;
0 if HomeLang2=1 or if home language is a mix of English and Spanish, a mix of English and some other language, or if some other language is the primary language in the home

$HomeLang2_{ijk}$ = 1 if home language of student *i*, nested in classroom *j* nested in block *k* is Spanish only or a mix of English and Spanish;
0 if HomeLang1=1 or if home language is a mix of English and Spanish, a mix of English and some other language, or if some other language is the primary language in the home

Trt_{jk} = 1 if classroom *j* nested in block *k* was in Treatment Groups 1, 2, or 3;
= 0 if control group.

$MeanLapD\_CT_{jk}$ = Class-level mean LapD Cognitive Total Score of class *j* nested in block *k* , calculated from tests administered in the fall of 2003 and fall of 2004.

The parameter estimate $\hat{\gamma}_{010}$ from the model above is the estimated treatment effect. The value of $\hat{\gamma}_{010}$ is entered in Exhibit 8-1 in the column labeled "Mean Difference T-C". The values shown in the column labeled "Control Mean (SD)" were calculated as the simple mean and standard deviation of the TOPEL outcome measure values of the children in the control group. The value of the mean shown in the column labeled "Treatment Mean (SD)" was calculated as the sum of the treatment effect, $\hat{\gamma}_{010}$, and the control group mean. The treatment group standard deviation was calculated as the standard deviation of TOPEL outcome measure values of the children in the combined group of the three treatment groups. The effect size was calculated by dividing the treatment effect, $\hat{\gamma}_{010}$, by the Year 2005 control group standard deviation. The p-value corresponds to a two-sided test of the null hypothesis that the treatment effect is equal to zero.

To estimate the impacts of each of the three treatments, contrasted with control, the data were analyzed in same model as specified above, except that three dummy variables representing the contrasts of each of the three treatment groups to the control group were entered in the level-2 model instead of the single treatment dummy that was utilized in the model above.

Other than the modifications to the level-2 model, shown below, all other model terms were identical to those used in the previously model described.

Level-2 Model:

$$\pi_{0jk} = \beta_{00k} + \beta_{01k}(Trt1_{jk}) + \beta_{01k}(Trt2_{jk}) + \beta_{01k}(Trt3_{jk}) + \beta_{02k}(MeanLapD\_CT_{jk}) + r_{jk}$$

where,

$Trt1_{jk}$ = 1 if classroom *j* nested in block *k* was in Treatment Group 1; = 0 else.

$Trt2_{jk}$ = 1 if classroom *j* nested in block *k* was in Treatment Group 2; = 0 else.

$Trt3_{jk}$ = 1 if classroom *j* nested in block *k* was in Treatment Group 3; = 0 else.

Additional models were fit to the data where treatment-groups 1 and 3 combined were contrasted to the control group. Data from Treatment Group 2 data were omitted from these analyses. Other than the minor modification to the level-2 model, shown below, all other model terms were the same as previously described.

Level-2 Model:

$$\pi_{0jk} = \beta_{00k} + \beta_{01k}(Trt13_{jk}) + \beta_{02k}(MeanLapD\_CT_{jk}) + r_{jk}$$

where

$Trt13_{jk}$ = 1 if classroom *j* nested in block *k* was in Treatment Groups 1or 3;
= 0 if control group.

Impacts on subgroups were estimated by creating subsets of data, and fitting the models specified above to the subsets.

Non-experimental Analyses—Relationship of teacher education to teacher behavior and classroom environment

Since the experimental design did not manipulate the levels of teacher education, the analyses of relationships between teacher education and teacher behavior and classroom environment were non-experimental.

Relationships of teacher education to teacher behavior and classroom environment were estimated from:

- The full sample
- The sample of English-dominant teachers
- The sample of Spanish-dominant teachers

The data were analyzed in two-level HLM models, where teachers (Level-1) were nested in randomization blocks (Level-2). The two-level random intercept HLM models were of the form:

Level 1

$$Y_{ij} = \beta_{0j} + \beta_1(TeacherBA) + r_{ij}$$

Level 2

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

where Y_{ij} is a 2003 OMLIT measure on the i^{th} class nested in the j^{th} block, the β_{0j} are random intercept terms for the *j* blocks, and TeacherBA is coded as 1 if the teacher has a bachelor degree or higher and zero otherwise.

In Exhibit 8-10, the column labeled "Estimate of Effect" shows the parameter estimate $\hat{\beta}_1$, the column labeled "standard error of effect" gives the standard error of $\hat{\beta}_1$. The column labeled "effect size" shows the estimate, $\hat{\beta}_1$, divided by the Year 2004 control group standard deviation of the measure, i.e., 10. The p-value is a two-sided test of the null hypothesis that $\beta_1 = 0$.

Estimates of relationships for subgroups were estimated by creating subsets of data, and fitting the models specified above to the subsets.

Chapter 7: Impact of the Interventions on Teachers and Classrooms

A basic assumption underlying the design of the study, including the strategy for collecting and analyzing data, was that child outcomes are mediated by the actions and behavior of their teachers. Therefore, significant impacts on teacher behavior and the literacy environment would be necessary precursors of improved child outcomes. Below, we present, first, the initial and later findings about the impact of the interventions on teachers and classrooms, and on the pattern of activities in the classroom.

Impact of the Interventions on Teacher Behavior and the Literacy Environment

We examined the effect of the interventions on teacher behavior and interactions with children using four constructs, representing support for the four building blocks of emergent literacy: a) support for oral language, b) support for phonological awareness; c) support for print knowledge; and d) support for print motivation. Each of the constructs is built from a range of observational variables, drawn from the OMLIT battery of measures.

Support for oral language incorporates the amount of read-aloud activities as well as measures of their quality in terms of: the use of open-ended questions; information about text concepts; introduction of new vocabulary; linking story elements to children's own experiences; post-reading discussions; and the amount of teacher-child language interaction. ***Support for phonological awareness*** is a measure of the ways teachers draw children's attention to the sounds of words through singing and rhymes, and help them blend one-syllable words into different two-syllable words (blending) and, conversely, break apart two-syllable words into their single-syllable component words (elision). ***Support for print knowledge*** incorporates the amount of time spent in teaching letters and the correspondence between letters and sounds and in helping children with writing, and extent to which the teacher encourages children to integrate print into other activities including daily routines. ***Support for print motivation*** measures the strategies teachers use to motivate children to want to read.

In addition to these teacher-focused constructs, two additional constructs were used to assess the impact of the interventions on the classroom environment: ***literary resources*** measures the amount of environmental print and text materials present in the classroom, as well as the extent to which literacy resources are integrated into various activity centers; ***literacy activities*** is a measure of all the classroom activities that incorporate literacy.

As Exhibit 7-1 shows, after less than six months' implementation of the curricula, there were significant impacts on teachers' support for oral language, print knowledge and print motivation, and on the number of activities that incorporated literacy. Teachers in treatment group classrooms were providing more opportunities for oral language development and learning about print, and they were engaging in more of the activities that foster children's desire to read and use print. At this point, there were no significant effects on the classrooms' literacy resources (probably because all classrooms in the study received a comprehensive package of materials to support literacy activities at the beginning of the study), or on teachers' support for phonological awareness. Two of the three interventions delayed training on this element until spring 2004, to ensure that the other elements were in place.

Exhibit 7-1

Overall Impact of the Interventions on Teacher Behavior and the Classroom Environment (OMLIT, Spring 2004)

Construct	Control Mean (SD)	Treatment Mean (SD)	Mean Difference T-C	Effect Size	P-value
Support for Oral Language	50 (10)	55.86 (8.60)	5.86	.59	.000***
Support for Phonological Awareness	50 (10)	52.12 (9.11)	2.12	.21	.181
Support for Print Knowledge	50 (10)	55.30 (10.40)	5.30	.53	.002**
Support for Print Motivation	50 (10)	55.84 (9.10)	5.84	.58	.000***
Literacy Resources	50 (10)	50.85 (9.54)	.85	.08	.586
Literacy Activities	50 (10)	53.90 (9.76)	3.90	.39	.018*
Sample Size (centers/classrooms)	***n = 54***	***n = 106***			

The effect sizes are standardized measures of the magnitude (size) of treatment effects. The standardization makes possible the comparison of the sizes of treatment effects, between different outcome measures. For example, if the effect sizes of a treatment on outcome measures A and B are 0.50, and 0.25, respectively, then the size of the treatment impact on A is twice the size of the impact on B. For each outcome measure, the effect size is equal to the estimated treatment impact, divided by the control group standard deviation.

*** = p<.001, ** = p<.01, * = p<.05, NS = not significant.

Even at this early stage, there were some differences in the pattern of effects on teachers. Treatments 1 and 3 had more significant effects on teacher behavior (Exhibit 7-2) and the impacts on teacher behavior are almost entirely driven by the effect of the intervention on Spanish-dominant teachers (Exhibit 7-3).

By Spring 2005, there were significant positive impacts on all six constructs. Teachers in the treatment groups learned about and conducted many more activities to promote phonological awareness, such as singing, playing rhyming games, reading poems.

While by the end of the study, all three interventions had significant effects on aspects of teacher behavior and the classroom environment, Exhibit 7-5 suggests that the three curricula had different strengths and weaknesses. Treatments 1 and 3, which had larger impacts on some aspects of teacher behavior and on the number of literacy activities, showed no significant effects on the literacy resources in the classrooms. Treatment 1, which significantly increased support for print motivation, is the only one of the three that used authentic children's literature (trade books) rather than controlled-language books. Treatment 2, which had slightly weaker effects on most aspects of teacher behavior, had strong effects on teacher support for phonological awareness and on literacy resources. This intervention introduced the concepts of blending and elision at the initial training and continued to emphasize them. In addition, the curriculum stressed building thematic connections into the classrooms' activity centers, increasing the richness of the print environment.

Exhibit 7-2

Differential Impact of the Three Interventions on Teacher Behavior and the Classroom Environment (OMLIT, Spring 2004)

Treatment 1 (Ready, Set, Leap)

Construct	Control Mean (SD)	Treatment Mean (SD)	Mean Difference T-C	Effect Size	P-value
Support for Oral Language	50 (10)	57.15 (7.70)	7.15	.72	.000***
Support for Phonological Awareness	50 (10)	53.23 (8.48)	3.23	.32	.030*
Support for Print Knowledge	50 (10)	59.03 (8.10)	9.03	.90	.000***
Support for Print Motivation	50 (10)	57.16 (9.24)	7.16	.72	.000***
Literacy Resources	50 (10)	52.13 (9.38)	2.13	.21	.279
Literacy Activities	50 (10)	55.77 (9.37)	5.77	.58	.005**
Sample Size (centers/classrooms)	***n = 54***	***n = 36***			

Treatment 2 (B.E.L.L.)

Construct	Control Mean (SD)	Treatment Mean (SD)	Mean Difference T-C	Effect Size	P-value
Support for Oral Language	50 (10)	53.00 (10.30)	3.00	.30	.134
Support for Phonological Awareness	50 (10)	52.21 (8.68)	2.21	.22	.289
Support for Print Knowledge	50 (10)	48.66 (9.30)	-1.34	-0.13	.517
Support for Print Motivation	50 (10)	54.45 (8.25)	4.45	.45	.034*
Literacy Resources	50 (10)	51.17 (7.81)	1.17	.12	.567
Literacy Activities	50 (10)	50.13 (10.20)	.13	.01	.952
Sample Size (centers/classrooms)	***n = 54***	***n = 34***			

Treatment 3 (Breakthrough to Literacy)

Construct	Control Mean (SD)	Treatment Mean (SD)	Mean Difference T-C	Effect Size	P-value
Support for Oral Language	50 (10)	57.12 (7.07)	7.12	.71	.000***
Support for Phonological Awareness	50 (10)	50.86 (10.16)	0.86	.09	.721
Support for Print Knowledge	50 (10)	57.43 (10.50)	7.43	.74	.000***
Support for Print Motivation	50 (10)	55.74 (9.75)	5.74	.57	.005**
Literacy Resources	50 (10)	49.13 (11.13)	-0.87	-0.09	.666
Literacy Activities	50 (10)	55.29 (8.84)	5.29	.53	.010*
Sample Size (centers/classrooms)	***n = 54***	***n = 36***			

Exhibit 7-3

Overall Impact of the Interventions on Teacher Behavior and the Classroom Environment for Spanish-dominant Teachers vs. English-dominant Teachers (OMLIT, Spring 2004)

Spanish-dominant Teachers							
Construct	**Control Mean (SD)**		**Treatment Mean (SD)**		**Mean Difference T-C**	**Effect Size**	**P-value**
Support for Oral Language	48.46	(9.82)	55.47	(8.08)	7.01	.71	.001***
Support for Phonological Awareness	47.79	(7.33)	52.59	(8.51)	4.80	.48	.015*
Support for Print Knowledge	49.16	(9.87)	57.13	(9.98)	7.97	.80	.000***
Support for Print Motivation	47.87	(10.48)	56.03	(8.37)	8.16	.81	.001**
Literacy Resources	50.04	(8.73)	52.34	(8.56)	2.30	.23	.257
Literacy Activities	48.82	(10.12)	53.66	(10.12)	4.84	0.48	.042*
Sample Size (centers/classrooms)	***n = 26***		***n = 49***				

English-dominant Teachers							
Construct	**Control Mean (SD)**		**Treatment Mean (SD)**		**Mean Difference T-C**	**Effect Size**	**P-value**
Support for Oral Language	51.43	(10.13)	56.01	(9.09)	4.59	.46	.045*
Support for Phonological Awareness	52.05	(11.72)	51.73	(9.60)	-0.33	-0.03	.089
Support for Print Knowledge	50.78	(10.23)	53.92	(10.71)	3.13	.31	.219
Support for Print Motivation	51.98	(9.29)	55.75	(9.72)	3.78	.38	.101
Literacy Resources	49.97	(11.21)	49.65	(10.06)	-0.31	-0.03	.894
Literacy Activities	51.10	(9.94)	54.15	(9.53)	3.05	0.31	.182
Literacy Activities (centers/classrooms)	***n = 28***		***n = 58***				

Exhibit 7-4

Overall Impact of the Interventions on Teacher Behavior and the Classroom Environment (OMLIT, Spring 2005)

Construct	Control Mean (SD)		Treatment Mean (SD)		Mean Difference T-C	Effect Size	P-value
Support for Oral Language	49.81	(10.38)	55.92	(9.36)	6.11	.61	.000***
Support for Phonological Awareness	48.16	(7.73)	53.05	(9.90)	4.89	.49	.001**
Support for Print Knowledge	48.41	(11.18)	55.83	(9.99)	7.42	.74	.000***
Support for Print Motivation	50.90	(10.85)	55.19	(9.48)	4.29	.43	.012*
Literacy Resources	48.91	(9.04)	51.69	(8.40)	2.77	.28 .	.045*
Literacy Activities	47.38	(11.37)	55.38	(8.50)	8.00	.80	.000***
Sample Size (centers/classrooms)	***n = 53***		***n = 104***				

Exhibit 7-5

Differential Impact of the Three Interventions on Teacher Behavior and the Classroom Environment (OMLIT, Spring 2005)

Treatment 1 (Ready, Set, Leap)

Construct	Control Mean (SD)	Treatment Mean (SD)	Mean Difference T-C	Effect Size	P-value
Support for Oral Language	49.81 (10.38)	56.13 (8.95)	6.32	.63	.003**
Support for Phonological Awareness	48.16 (7.73)	52.99 (11.20)	4.83	.48	.013*
Support for Print Knowledge	48.41 (11.18)	57.88 (8.94)	9.46	.95	.000***
Support for Print Motivation	50.90 (10.85)	56.99 (10.95)	6.09	.61	.005**
Literacy Resources	48.91 (9.04)	51.62 (8.80)	2.71	.27	.116
Literacy Activities	47.38 (11.37)	57.12 (8.01)	9.75	.98	.000***
Sample Size (centers/classrooms)	***n = 53***	***n = 36***			

Treatment 2 (B.E.L.L.)

Construct	Control Mean (SD)	Treatment Mean (SD)	Mean Difference T-C	Effect Size	P-value
Support for Oral Language	49.81 (10.38)	54.14 (9.97)	4.33	.43	.047*
Support for Phonological Awareness	48.16 (7.73)	53.99 (9.68)	5.83	.58	.004**
Support for Print Knowledge	48.41 (11.18)	51.75 (7.32)	3.34	.33	.131
Support for Print Motivation	50.90 (10.85)	53.55 (9.84)	2.65	.27	.236
Literacy Resources	48.91 (9.04)	53.98 (7.97)	5.07	.51	.005**
Literacy Activities	47.38 (11.37)	52.40 (8.31)	5.02	.50	.000***
Sample Size (centers/classrooms)	***n = 53***	***n = 33***			

Treatment 3 (Breakthrough to Literacy)

Construct	Control Mean (SD)	Treatment Mean (SD)	Mean Difference T-C	Effect Size	P-value
Support for Oral Language	49.81 (10.38)	57.39 (9.08)	7.58	.76	.001**
Support for Phonological Awareness	48.16 (7.73)	52.26 (8.82)	4.11	.41	.036*
Support for Print Knowledge	48.41 (11.18)	57.52 (11.81)	9.11	.91	.000***
Support for Print Motivation	50.90 (10.85)	54.84 (7.24)	3.94	.39	.072
Literacy Resources	48.91 (9.04)	49.62 (8.26)	.71	.07	.685
Literacy Activities	47.38 (11.37)	56.24 (8.57)	8.86	.89	.000***
Sample Size (centers/classrooms)	***n = 53***	***n = 35***			

Exhibit 7-6

Impact of the Interventions on Teacher Behavior and the Classroom Environment by Language of Teacher (OMLIT, Spring 2005)

Spanish-dominant Teachers

Construct	Control Mean (SD)	Treatment Mean (SD)	Mean Difference T-C	Effect Size	P-value
Support for Oral Language	49.26 (10.08)	55.59 (9.77)	6.33	.63	.009**
Support for Phonological Awareness	48.07 (7.33)	52.40 (9.28)	4.33	.43	.041*
Support for Print Knowledge	46.99 (10.73)	55.98 (10.23)	8.99	.90	.001**
Support for Print Motivation	48.30 (9.99)	54.21 (9.84)	5.91	.59	.014*
Literacy Resources	49.34 (8.87)	52.79 (7.85)	3.45	.34	.075
Literacy Activities	45.75 (10.33)	53.75 (8.10)	8.00	.80	.000***
Sample Size	***n = 26***	***n = 49***			

English-dominant Teachers

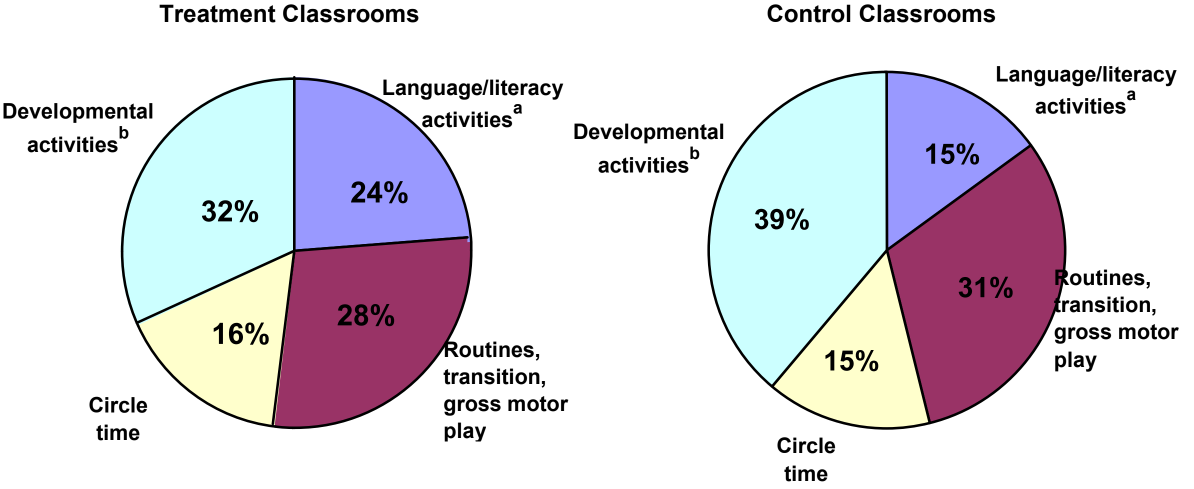
Construct	Control Mean (SD)	Treatment Mean (SD)	Mean Difference T-C	Effect Size	P-value
Support for Oral Language	50.34 (10.83)	55.85 (9.06)	5.51	.55	.023*
Support for Phonological Awareness	48.25 (8.24)	53.46 (10.48)	5.22	.52	.018*
Support for Print Knowledge	49.78 (11.63)	55.18 (9.84)	5.40	.54	.032*
Support for Print Motivation	53.41 (11.24)	55.64 (9.16)	2.23	.22	.358
Literacy Resources	48.50 (9.36)	50.84 (8.71)	2.34	.23	.253
Literacy Activities	48.94 (12.27)	56.61 (8.72)	7.67	.77	.002**
Sample Size	***n = 27***	***n = 55***			

Impact of the Interventions on Classroom Activities

One of two questions addressed in additional analyses was whether the increased focus on language and literacy activities might come at the expense of other important developmental activities. The interventions did indeed increase time spent on language and literacy activities substantially. However, Exhibit 7-7 shows that, while there are some resulting differences in the proportion of time allocated to different activities, these differences were not large. Children in the treatment group spent 9% more time in language and literacy activities than children in the control group (a 64% increase), 7% less time in other developmental activities[26] and 3% less time in routines, transitions and gross motor play.

Exhibit 7-7

Children's Activities (OMLIT – Spring 2005)



[a] Literacy/language activities include: reading (read-aloud, shared reading, child reading by himself), letters, letter-sound correspondence, writing (emergent tracing/copying, computer language programs).

[b] Developmental activities include: dramatic play, creative play, sensory play, blocks, fine motor play, games

Source: OMLIT Snapshot of Classroom Activity, one day, in-class observations

[26] Time on any single activity was reduced by 1% or less.

Chapter Eight: Impact of the Interventions on Child Outcomes

The effects of the interventions on children's language development and emergent literacy skills were assessed in Spring 2004 at the end of the four-year-old year, for children who had been in the classrooms between two and ten months. The average number of children enrolled in the four-year-old classrooms in 2004-2005 ranged from 16 to 24. The percentage of children assessed in Spring 2005 ranged from 50% to 55% of the enrollment. There was no sampling of children to be tested. The Early Learning Coalition obtained parental permission for assessment in Fall 2004. Almost no parents refused permission for assessment. However, because it was not always possible to obtain permission to test children more recently enrolled in time for the testing, in practice many recently enrolled children were not able to be tested. In addition, as the end of the testing period approached, in May 2005, parents began to make their summer arrangements and some 4-year-olds left the program before they could be assessed.

Children's language and literacy skills were assessed using three subtests from the *TOPEL*:

- **Definitional Vocabulary:** This is a test of vocabulary in which the child is asked to identify a pictured item (target word) and produce an entailment (i.e., answer questions such as: What is it for? What does it do? Where is it found?) in which associated verbs, adjectives, and nouns are elicited.
- **Phonological Awareness:** This test of phonemic sensitivity combines blending, specifically the ability to blend sounds (put sounds together – e.g., hay +stack is -- haystack) and elision, specifically the ability to remove sounds from words (e.g., what word is left when you take stack away from haystack?). The test moves from word-level, to syllable-level, to sub-syllable level and from receptive (multiple choice, identification) to productive (free response) skills.
- **Print Knowledge:** This subtest measures early print knowledge (print concepts, letter discrimination, word discrimination, letter-name identification and production, letter-sound identification and production).
- **Early Literacy Index:** Scores from the three subtests were combined to produce an index of early literacy.

Taken together, the three curricula interventions had significant effects on all four outcome measures (Exhibit 8-1). However, the findings are driven by the two interventions that showed impacts on children's language and literacy development (Exhibit 8-2). Treatment 1, Ready, Set, Leap and Treatment 3, Breakthrough to Literacy, had significant effects on all of the measures; Treatment 2, Building Early Language and Literacy had no significant impacts on any of the measures. When we combine the impacts for Treatments 1 and 3 (Exhibit 8-3), we can see that these two curricula taken together significantly improved outcomes for children. For the remainder of the discussion, we have combined the two curricula to improve statistical power and because the impacts of each were quite similar.

Exhibit 8-1

Overall Impact of the Interventions on Child Outcomes (TOPEL, Spring 2005)

Subtest	Control Mean (SD)		Treatment Mean (SD)		Mean Difference T-C	Effect Size	P-value
Definitional Vocabulary	78.79	(16.43)	82.43	(17.77)	3.64	.22	.017*
Phonological Awareness	88.74	(16.19)	93.28	(15.95)	4.54	.28	.003**
Print Knowledge	95.89	(15.31)	102.82	(14.66)	6.93	.45	.000***
Early Literacy Index	84.93	(16.32)	91.12	(16.70)	6.19	.38	.000***
Sample Size (children)	***n = 509***		***n = 1014***				

*** = p<.001, ** = p<.01, * = p<.05, NS = not significant.

Exhibit 8-2

Differential Impact of the Three Interventions on Child Outcomes (TOPEL, Spring 2005)

Treatment 1 (Ready, Set, Leap)

Subtest	Control Mean (SD)		Treatment Mean (SD)		Mean Difference T-C	Effect Size	P-value
Definitional Vocabulary	78.79	(16.43)	83.40	(18.11)	4.61	.28	.017*
Phonological Awareness	88.74	(16.19)	94.36	(16.13)	5.62	.35	.003**
Print Knowledge	95.89	(15.31)	105.83	(13.03)	9.94	.65	.000***
Early Literacy Index	84.93	(16.32)	93.20	(15.77)	8.27	.51	.000***
Sample Size (children)	***n = 509***		***n = 320***				

Treatment 2 (B.E.L.L.)

Subtest	Control Mean (SD)		Treatment Mean (SD)		Mean Difference T-C	Effect Size	P-value
Definitional Vocabulary	78.79	(16.43)	79.88	(17.87)	1.09	.07	.577
Phonological Awareness	88.74	(16.19)	89.31	(15.53)	0.57	.04	.767
Print Knowledge	95.89	(15.31)	97.01	(15.45)	1.11	.07	.565
Early Literacy Index	84.93	(16.32)	85.93	(16.93)	0.99	.06	.637
Sample Size (children)	***n = 509***		***n = 340***				

Treatment 3 (Breakthrough to Literacy)

Subtest	Control Mean (SD)		Treatment Mean (SD)		Mean Difference T-C	Effect Size	P-value
Definitional Vocabulary	78.79	(16.43)	83.83	(16.90)	5.04	.31	.009**
Phonological Awareness	88.74	(16.19)	95.82	(15.63)	7.08	.44	.000***
Print Knowledge	95.89	(15.31)	105.13	(13.63)	9.24	.60	.000***
Early Literacy Index	84.93	(16.32)	93.81	(16.12)	8.88	.54	.000***
Sample Size (children)	***n = 509***		***n = 354***				

*** = p<.001, ** = p<.01, * = p<.05, NS = not significant.

Exhibit 8-3

Impact of Treatments 1 and 3 Combined on Child Outcomes (TOPEL, Spring 2005)

Subtest	Control Mean (SD)		Treatment Mean (SD)		Mean Difference T-C	Effect Size	P-value
Definitional Vocabulary	78.79	(16.43)	83.72	(17.49)	4.93	.30	.001***
Phonological Awareness	88.74	(16.19)	95.09	(15.86)	6.35	.39	.000***
Print Knowledge	95.89	(15.31)	105.51	(13.38)	9.62	.63	.000***
Early Literacy Index	84.93	(16.32)	93.59	(15.94)	8.66	.53	.000***
Sample Size (children)	***n = 509***		***n = 674***				

*** = p<.001, ** = p<.01, * = p<.05, NS = not significant.

Exhibit 8-4 shows that the impacts were different for children in classrooms with teachers whose primary language was Spanish (where the children also spoke Spanish as their home language) vs. children in classrooms with teachers whose primary language was English. The exhibit shows that, for children in classrooms with Spanish-dominant teachers, there were impacts on more of the measures and that the impacts were greater than for children with English-dominant teachers. This finding reflects the earlier finding that the curricula had a larger impact on the behavior of Spanish-speaking teachers. The results are quite similar when we look at the difference in outcomes specifically for children with a home language other than English and those whose home language was English (Exhibit 8-5).[27] Some of the English-language learners (and all of the Haitian-Creole speakers) were in classrooms with English-speaking teachers, on whom the effects of the interventions were less pronounced.[28]

It is important to remember that these outcomes are for tests administered in English. An important goal for the curricula was to help English-language learners progress in English before they entered English-only kindergarten classes, and the two interventions appear to have been quite effective in doing that.

[27] Note that this is a non-experimental comparison, since children's language was not taken into account in the random assignment process.

[28] Spanish-dominant teachers were always in classrooms with children whose home language was Spanish. Almost all the interactions in these classrooms were in Spanish. However, some children whose home language was Spanish or Haitian-Creole were in classrooms with English-dominant teachers (as well as children whose home language was English). In these mixed classrooms, there was usually an aide who spoke Spanish (or Haitian-Creole) but the dominant classroom language was English. The impacts of the interventions on Spanish-speaking children show the same pattern, regardless of the classroom language, but the effects were larger for Spanish-speaking children in classrooms with Spanish-dominant teachers.

Exhibit 8-4

Impact of Treatments 1 and 3 Combined on Child Outcomes by Language of Teacher (TOPEL, Spring 2005)

Spanish-dominant Teachers

Subtest	Control Mean (SD)	Treatment Mean (SD)	Mean Difference T-C	Effect Size	P-value
Definitional Vocabulary	73.52 (17.13)	79.88 (18.64)	6.36	.39	.007***
Phonological Awareness	84.64 (16.35)	93.54 (16.09)	8.90	.55	.000***
Print Knowledge	92.69 (15.23)	105.79 (13.68)	13.10	.86	.000***
Early Literacy Index	79.46 (16.81)	91.20 (16.64)	11.75	.72	.000***
Sample Size (children)	***n = 281***	***n = 332***			

English-dominant Teachers

Subtest	Control Mean (SD)	Treatment Mean (SD)	Mean Difference T-C	Effect Size	P-value
Definitional Vocabulary	83.83 (14.01)	87.52 (15.14)	3.69	.22	.069
Phonological Awareness	93.47 (14.67)	97.16 (15.23)	3.69	.23	.086
Print Knowledge	99.84 (14.49)	106.13 (12.99)	6.30	.41	.001**
Early Literacy Index	90.16 (13.99)	95.99 (14.54)	5.84	.36	.010*
Sample Size (children)	***n = 228***	***n = 342***			

*** = p<.001, ** = p<.01, * = p<.05, NS = not significant.

Exhibit 8-5

Impact of Treatments 1 and 3 Combined on Child Outcomes for Children with Spanish or Haitian Creole as Their Home Language (TOPEL, Spring 2005)

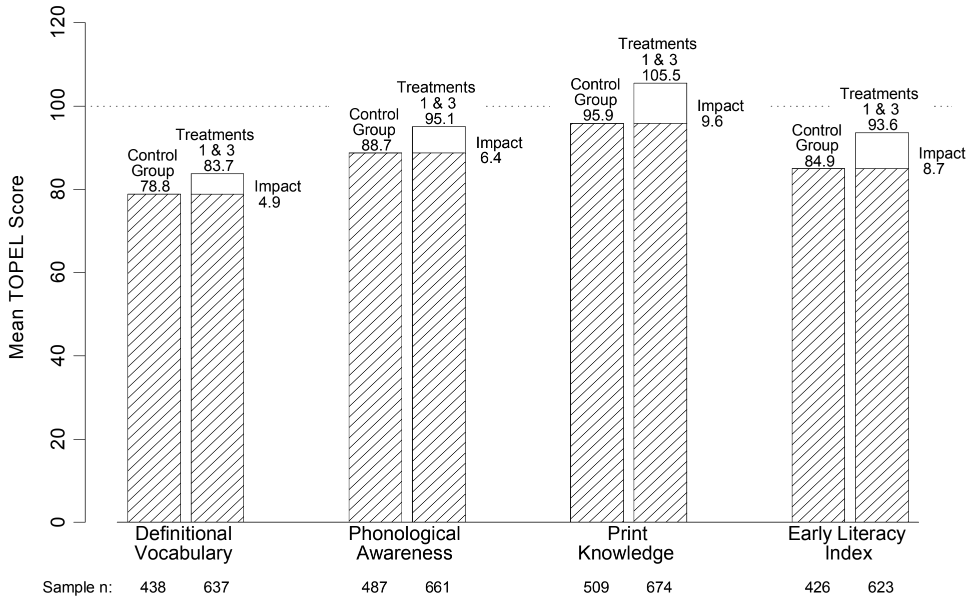
Subtest	Control Mean (SD)	Treatment Mean (SD)	Mean Difference T-C	Effect Size	P-value
Definitional Vocabulary	76.37 (16.69)	81.49 (17.90)	5.12	.31	.004**
Phonological Awareness	87.94 (16.62)	94.56 (16.20)	6.62	.41	.000***
Print Knowledge	95.09 (15.43)	105.57 (13.32)	10.48	.68	.000***
Early Literacy Index	83.33 (16.82)	92.62 (16.26)	9.29	.57	.000***
Sample Size (children)	***n = 404***	***n = 525***			

*** = p<.001, ** = p<.01, * = p<.05, NS = not significant.

Another way to look at the impact of the curricula on children's outcomes is to see where they are in terms of national norms. As part of ongoing work for the Office of the Assistant Secretary for Planning and Evaluation (ASPE), we have calculated that children from low-income families are about a year behind the national norms on a test of language at the end of the four-year-old year, as they prepare to enter kindergarten (Layzer, in preparation). While the interventions had significant impacts, it seems important to ask the question, “How much of the gap was closed?” On all four measures, children in the control group scored considerably below the norms. On the overall index, the interventions succeeded in closing more than half the gap in achievement. On the individual subscales, the interventions succeeded in halving the gap for Phonological Awareness and outperforming the norming sample (a nationally-representative sample of children). On the Definitional Vocabulary subtest, although the children in the two treatment groups made significant gains, there remained a large gap in achievement (Exhibit 9-6). As part of the analysis, we investigated a possible age-by-treatment interaction but found none.

These gains made by children in the two treatment groups can be described in another way. The discussion above shows that, on all three subtests the gap was reduced or eliminated. How many months of growth do these impacts represent? Exhibits 8-6, 8-7, 8-8, and 8-9 show that the impacts range from a low of almost five months for Definitional Vocabulary to nine months for Print Knowledge.[29]

Exhibit 8-6
Impact of Treatments 1 and 3 Combined on Child Outcomes (TOPEL, Spring 2005)



Read: The control group mean TOPEL score for definitional vocabulary was 78.8. The model-estimated mean for treatment groups 1 and 3 was 83.7. The treatment impact on definitional vocabulary was 4.9 scale score points.
Note: TOPEL scores are standardized such that the population mean and standard deviation are 100 and 15, respectively.

[29] Exhibit 8-7 addresses the question “How does a difference of 10 standardized score points equate to the changes in Print Knowledge associated with normal growth?” The exhibit was created as follows: The model-estimated standardized score means for Treatment and control groups were 106 and 96 respectively. (The model estimates a common treatment effect across children of different ages, since no age-by-treatment effect was found in the earlier analyses). In the exhibit, the plotting symbols shown as boxes, and connected with blue lines show the raw scores corresponding to a standardized score of 106 for children of different ages. The open circles connected by orange lines indicate the raw scores corresponding to a standardized score of 96. We began by finding the raw score for a 37-month-old child that corresponds to a standardized score of 96 (9.5). We then found the age at which a raw score of 9.5 corresponds to a standardized score of 96 in the control group (44.8 months). This suggests that the impact is roughly equivalent to almost 8 months of growth. The other exhibits were constructed in the same way.

Exhibit 8-7 Topel Print Knowledge: Impact of Treatments 1 and 3 Relative to Growth

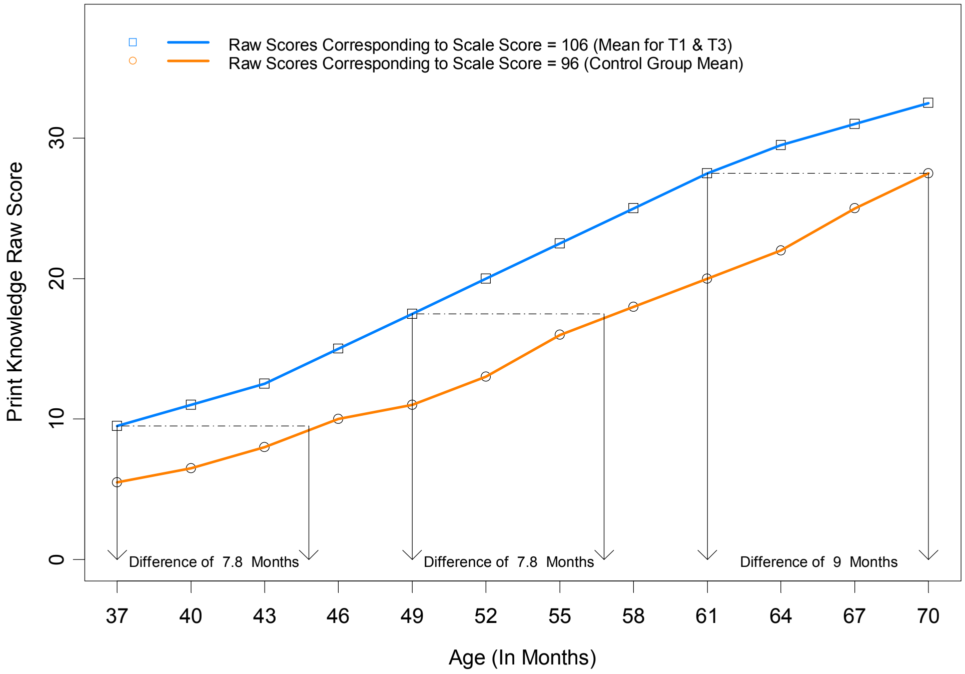


Exhibit 8-8. Topel Definitional Vocabulary: Impact of Treatments 1 and 3 Relative to Growth

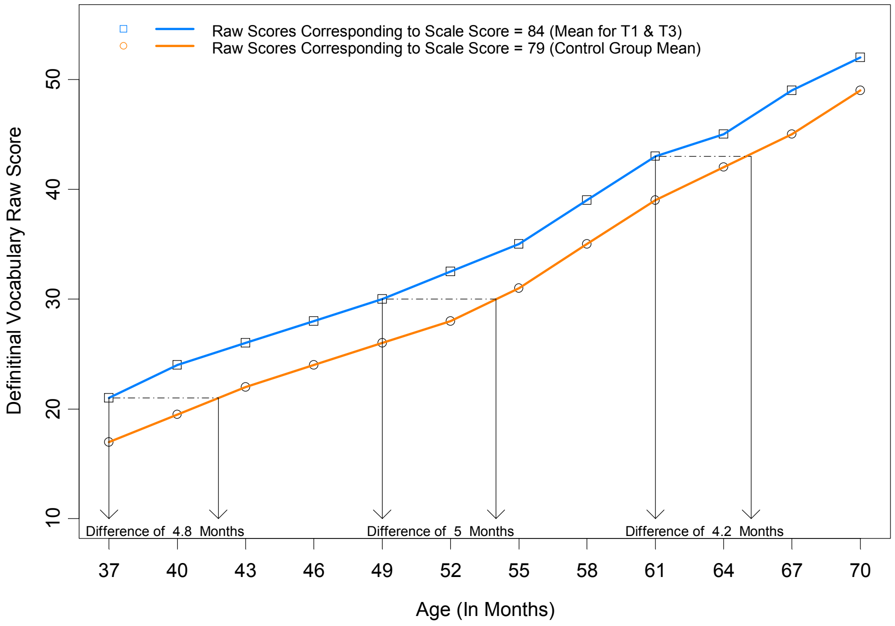
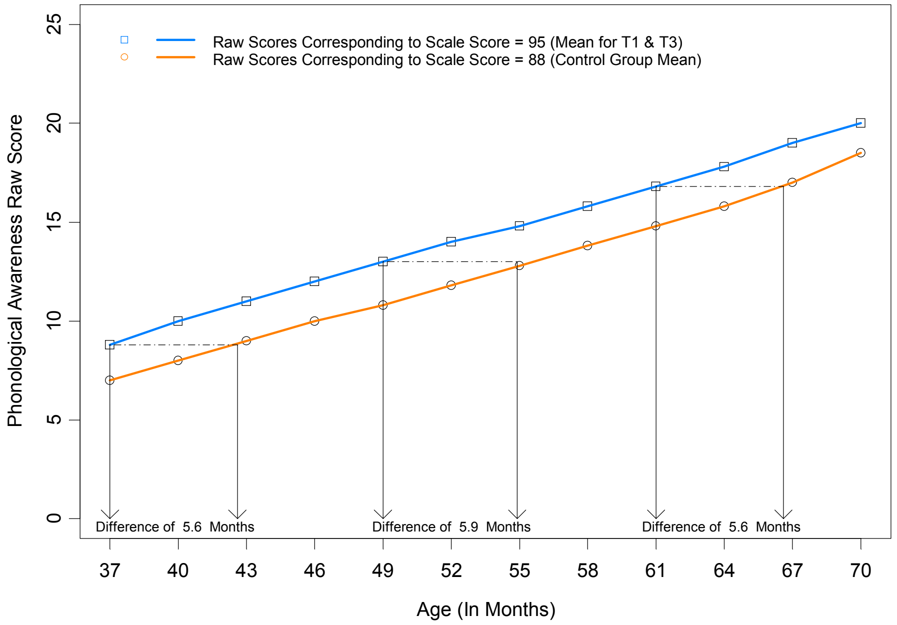


Exhibit 8-9. Phonological Awareness:
Impact of Treatments 1 and 3 Relative to Growth



Relationship Between Staff Educational Background and Teacher and Child Outcomes

Because of the national discussion about the importance of teacher educational credentials in early childhood education, which is increasingly reflected in states' systems for improving quality, we were interested in investigating two related questions:

- What is the relationship between teacher educational background and teacher behavior and interactions in the classroom?

- Does the educational level of teachers make a difference to the impact of the interventions on teacher behavior and interactions, the classroom environment and child outcomes?

To answer the first question, we used information on teacher education from the staff background questionnaire and observational data from the baseline data collection in 2003. The analysis investigated the relationship between having a bachelor's degree and teacher behavior and interactions with children.

We found small but significant relationships between a bachelor's degree and teachers' support for print knowledge and for teacher's positive affect toward children. The size of the effect is comparable to the effect size found by Barnett in his recent meta-analysis (Barnett and Ackerman, 2006). However, analysis of the relationships for teachers whose primary language was English vs. Spanish found significant relationships only for Spanish-dominant teachers (Exhibit 8-10).

Exhibit 8-10

Relationship of Teacher Education to Teacher Behavior and the Classroom Environment Overall and by Language of Teacher (OMLIT, Arnett, Fall 2003)

Construct	Estimate of Effect	Standard Error of Effect	Effect Size	P-value
Full Sample				
Support for Oral Language	0.01	1.67	.00	.997
Support for Print Knowledge	1.05	0.48	.10	**.031***
Support for Print Motivation	-1.25	1.32	.12	.345
Literacy Resources	1.04	0.83	.10	.213
Arnett: Positive Affect	2.99	1.51	.30	**.049***
Arnett: Not Punitive	-0.03	1.18	-.00	.981
Arnett: Engaged	3.44	2.20	.30	.121
English-Dominant Teachers				
Support for Oral Language	0.23	2.72	.02	.932
Support for Print Knowledge	1.02	0.77	.01	.193
Support for Print Motivation	-2.17	2.26	.02	.342
Literacy Resources	-0.19	1.40	.02	.893
Arnett: Positive Affect	4.27	2.32	.04	.070
Arnett: Not Punitive	-1.38	1.77	.01	.438
Arnett: Engaged	2.01	3.43	.02	.560
Spanish-Dominant Teachers				
Support for Oral Language	0.46	2.32	.00	.841
Support for Print Knowledge	1.61	0.64	.02	**.015***
Support for Print Motivation	0.15	1.74	.02	.937
Literacy Resources	2.68	1.07	.03	**.015***
Arnett: Positive Affect	1.45	2.22	.01	.515
Arnett: Not Punitive	1.78	1.82	.02	.331
Arnett: Engaged	4.34	3.36	.04	.202

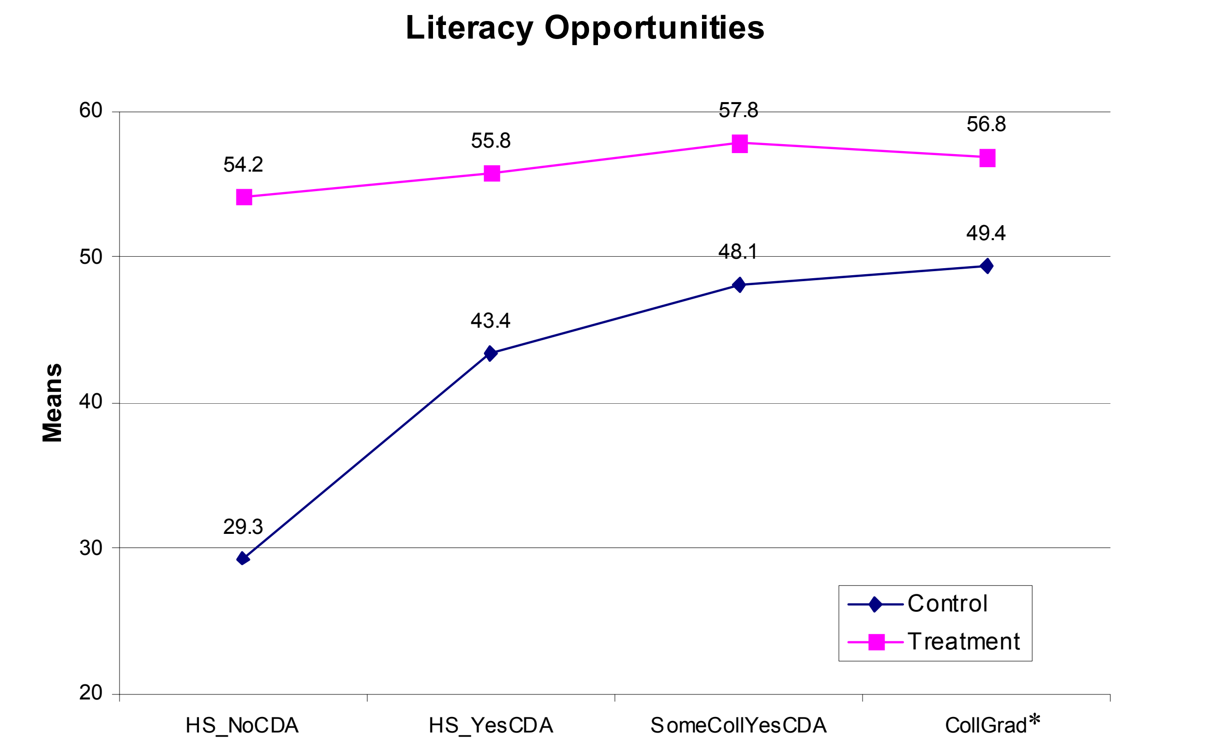
Note: Overall sample of 157 includes 82 English-dominant teachers and 75 Spanish-dominant teachers.

*** = p<.001, ** = p<.01, * = p<.05, NS = not significant.

Underlying the second question is the idea that better-educated teachers would be doing better in the classroom to begin with, might be better prepared to grasp and implement a new curriculum, would therefore demonstrate more of the behaviors and interactions that support language and literacy development and would produce greater impacts on children’s outcomes. To examine whether this was indeed the case, we looked first at the 2005 observational data from the OMLIT, to determine whether teachers’ educational achievement affected the impact of the treatment on teacher behavior. An interaction effect was found for one construct on the OMLIT – Literacy Opportunities (the number and type of activities and opportunities, either teacher-or child-initiated, that supported literacy), but it was not the hypothesized effect. Rather than heightening the effect of the interventions on better-educated teachers, the effect of the interaction was to eliminate the differences between less-educated teachers and their better-educated counterparts (Exhibit 8-11). In the treatment group, teachers at all educational levels look remarkably similar in the extent to which they provide or facilitate such opportunities, compared with quite dramatic differences in the control group teachers. As Exhibit 8-12 shows, the interaction effect was found in the sample of teachers for whom English was the dominant language. There were no significant interaction effects for Spanish-dominant teachers.

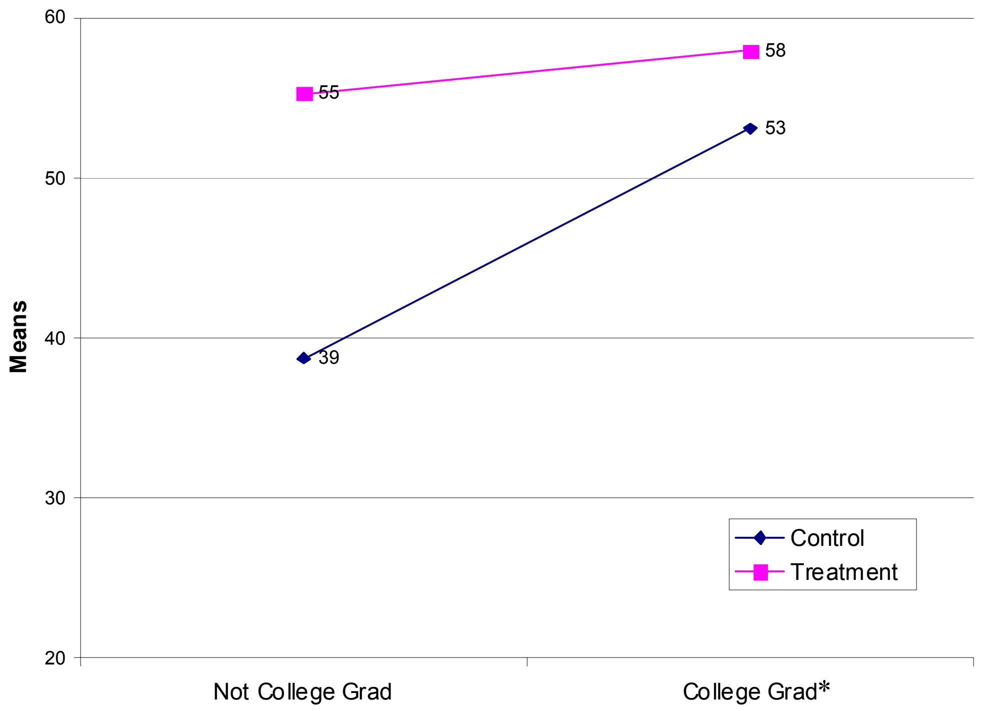
There were no interaction effects on child outcomes, however: the impacts of the treatment were similar for children, regardless of the educational background of the teacher.

Exhibit 8-11. Literacy Opportunities
Treatment-Teacher Education Interaction Effect (Full Sample): P=0.0408



* BA/BS Degree

Exhibit 8-12. Literacy Opportunities
Treatment-Teacher Education Interaction Effect (Teacher Prefers English): P=0.040



* AA/BA/BS Degree: Categorieswere combined for this analysis because of small sample sizes in some groups

Chapter Nine: Cost Effectiveness Analysis

Both Ready, Set, Leap (RSL) and Breakthrough to Literacy (BTL) had significant effects on all four measures of emergent literacy outcomes for children. Effect sizes (measured in standard deviations of the outcome measure for the control group) ranged from a low of 0.28 (definitional vocabulary, RSL) to a high of 0.65 (print knowledge, RSL). While the magnitude of the effect sizes for each outcome was comparable for RSL and BTL, the effect size for BTL was slightly higher than for RSL on three of the four child outcome measures examined.

However, the costs of implementing the two curricula over the two-year period of the evaluation were quite different. The average costs per classroom for each curriculum are shown in Exhibit 9-1. Cost information was extracted from ELC expenditure records.

Exhibit 9-1

Average Marginal Costs per Classroom Over the Two-Year Experiment: RSL and BTL[1]

Measure	RSL	BTL
Mentors	$3,069	$3,069
Curriculum materials and software, teacher training, take-home materials	$7,953	$13,500
Computers and linked printers with installation		$1,485
Computer maintenance		$120
Total marginal cost per classroom	$11,022	$18,174

[1] These are the incremental costs per classroom compared with the control group, over and above the cost of child care.

Both BTL and RSL used mentors that visited each classroom twice/month for 17 months, at a cost/classroom of $3,069 over the two-year period. The cost of the BTL curriculum materials, software, teacher training and take-home materials was $13,500/classroom, compared with $7,953/classroom for the RSL curriculum. In addition, each BTL classroom required two computers and a linked printer at a cost of $1.485, plus a $120 maintenance contract for this equipment. The total marginal cost/classroom was $18,174 compared with $11,022 for RSL, a difference of 65 percent.

Cost effectiveness analysis is a technique for comparing the relative value of alternative strategies for achieving the same effects. Exhibit 9-2 presents the RSL and BTL cost effectiveness ratios (CERs) for each of the child outcome measures used in the evaluation. These ratios are computed as the average marginal cost per classroom of a curriculum divided by the effect sizes. The CERs should be interpreted as the additional cost per classroom to obtain a one standard deviation effect size on an outcome measure. As one would expect given the very small differences in effect sizes and very large difference in costs between RSL and BTL, RSL is considerably more cost effective than BTL on each of the child outcome measures examined. Taking a simple average of the CERs across the four outcome measures, the average CER for RSL is $24,493 compared with $39,764.

The costs of early childhood education programs are best compared in terms of the annual cost per full time equivalent child or cost per child hour of care. Using an average of 20 FTE children per classroom (this is a common class size for four-year-olds nationally, in child care, prek and Head Start programs, even in Florida where group size in child care is not regulated), RSL added an average of $551 to the annual cost of an FTE, while BTL added an average of $909 per FTE and produced essentially the same effects on emergent literacy outcomes for children.

Exhibit 9-2

Cost Effectiveness of RSL and BTL

	RSL Marginal Cost = $11,022/Classroom		**BTL Marginal Cost = $18,174/Classroom**	
Measure	**Effect size**	**CE Ratio ($/S.D.)**	**Effect size**	**CE Ratio ($/S.D.)**
Definitional vocabulary	.28*	$39,364	.31**	$58,626
Phonological awareness	.35**	$31,491	.44***	$41,304
Print knowledge	.65 ***	$16,957	.60 ***	$30,290
Early literacy index	.51 ***	$21,612	.54***	$33,655
Average	.45	$24,493	.47	$38,668

*** = p<.001, ** = p<.01, * = p<.05, NS = not significant.

Replicating the Interventions

In order to achieve effects similar to those reported here, it would be necessary to replicate the intervention models described earlier in the report. While both of the successful curricula now include in their pricing several support visits in addition to initial and follow-up training, the level of support added is probably not sufficient to achieve the substantial impacts on teacher behavior and the classroom environment that, in turn, led to the reported impacts on outcomes for children.

Above, we described the per child cost of achieving the impacts on children ***in the experiment*** in Miami-Dade child care centers. These costs go down dramatically over time in the same child care centers, since the initial costs of curricula, computer equipment, mentors and training are eliminated. Costs to train and mentor replacement teachers remain, as do the costs of replacing lost or damaged curriculum materials, including take-home materials and, eventually, replacing outmoded computer equipment.

To replicate the intervention model elsewhere requires initial investment in curricula and first year training (initial and follow-up), the cost of which will depend on the number of classrooms and teachers involved.[30] In addition, for BTL, the initial investment includes the cost of computer equipment and installation (two computers and a linked printer per classroom). Added to these costs are the costs for a full-time mentor for every 18 classrooms for a period of 18 months on average. In the experiment described in this report, teachers varied in the speed with which they were able to implement the curricula fully. In a replication, mentors would move on once a teacher had achieved full implementation. For some teachers this might be achieved in a year or less; others might need two years, or a center might experience teacher turnover, so that mentors would start from scratch in that classroom. In the experiment, mentors all had an undergraduate degree and experience in early childhood education.

[30] Curriculum developers typically negotiate a price for their curricula based on the number of classrooms, children in the classroom, teachers to be trained, etc.

To oversee and monitor implementation would require an additional staff person, who could also be trained as a trainer, to provide training for teachers in the second year, to train replacement teachers, and to train new teachers as the replication is extended to more centers. In the experiment, the on-site coach/trainer managed two mentors and monitored 36 classrooms; in a larger replication, this staff person could probably manage more mentors and oversee more classrooms. The coach/trainers in the experiment had Master's degrees in early childhood education.

In a replication, the group of classrooms in which the curricula is implemented would have some ongoing costs. Most significant are the costs for training and mentoring replacement teachers; in some places these could be substantial. In the Miami-Dade centers, however, recent contacts showed that, two years after the experiment ended, the same teachers were continuing to implement the curricula in 75%-80% of the study classrooms. All classrooms will need replacement of broken, damaged or lost materials annually, as well as take-home materials for a new group of four-year-olds.

Discussion

The findings show that this model of professional development, in which initial and follow-up training sessions were supported by bi-monthly mentoring over an 18-month period, was effective in changing teachers' classroom practices and the classroom environments in ways that fostered early language and literacy development. This finding does not imply that all types of mentoring are equally effective. For all three of the interventions, mentoring activities were directly linked to research on early literacy and to teachers' actual classroom activities.

Importantly, this focused training and ongoing support eliminated the effects of teachers' educational background on their support for children's literacy. However, impacts on children did not differ by teachers' educational levels.

In most classrooms, the elements of each curriculum were securely in place at the end of the 18-month period. However, even after 18 months, many teachers were still not comfortable working with small groups for most of the time, as the mentors encouraged them to do. Much of the reading aloud that teachers did was with somewhat larger groups than was optimal. Mentors reported that teachers worried that some children would "miss out' on reading time if they worked mostly with small groups.

The impacts on children are also encouraging, given the size of the achievement gap for low-income children that is revealed as they prepare to enter school. On all but one of the measures, children in the treatment group moved close to the national norm or went beyond it. It is troubling that the gaps in children's vocabulary did not come close to being closed. A major reason for this is that Spanish-speaking children began with English-language scores well below the norms and below their English-speaking peers. Even though they made substantial progress as a result of the interventions, a large gap remained. It seems that the gap in this area may be too great to be closed in one year.

Nevertheless, the impacts on children's outcomes are substantially larger than we are used to seeing in large-scale, "real-life" studies. There are no comparable randomized experiments in child care centers against which to compare this study; the Head Start Impact Study may provide the closest comparison. On similar measures for 4-year-olds, the Impact Study found no impact on oral comprehension and phonological awareness, and a relatively modest effect (.22 of a standard deviation) on a letter-word identification test. This effect size is identical to the overall average effect of any organized preschool experience (center-based child care, Head Start, private pre-k and public prekindergarten) reported by Magnuson and her colleagues from their analysis of ECLS-K data (Magnuson et al., 2006). On the other hand, the impacts of the Project Upgrade interventions are similar to those reported for school-based prekindergarten programs. Using a regression-discontinuity design and data from five states, Barnett et al. (2005) found an impact of preschool on print awareness of .64 of a standard deviation. The effect of the Project Upgrade interventions seems to have been to focus the attention of child care staff on aspects of children's development that early childhood teachers in school-based programs recognize as critical elements of school readiness.

Finally, there is the finding that one of the interventions, though it had positive effects on teachers and classrooms, had no impact on children's outcomes. There are some possible explanations for this: this intervention featured two 15-20-minute add-on sessions each day in contrast to the other two which were intended to be woven into activities throughout the day. It seems likely that, B.E.L.L. teachers, though they engaged in the behaviors and interactions that promote literacy, spent less time on them than teachers in the other two groups, and that the exposure was not sufficient to affect children's outcomes.

In addition, the two successful interventions both used computer-based technology or electronic aids to act as a "second teacher" in the classroom; children could work by themselves in activity centers and ***receive feedback on what they were doing***. In classrooms with Spanish-dominant teachers, these electronic aids were key elements in children's learning English vocabulary.

In both cases, the result was greater exposure to the treatment. Since teachers liked all of the interventions, and benefited from all of them, it might be possible for the B.E.L.L. developer to modify the curriculum strategy in ways that would increase the intensity of exposure, by using electronic aids, dramatic play or fine motor materials to underscore the lessons learned in the 15-20 minute literacy activity periods.

None of the interventions was inexpensive; in particular, the two successful curricula had significant costs for electronic devices and supports. However, even for these two, the added costs per child hour amount to between 28 cents and 45 cents, an increase of 7-12%. For this, the interventions achieved large short-term impacts and reduced the gap in school readiness significantly. If we are able to follow the children into school and examine their later academic progress, that information will help determine the longer-term value of that investment.

While these findings provide the guidance that the Early Learning Coalition hoped for, the question of the longer-term meaning of these effects needs to be addressed. Did the interventions reduce the gaps in achievement sufficiently that children are better able to take advantage of the school experience? For teachers, are the effects on their behavior sustained in the absence of continued support from mentors? Do they continue to build on what they have learned? Does teacher turnover mean that later four-year-old cohorts have less exposure to the successful curricula? These questions haunt all early childhood interventions; they are especially important for interventions that have such powerful short-term effects. We are hopeful that we will be able to examine the longer-term impact of the interventions on children who remain in the Miami-Dade public schools and address at least the first of these questions.

References

Barnett, W. Steven, Cynthia Lamy, and Kwang-Hee Jung. (2005). The effects of state prekindergarten programs on young children's school readiness in five states. New Brunswick, NJ: National Institute for Early Education Research.

Bryk, A. S. & Raudenbush, S. W. (1992). Hierarchical linear models. Newbury Park, CA: Sage.

Dickinson, D.K., & Tabors, P.O. (2001). Beginning literacy with language: Young children learning at home and school. Baltimore, MD: Paul H. Brookes.

Lonigan, C.J., Burgess, S.R., & Anthony, J.L. (2000). Development of emergent literacy and early reading skills in preschool children: Evidence from a latent variable longitudinal study. Developmental Psychology, 30(5), 596-613.

Magnuson, Katherine, Christopher Ruhm, and Jane Waldfogel. (2006). Does prekindergarten improve school preparation and performance? Economics of Education Review (forthcoming).

National Research Council. (1999). Starting out right: A guide to promoting children's reading success. Washington, D.C.: National Academy Press.

Neuman, S.B., Copple, C., & Bredekamp, S. (2000). Learning to read and write: Developmentally appropriate practices for young children. Washington, D.C. National Association for the Education of Young Children (NAEYC).

Neuman, S.B., & Roskos, K. (1998). Children achieving: Best practices in early literacy. Newark, DE: International Reading Association.

Whitehurst, G.J., & Lonigan, C.J. (1998). Child development and emergent literacy. Child development, 69(3), 848-872.

Whitehurst, G.J., & Lonigan, C.J. (2001). Emergent literacy: Development from prereaders to readers. In Neuman & Dickinson (Eds.), Handbook of Early Literacy Research (pp. 11-29). New York: Guilford Press.

Appendix A

Reliability of the OMLIT

Two kinds of reliability have been established for the OMLIT measures, based on data from two national observation studies:

- ***Inter-rater reliability***: the degree of agreement between two trained observers administering the observation measures at the same time in the same classroom.

- ***Agreement with a criterion***: the extent of agreement between coding by trained observers and "master" or "correct" coding by experts of a standardized stimulus (e.g., a videotape of a classroom, written examples, etc.).

The discussion below presents data on these two types of reliability. Future waves of observations will provide additional data to increase the accuracy of our estimates of the reliability of the measures. The third type of reliability will depend on different data collection designs planned to occur in the near future.

Agreement with Criterion Coding

Paper and Pencil Tests

Reliability was assessed via paper and pencil tests on two of the OMLIT measures—the *Snapshot* and the *QUILL*. Written scenarios describing classroom events were prepared and coded in advance by the OMLIT developers (the "criterion" coding). The accuracy of observers' coding of written scenarios was determined by comparing it to the criterion coding of the same scenarios. Although this type of paper-and-pencil test does not simulate the "live" action in a classroom, it does provide a measure of how well observers understand the coding definitions for the various activities and specialized literacy data.

On the *Snapshot*, observers coded 15 written scenarios, and their coding was compared to criterion coding of the same written scenarios done in advance by three of the OMLIT developers. A high level of agreement was achieved between the coding done by the observers and the criterion coding (Exhibit A-1). On average, the coding of the written scenarios by the observers agreed almost perfectly with the criterion coding by the trainers. Further, each of the individual observers scored 95% or higher on the agreement between their coding and the criterion coding.

On the *QUILL*, the agreement ranged from 69% to 84% when agreement was defined as an exact match in ratings (Exhibit A-1). The agreement was substantially higher when the definition of agreement was expanded to agreement within a point.

Exhibit A-1
Average Agreement on Coding Written Scenarios on the *Snapshot* and *QUILL*
(% Agreement between 14 Observers and Criterion Coding)

Codes/Variables	**Average% Agreement with Criterion Coding**	
Snapshot	%	
Environment (all codes)	**98%**	
• Total # children present	93	
• Total # adults present	98	
• Type of adults present: teachers and aides	99	
• Type of adults present: other adults	99	
Activities (all codes)	**98%**	
• Type of activity	98	
• Number of children in activity	99	
• Number of teachers in activity	99	
• Number of aides in activity	98	
• Number of other adults in activity	99	
• Integration of literacy in other activities	96	
• Any language by children or adults	96	
All categories on *Snapshot*	**98%**	
QUILL	**Exact Agreement %**	**+/- 1 Pt on Scale %**
Overall average quality		
• Writing	.79	.83
• Letter/word knowledge	.70	.76
• Oral language	.69	.73
• Functions/features of print	.71	.76
• Print motivation	.82	.85
• Sounds	.84	.88

Coding Videotapes

Observers coded two videotape recordings of teachers reading aloud to a group of children using the *RAP*. The agreement between the observers' coding and the criterion coding by the developers was assessed in four areas:

- Instructional behavior in the pre-reading (set-up) period.
- Instructional behavior while reading the book.
- Post-reading instruction.
- Quality ratings on (a) introduction of new vocabulary, (b) depth of story-related discussion, including use of open-ended questions that invite children to engage in prediction, imagination, and/or rich description, and (c) the depth of any post-reading book-related activities that the adult organizes (beyond oral discussion).

Agreement between the observers and the criterion coding was computed as the average agreement across the two videotapes. The average percent agreement was very high on coding the instructional strategies used by the teacher during the read-aloud (Exhibit A-2). Average percent agreement on the Quality Indicators also was high (88%).[31]

Exhibit A-2
Average Agreement on Coding Videotaped Read Alouds with the *RAP*
(% Agreement between 14 Observers and Criterion Coding)

Codes on the *RAP*	**Average % Agreement with Criterion Coding**
Instructional Behavior	%
• Pre-reading strategies	96%
• Reading strategies	95
• Post-reading strategies	98
All Pre-reading, reading, post-reading codes	96
Quality Indicators	%
• Vocabulary links	100%
• Adult use of open-ended questions	94
• Depth of post-reading activity	91
All Quality Indicators	92%

Inter-Rater Agreement from Live Observations

Inter-rater agreement on the OMLIT was assessed as part of the training process (14 paired observations), and, subsequent to training, as part of the actual data collection (17 paired observations). The calculation of inter-rater reliability used data from both of these sources.

Classroom Literacy Opportunities Checklist (CLOC)
Scores on the *CLOC i*nclude an average score across all items and average scores on each of six components of literacy resources. Inter-agreement on the *CLOC* was based only on data from the double-coding in 17 Even Start classrooms.

The average *CLOC* rating by the two observers agreed *exactly* in 80% of the pairs (Exhibit A-3). Nine of the ten sections on the *CLOC* had reliabilities above 70%; the ratings on "literacy materials in other centers" had a lower reliability of 59%. Discussions with observers suggest that the low reliability was attributable to the difficulty of noticing individual literacy resources (a book, pencils and paper) in other centers. We will strive to increase the reliability of this section through (a) improving the definition of the item to help observers understand what they are looking for, and (b) focusing training on these items to heighten observer awareness of isolated materials in different areas of the classroom.

[31] Two quality indicators were dropped from the *RAP*, based on low agreement. "Level of child engagement" and "Depth of adult discussion" were eliminated, because the average agreement on the coding of videotapes was below 75% for each.

Exhibit A-3
Inter-Rater Agreement on the *CLOC*
(17 Paired Observations of Early Childhood Classrooms)

CLOC Codes[a] (# items)	Average % Agreement[b]	Range in % Agreement Across Observer Pairs
Total across all items (56)	80%	57% – 100%
• Physical layout of classrooms (5)	91	20% – 100%
• Print environment (8)	77	38% – 100%
• Books/reading area/listening area (16)	78	50% – 100%
• Writing resources (5)	81	25% – 100%
• Literacy toys and materials (7)	82	25% – 100%
• Cultural diversity (3)	73	19% – 100%
• Literacy in other centers (3)	71	20% – 100%
• Curriculum theme (9)	76	10% - 100%

[a] Each item rated on a scale of 1 - 3

[b] Based on *exact* agreement between the ratings assigned to *CLOC* items by paired observers.

Quality of Instruction in Language and Literacy (QUILL)

Inter-rater reliability on the frequency of the different types of language/literacy activities was defined as two observers selecting the exact same rating ("none," "one," "a few," or "many" instances of the literacy activity). On the quality ratings, agreement was defined as two observers selecting a quality rating that was *within one point* (on the 5-point scale).

Inter-rater agreement on the frequency of literacy activities ranged from 67% to 83%, with average agreement of 76% (Exhibit A-4). Coders agreed least often on the frequency of activities that promoted oral language and that called children's attention to the functions and features of print. On the quality ratings, agreement ranged from 68% to 94%.

Exhibit A-4
Inter-Rater Agreement on the *QUILL*
(31 Paired Observations of Early Childhood Classrooms)

QUILL Codes	Average % Agreement
Frequency of literacy activities	**Exact**
All literacy/language activities	82%
Writing activities	88
Activities to promote letter/word knowledge	82
Activities to promote oral language	67
Activities to promote functions/features of print	67
Activities to promote understanding of sounds	71
Quality of instruction in literacy	**+/- 1 Pt**
All language and literacy activities	94%
Writing activities	85
Activities to promote letter/word knowledge	85
Activities to promote oral language	87
Activities to promote functions/features of print	68
Activities to promote understanding of sounds	69

Reading Aloud Profile—The RAP

The rate of agreement on the *RAP* when coding read-alouds in actual classrooms was quite high, regardless of the fact that most 3-hour paired observations typically involved only 1 or 2 read-alouds. Agreement on instructional behavior before, during and after reading a book ranged from 85% to 97%, with an overall average of 90% (Exhibit A-5). (The inter-rater agreement on individual instructional codes during reading ranged from 53% to 93%.) The overall quality ratings also had high inter-rater agreement. The inter-rater agreement was around 85%, when agreement was defined as within one point; the agreement dropped substantially when agreement required both coders to derive exactly the same quality rating.

Exhibit A-5
Inter-Rater Agreement on the *RAP*
(31 Paired Observations of Early Childhood Classrooms)

***RAP* Codes**	**Average % Agreement**		**Range in % Agreement Across Observers**
Adult Behavior			
Pre-reading strategies used by teacher	89%		73% – 100%
Reading strategies used by teacher	85		64% – 100%
Post-reading strategies used by teacher	97		73% – 100%
Pre-reading, Reading, Post-reading codes combined	90		76% – 98%
Quality Indicators	**+/- 1 pt**	**Exact**	
Vocabulary links	83%	76%	NA[a]
Adult use of open-ended questions	83	64	NA
Depth of post-reading activity	85	76	NA

[a] An observer either agreed or not with the rating on the criterion coding, which means there is not a continuous range of agreement.

Classroom Literacy Instruction Profile: The CLIP

Inter-agreement on the *CLIP* was based only on data from the double-coding in 17 Even Start classrooms.

The *CLIP* measure involves a two-stage coding protocol. First, the observer determines if any of the classroom staff are involved in a literacy activity. If so, then the observer codes seven characteristics of the literacy activity. If no staff member is involved in a literacy activity, the observer records only the type of non-literacy activity that the classroom is involved in. The first aspect of inter-rater reliability that was computed for the *CLIP* was the extent to which the two coders agreed on whether or not a staff member was involved in a literacy activity during the *CLIP* coding period. For observation segments where the two raters agreed that the teacher was involved in a literacy activity, the percent agreement was computed on the seven characteristics of the literacy activity.

On average, the inter-rater agreement on the occurrence of a literacy event was 85% (Exhibit A-6). When both observers identified a literacy activity, there was very high agreement on the characteristics of that activity. The two most critical categories are the type of literacy activity and the literacy knowledge afforded, and the inter-rater agreement on these codes was above 95%. The inter-rater agreement on the quality ratings was also very high.

Exhibit A-6
Inter-Rater Agreement on the *CLIP*
(17 Paired Observations of Early Childhood Classrooms)

CLIP Codes	Average % Agreement	Range in % Agreement Across Pairs
Occurrence of literacy event		
Staff involved in literacy event or not	85%	50% – 100%
Rate of literacy activities (total # literacy events/# *CLIP*s)	94	76% – 100%
Characteristics of literacy events		
Type of literacy activity	98%	50% – 100%
Number of children involved	96	0/1
Language spoken by teacher	97	71% – 100%
Language spoken by children	97	67% – 100%
Instructional style	97	57% – 100%
Text support	98	61% – 100%
Literacy knowledge afforded	96	56% – 100%
Quality ratings		
Cognitive challenge	92%	NA
Depth of discussion	93	NA

Snapshot of Classroom Activities—The Snapshot

High inter-rater agreement was not expected for many of the *Snapshot* codes, since the allocation of children to activities could vary depending on the direction of rotation of the observer's scan of the classroom. For this reason, while we expected that observers might agree on the activities taking place in the classroom, they were much more likely to differ on the number of children they assigned to each activity. This also leads us to believe that the inter-rater reliability estimates for the *Snapshot* present an underestimate of the true level of agreement across trained observers in how they would code an idealized "stationary" classroom.

The Environment section on the *Snapshot* includes a count of the numbers of children and adults present in the classroom. There was a high level of agreement—above 80%—on all codes on the Environment (Exhibit A-7). On the Activities section of the *Snapshot*, children and adults are allocated into activities. This is the part of the *Snapshot* where small differences in timing between observers could adversely affect their agreement. As predicted, the inter-rater agreement was lowest for the categories involving numbers of children in an activity. The level of agreement on the numbers of adults in each activity also was low. On the other hand, the types of activities that each observer coded had higher inter-rater agreement (82%), as did the integration of literacy in activities (88%). Although the level of agreement at the activity level on whether or not children or adults were talking was only 71%, agreement was very high—100%— on whether or not there were ***any*** adults or children talking in any of the activities coded on a *Snapshot.*

Exhibit A-7
Inter-Rater Agreement on the *Snapshot*
(31 Paired Observations of Early Childhood Classrooms)

***Snapshot* Codes**	**Average % Agreement**	**Range in % Agreement Across Pairs[b]**
Environment		
Total # children present	88%	71% – 100%
Type of adults present: teachers/aides	81	71% – 100%
Type of adults present: other	87	71% – 100%
All codes on Environment	85	65% – 100%
Activities on *Snapshot*		
Type of activity	82%	79% – 100%
Number of children in activity	57	33% – 79%
Number of teachers in activity	80	33% – 78%
Number of aides in activity	81	55% – 92%
Number of other adults in activity	91	60% – 100%
Literacy in other activities	89	76% – 100%
Any language by child/adult in each activity	71	51% – 84%
***Snapshot*-level Codes**		
Any adult talk in *Snapshot*	100%	NA
Any child talk in *Snapshot*	100	NA
Any adult/child talk in *Snapshot*	100	NA

IRT Scaling

The QUILL ratings and CLOC constructs have undergone IRT scaling by Futoshi Yamoto, a psychometrician at Abt, which shows these constructs to have very high reliability. A separate technical report has been prepared on the IRT scaling, and this will be available soon.