Evaluation of Child Care Subsidy Strategies

Findings from Project Upgrade in Miami-Dade County

March 2007

Prepared for
Ivelisse Martinez-Beck
Richard Jakopic
U.S. Dept. of Health and Human Services
Administration for Children and Families
Office of Planning, Research and Evaluation and
The Child Care Bureau
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

Acknowledgements

This project would not have been possible without the energy and determination of two people: David Lawrence, currently President of the Early Learning Initiative Foundation and a tireless advocate for children, who as Chairman of the Board of the Miami-Dade County School Readiness Coalition recognized the importance of the study and convinced the Coalition to support it; and Lisa Blair, now President of the Family Learning Partnership, who managed to recruit centers for the study in what must be a record time of less than two months and then, as Project Upgrade Coordinator for the Coalition, oversaw the crucial first three months of implementation.

Ardene Bachoo and John Alena from Child Development Services, together with Barbara Weinstein and Ann de las Pozas from Family Central helped recruit centers and were a constant source of help and advice throughout the study. Karen Egozi and Yvette Medina worked with us over the two years of the study to solve problems as they arose and to ensure the Coalition's continued support. Our on-site research partner, Dr. Charles Bleiker, Kelly Hearn and other staff at FIU played important roles in coordinating on-site data collection activities and providing a prompt response to problems.

We would like to thank the staff at OPRE and the Child Care Bureau who made the decision to embark on this experiment and provided support and encouragement throughout the study. We are grateful to the external experts on our Technical Advisory Group and to Dr Ann Witte for constructive feedback and advice.

The positive findings in this study are a testament to the efforts made by early childhood teachers to make a meaningful difference in the lives of children for whom they care, and to the hard work of the mentors and trainers who coped with a myriad of implementation challenges with determination and intelligence. We owe them our gratitude.

Contents

This report summarizes 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 funds (4%) is set aside, with state matching funds, 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-two 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. For each outcome reported, tables showing more detailed statistical data are provided in Attachment A.

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.
- 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 of print; 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 effects of the interventions are substantially larger

[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.

than those found on similar measures in the Head Start Impact Study and more closely resemble the effects of school-based prekindergarten programs.

- 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 effect of the training and on-going mentoring provided as an integral part of the interventions was to eliminate 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. Consequently, the impacts of the interventions on child outcomes were not affected by teachers' educational achievement.

Exhibit 1

Key Impact Findings

Domain/Construct (measure)	All Teachers	Spanish-dominant Teachers	English-dominant Teachers
	Effect size	Effect size	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	**Children in Classrooms with Spanish-dominant Teachers**	**Children in Classrooms with English-dominant Teachers**
	Effect Size	**Effect Size**	**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

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

[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.

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 Miami-Dade County, the School Readiness Coalition (SRC)[3] acts as the county's fiscal agent for CCDF subsidy and quality improvement funds. In response to the President's initiative and the anticipated advent of statewide voluntary prekindergarten, the SRC embarked on an effort to improve the school readiness of low-income children. In the first phase of this effort (Spring 2003), the SRC 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 children's language development. For that reason, the SRC's stakeholder advisory committee recommended that program interventions focus on language development and early literacy. Working closely with staff at the SRC, 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 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'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. However, these challenges, while they may differ in degree, are those found in many large US cities. A successful intervention in Miami-Dade County could provide guidance for many communities beyond its borders.

---

[3] In 2005, this entity 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 LAP-D.

Research Context

This experiment focuses specifically on the development of language and emergent literacy skills. In part, this reflects the SRC's identification of 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).

The Interventions

Three language/literacy interventions were selected for the study by the SRC after a systematic and comprehensive review of potential curricula had been conducted. 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 SRC was interested in introducing a curriculum in three-year-old, as well as four-year-old classrooms);
- Has some preliminary evidence of effectiveness; and
- Can be implemented by child care staff.

SRC staff met with developers whose curricula met all or most of the criteria and selected three. 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 also included take-home components (books and materials to be used by families with children at home). The three were:

- ***Ready, Set, Leap!*** (RSL; LeapFrog SchoolHouse), a curriculum that uses interactive electronic technology and thematically-grouped children's trade books. It is a comprehensive program with activities throughout the day, and targets oral language development, phonological and print knowledge.
- ***Building Early Language and Literacy*** (B.E.L.L.; not published), 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- to 20-minute lessons.
- ***Breakthrough to Literacy*** (BTL; Wright Group/McGraw-Hill), an integrated language and literacy curriculum for preschool children built around a series of weekly books with a focus on reading aloud and answering questions about the book. Computer software provides individualized literacy activities for children, also organized around the weekly book, that focus on phonological and print knowledge. It is a comprehensive program with activities throughout the day.

Research Questions and Study Design

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 whether it is possible to train child care staff, many of whom have limited education beyond high school, to deliver such curricula with fidelity, what level of support is needed to accomplish this, and what impact the interventions had 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: given this level of training and support, teacher knowledge and attitudes will change, these changes will be reflected in their behavior and interactions with children and in the classroom environment that they create; and these changes in behavior and interactions with children, and changes in the classroom environment will result in positive impacts on children's language and emergent literacy skills. We assumed 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 differential 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 answer these questions, 164 child care centers,[5] randomly selected from a group of 200 that expressed interest and were eligible to participate,[6] were randomly assigned to one of three selected curricula or to a control group. Thirty-six centers were assigned to each curriculum group[7] and 55 to the control group. An unbalanced design was chosen because of budget constraints that limited both the number of curricula that could be tested and the number of centers that could be included in the treatment groups. One four-year-old classroom was selected in each center[8]. All children in the classroom were eligible to participate, whether or not they were receiving a subsidy.

In the treatment centers, classroom staff (the teacher and an aide, where one was present) were trained to implement the curriculum to which they were assigned. The initial training was supplemented by two refresher trainings and supported by ongoing mentoring visits. When teachers left a center, developers trained their replacements. To address the concerns of coalition staff and curriculum developers about the lack of basic literacy materials and other resources in the classrooms, the SRC provided every classroom in the study, including classrooms in the control group, with a package of literacy materials that included books, paper, pencils, crayons and markers, audio-cassette players and

---

5 164 centers were actually assigned, two more than the design called for.

6 To be eligible, a center had to: serve primarily low-income children (75% or more eligible for free-and reduced-price meals in the CACFP), including some whose care was subsidized; and have at least one four-year-old classroom with at least five children. In addition, the center could not already be testing or implementing a literacy curriculum.

7 One group had 37 centers assigned to it.

8 In centers with more than one four-year-old classroom, the one with the most subsidized children was chosen. If there were equal numbers of subsidized children in the classrooms, the one with the most children was chosen. In both cases, the selection was designed to ensure that the maximum number of low-income children had the opportunity to benefit from the intervention.

tapes. In addition, as an incentive to participate in the study, control centers received a package of materials for their infant-toddler classrooms or a set of outdoor play materials. In an effort to reduce staff turnover during the study, the coalition offered a stipend of $500 to teachers who remained at the same center, to be paid in July of each study year.

The experiment was conducted over a two-year period. Centers were recruited and randomly assigned between August and October 2003. Baseline observations were conducted before training in the interventions took place, from October to late November.[9] Initial training in the curricula took place in November and early December; refresher trainings were conducted in Spring 2004 and late August 2004. Mentors were hired and trained in late Fall 2003 and began visiting classrooms in December. Classrooms were observed in late Spring 2004 and again in late 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[10]. 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.

Study Measures

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,[11] 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 School Readiness Coalition 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. Although random assignment allows us

[9] The SRC conducted assessments of subsidized four-year-olds at the same time. These data were used to assess comparability of centers at baseline and as covariates in the outcome analysis.

[10] 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.

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

to attribute treatment-control differences in children's outcomes to the interventions, without knowing anything about the center environments, the impacts on children will be better understood if we know about the extent to which the centers themselves changed. In the worst-case scenario, if we failed to find any impacts on children, it would be important to know if the lack of impacts is the result of the failure of the interventions to effect significant changes in the centers. Further, in the event that there are child impacts, we wanted to know how these were achieved—what types of changes did occur in the centers and how much change did it take to translate into benefits for children? Therefore, the design of the study called for measuring treatment-control differences in the center environments, in addition to measuring differences in child outcomes.

If 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 as a result of the interventions.[12] 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.

[12] 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. 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.

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. Dept. of Education. The CLIO[13] 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 ELLCO and the ECERS-R, and determined that new 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. 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 than 200 classrooms in each year, and calculation of inter-observer agreement for each group of observers.

Given the more than adequate reliability of the OMLIT battery (see discussion in Attachment B), 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 along 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 what we believed to be 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 versus other samples).

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 School Readiness Coalition 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 SRC 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 SRC 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:

---

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

[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. It is not appropriate as an evaluation tool.

- The testing had to impose as little additional burden as possible on the children (and classrooms), with the goal of less than 30 minutes of testing/child; and
- The testing should focus on outcomes that were not already assessed on the LAP-D.

Further, a high proportion of the children in the study classrooms came from Spanish-speaking homes and varied substantially in their English language skills. Despite the fact that the curricula were all English language/literacy curricula, all three also provided support for Spanish-speaking children. Therefore, the test battery had to have equivalent Spanish and English language versions (and had to articulate an acceptable policy about language of testing).

From the perspective of the study, other guidelines included:

- The outcome battery needed to be sensitive to the content of the curricula, to increase the chances of detecting impacts;
- The outcome battery needed to use standardized, norm-referenced measures that provided strong scores for multivariate analyses and allowed for comparison to normal development; 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 (e.g., Drs. Christopher Lonigan of Florida State University, David Kaplan of the University of Texas) 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 National Head Start 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. English and Spanish versions of the test were available. In light of the county's administration of the LAP-D, we recommended that the additional child assessments for the experiment should use the TOPEL, since it met all of the study criteria, as well as the SRC guidelines, and the recommendation was accepted.

[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. The norming data was expected to be available by late 2005, according to the test authors and Pro-Ed, although the raw scores would be appropriate for analytic purposes. As promised, Pro-Ed released the norming data in spring 2006, in time for the experiment to use 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. (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 on these same studies may be able to convert the raw scores to standardized scores.)

Recruitment and Random Assignment

To recruit centers for the study, SRC staff sent information about the study, translated into Spanish and Haitian Creole, to the approximately 850 centers that serve subsidized children in the county. SRC staff and staff from the two central subsidy agencies then made follow-up telephone calls and screening calls to determine eligibility and interest. Abt, MDRC and SRC staff held informational meetings for center directors and staff to answer questions and explain the random assignment process. After eligibility was determined, 180 centers were randomly assigned to the four groups, allowing for some replacement of centers that dropped out before knowing their assignment[16]. Notification of assignment was provided at one large meeting to which center directors and teachers were invited. Four centers decided against participation after being reminded about the random assignment process and were replaced. Directors who reiterated their willingness to participate were asked to review and sign a Memorandum of Understanding that laid out their responsibilities and the responsibilities of the research team, and were then informed of their assignment. No centers refused their assignment. 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, as we have seen, was very low and distributed quite evenly across the four groups.

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 SRC), the baseline observation measures and the LAP-D assessments of children administered in Fall 2003. There were no significant differences between treatment and control groups. of differences. We therefore concluded that, in terms of measurable aspects of the classrooms, random assignment was successfully carried out. Exhibits A1-A4 in Attachment A provide a detailed comparison of the baseline characteristics of the four groups.

[16] There are several reasons why the potential sample quickly diminished. The first and most important reason was lack of interest in participation. Many centers in Miami are either small, for-profit businesses or faith-based entities. While both groups are heavily represented in our sample, many Protestant faith-based centers use IBEKA, a religious literacy curriculum and are committed to it. Many small business owners did not want to participate in a government-sponsored study. Other reasons for reduction in the sample had to do with eligibility – centers were using High Scope or Creative Curriculum or another off-the-shelf curriculum. Because of the timing of our recruitment (late summer), some centers did not have a fully-enrolled 4-year-old classroom and could not assure us that they would. We also used zipcode information combined with a minimum number of subsidized children to eliminate middle-class centers that might have a few subsidized slots for low-income children. We were left with a pool of about 300 centers. 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. (In the case of the small number of centers that dropped from the study because they were sold, the new owner was not interested in having us or the mentors in the center). We invited the remaining 200 to a final meeting to have the terms of the study explained once more, and 180 came to the meeting. All of these were randomly assigned to a group and people who opted out before they heard their assignment were replaced. We actually assigned 164 centers rather than 162 (B.E.L.L. began with two extra), and no-one who attended the meeting and agreed to random assignment was rejected.

Data Collection

Staff questionnaires were completed prior to the first whole-group staff training. We were able to obtain similar information on replacement teachers in most cases.

Classroom observations were conducted at three time points: in fall 2003, before the treatment interventions were implemented; in spring 2004, after approximately six months of implementation of the curricula, and in spring 2005, after approximately 18 months of implementation. Observers with a background in early childhood education were trained to standardized reliability criteria (see Attachment B for a detailed discussion) on the observation system before being allowed to conduct the classroom observations for the study.

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. Assessors were trained child testers who had been trained to standardized reliability criteria (see Attachment B for a detailed discussion) on each child measure. The child assessors were bilingual in Spanish and English and provided instructions in Spanish for children as needed. All children were tested in English. In addition, the Spanish versions of subtests were used with children whose home language was Spanish.[17].

Implementation of the interventions was assessed in a variety of ways: trainers for each developer used measures tailored to the individual curriculum; mentors 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.

Analysis Methods

Impacts on classrooms and instructional practices were analyzed using two-level hierarchical models where classrooms (level 1) were nested within randomization blocks (level 2). Treatment impacts were estimated in models that controlled for year 2003 baseline measures of teacher behavior and classroom environment, dominant language of teacher and size of center. Impacts on children's language and literacy skills, as measured by the 2005 *TOPEL*, were analyzed in three-level hierarchical linear models in which children (level 1) were nested in classrooms (level 2) and classrooms were nested in randomization blocks (level 3). Treatment impacts were estimated in

[17] While all the children were tested in English, assessors were allowed to read instructions in Spanish for Spanish-dominant children. In addition to being tested in English, children whose home language was Spanish were also assessed with the Spanish-language version of the Definitional Vocabulary subtest and part of the Phonological Awareness subtest. While the primary question for all children was whether the interventions improved their English-language competence, we were interested in a secondary question: whether the interventions improved children's Spanish-language skills. Only the results from the English-language TOPEL are presented in this report. Analysis of the Spanish-language subtests will be reported in a subsequent document.

models that controlled for child's age, sex and language spoken at home, for classroom-level mean LAP-D *Cognitive Total* scores measured in Fall 2004, dominant language of teacher and size of center. More detailed information on the specification of models is provided in Attachment A.

The impact estimates presented in this report represent: first, the effects of the three treatments on teacher behavior and the classroom environment after six and eighteen months of participation in the study; and secondly, the effects on children of between two and ten months of exposure to the curricula.

Classrooms and Teachers in Fall 2003

Across all the 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 16% of classrooms there was a mix of languages. (Exhibit A4 shows the distribution by treatment group.) In classrooms with one or more Spanish-speaking children, at least one staff member spoke Spanish.

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 Staff

More than half of the 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. More than half (58%) reported having an Associate or BA degree[18]. 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 A-4).

[18] This is a higher proportion than we expected to find. Most of the more highly-educated teachers were Spanish-speaking and had obtained their credential outside the US. Child Care may be one of a small number of job opportunities for someone with limited English, regardless of educational attainment.

Implementation of the Interventions

Although the three curricula differed from each other in a variety of ways, teachers in all three groups received comparable levels of professional development. Each curriculum developer provided two to three in-service training sessions, off-site, for all teachers and aides who were involved in implementing the curricula, as well as interested directors.[19] 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, because, in spite of incentives, there was steady attrition of teachers, all three developers provided training sessions as new staff were hired for the classrooms.

During the year, each curriculum was assigned two mentor coaches, paid for by the SRC, and supervised by on-site coordinators employed by the developers. Each mentor was responsible for approximately 18 classrooms, which she visited twice a month, on average (some required more frequent visits, especially as teachers were replaced, while others were able to be visited monthly). 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, strategies for classroom management to help children focus.

On one visit to a BTL classroom, the mentor had been working with the teacher for most of the year on shifting from large-group to small-group activities. Over the two-hour period of the visit, children were engaged in activities in five or six small groups. The mentor told us that the teacher still had some misgivings because she wasn't sure the small groups were effective without her; she felt that she needed to supervise them more closely. In this situation, the teacher was following the mentor's recommendation, but still getting accustomed to a different teaching approach.

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

[19] 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 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, and the teacher's guide in the other).

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. Ready, Set, Leap and B.E.L.L mentors reported that about 11% (4 classrooms out of 36) were not implementing at a satisfactory level; Breakthrough to Literacy mentors judged that 22% (7 classrooms) were still at a beginning level of implementation at the end of the first year. At the end of the study, a similar number of centers (3 to 4) in each group were still not implementing the curricula at a satisfactory level. In some cases, this was a teacher, newly hired late in Year 2, 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 drop these resistant teachers, but often the on-site coordinator assumed responsibility for regular visits to attempt to change practices

In interviews, mentors from all three models reported independently the same features of successful implementers: a positive attitude towards instructional change; effective classroom management; and well-organized space and materials; healthy working relationships among directors, staff and parents; and frequent individual interactions between adults and children.

They reported similar barriers to implementation across the three models: resistance to instructional change; lack of trust and cooperation between teachers and administrative staff; difficulties encountered by teachers making the transition from Spanish-language to English-language instruction; and teacher turnover.

The latter problem seems to have been 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.

Findings

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 then findings about the impact on children's language and emergent literacy skills. Finally, we discuss the impact of the interventions on the pattern of activities in the classroom, and the effect of teachers' educational background on teacher and child outcomes.

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 through 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 2 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

Exhibit 2

Impacts of Interventions on Teacher Behavior and the Classroom Environment (OMLIT, Spring 2004, 2005)

Construct	Spring 2004 Effect size	Spring 2005 Effect size
Support for Oral Language	.59***	.61***
Support for Phonological Awareness	ns	.49**
Support for Print Knowledge	.53**	.74***
Support for Print Motivation	.58***	..43**
Literacy Resources	ns	.28*
Literacy Activities	.39*	.80***

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

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

While all three interventions had significant effects on aspects of teacher behavior and the classroom environment, Exhibit 3 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.

Taken together, the interventions had somewhat different effects on Spanish-dominant vs. English-dominant teachers (Exhibit 4). As we saw earlier, almost half of the teachers in the study classroom expressed their preference for training in Spanish. Initially, the presence of such a large group of Spanish-language teachers, including some who were monolingual in Spanish, was worrying for the curriculum developers. Although all three had the ability to train and provide on-going support in Spanish, and provided literacy materials in Spanish as well as English, all three curricula were intended to enhance children's English language development, and they were concerned that their training might not be as effective for teachers whose first language was not English. These worries proved unfounded. Exhibit 4 shows that the effects on Spanish-dominant teachers were as strong as and, in some cases, stronger than the effects on English-dominant teachers.

Exhibit 3

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

Construct	Treatment 1 (RSL) Effect size	Treatment 2 (B.E.L.L.) Effect size	Treatment 3 (BTL) Effect size
Support for Oral Language	.63**	.43*	.76***
Support for Phonological Awareness	.48*	.58**	.41*
Support for Print Knowledge	.95***	.33*	.91***
Support for Print Motivation	.61**	.ns	ns
Literacy Resources	ns	.51**	ns
Literacy Activities	.98***	.50*	.89***

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

Exhibit 4

Impacts of Three Interventions on Teacher Behavior and the Classroom Environment for English-dominant vs. Spanish-dominant Teachers (OMLIT, Spring 2005)

Construct	English-dominant teachers	Spanish-dominant teachers
	Effect size	Effect size
Support for Oral Language	.55*	.63**
Support for Phonological Awareness	.52*	.43*
Support for Print Knowledge	.54*	.90***
Support for Print Motivation	ns	.59**
Literacy Resources	ns	ns
Literacy Activities	.77***	.80***

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

Impact of the Interventions on Child Outcomes

The effects of the interventions on children's language development and emergent literacy skills were assessed 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 20. The percentage of children assessed in Spring 2005 ranged from 50% to 55% of the enrollment. 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 5). However, the findings are driven by the two interventions that showed impacts on children's language and literacy development (Exhibit 6). 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[20]. When we combine the impacts for Treatments 1 and 3 (Exhibit 7), 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 5

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

Measure	Effect size
Definitional vocabulary	.22*
Phonological awareness	.28**
Print knowledge	.45***
Early literacy index	.38***

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

Exhibit 6

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

Measure	Treatment 1 (RSL)	Treatment 2 (BELL)	Treatment 3 (BTL)
	Effect size	Effect size	Effect size
Definitional vocabulary	.28*	ns	.31**
Phonological awareness	.35**	ns	.44***
Print knowledge	.65 ***	ns	.60 ***
Early literacy index	.51 ***	ns	.54***

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

[20] B.E.L.L. was delivered in two 15-20 minute sessions each day, and was a less intense treatment than the other two curricula.

Exhibit 7

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

Measure	Effect size
Definitional vocabulary	.30**
Phonological awareness	.39***
Print knowledge	.63***
Early literacy index	.53***

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

Exhibit 8 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. Exhibit 8 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 9).[21] 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[22].

It is important to remember that these outcomes are for tests administered in English.[23] 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.

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

[22] 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.

[23] While all the children were tested in English, some were also tested in Spanish, but the results are not reported here.

Exhibit 8

Impacts of Treatments 1 and 3 Combined on Child Outcomes for Children in Classrooms with Spanish-dominant Teachers vs. Children in Classrooms with English-dominant Teachers (TOPEL, Spring 2005)

Measure	Spanish-dominant Teachers	English-dominant Teachers
	Effect size	Effect size
Definitional vocabulary	.39**	ns
Phonological awareness	.55***	ns
Print knowledge	.86***	.41**
Early literacy index	.72***	.36*

Exhibit 9

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

Subtest	Spanish-Creole speaking children	English-speaking children
	Effect size	Effect size
Definitional vocabulary	.31**	.28*
Phonological awareness	.41***	.31*
Print knowledge	.68***	.34*
Early literacy index	.57***	.36**

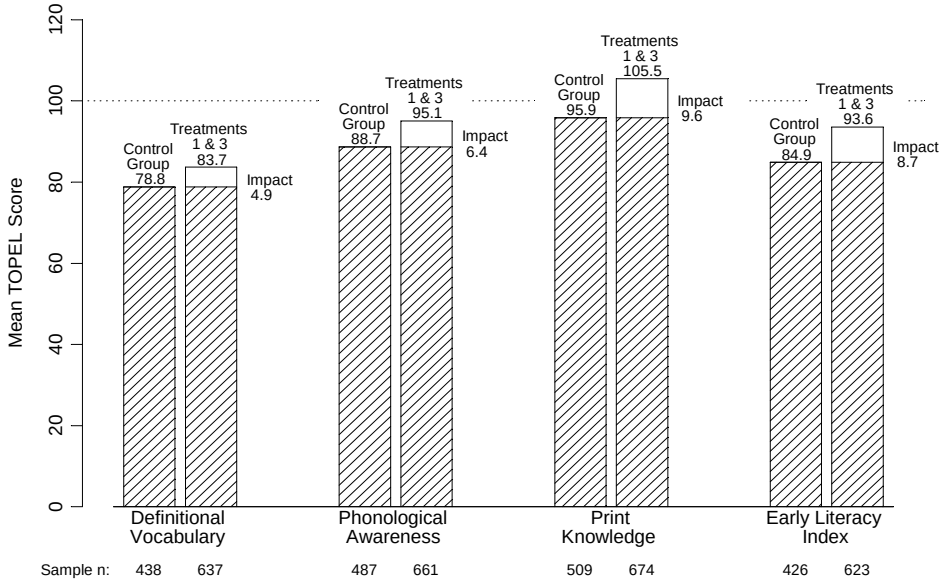
*** = p<.001, ** = p<.01, * = p<.05, NS = not significant.

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 10). 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 11, 12 and 13 show that the impacts range from a low of almost five months for Definitional Vocabulary to nine months for Print Knowledge.[24]

Exhibit 10
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.

[24] Exhibit 11 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 11. Topel Print Knowledge: Impact of Treatments 1 and 3 Relative to Growth

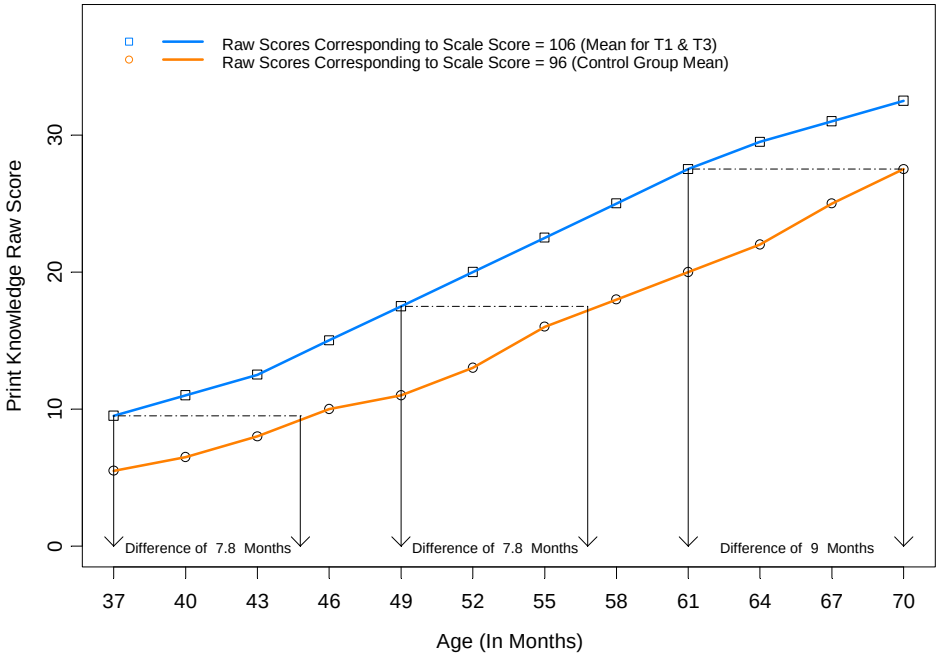


Exhibit 12. Topel Definitional Vocabulary: Impact of Treatments 1 and 3 Relative to Growth

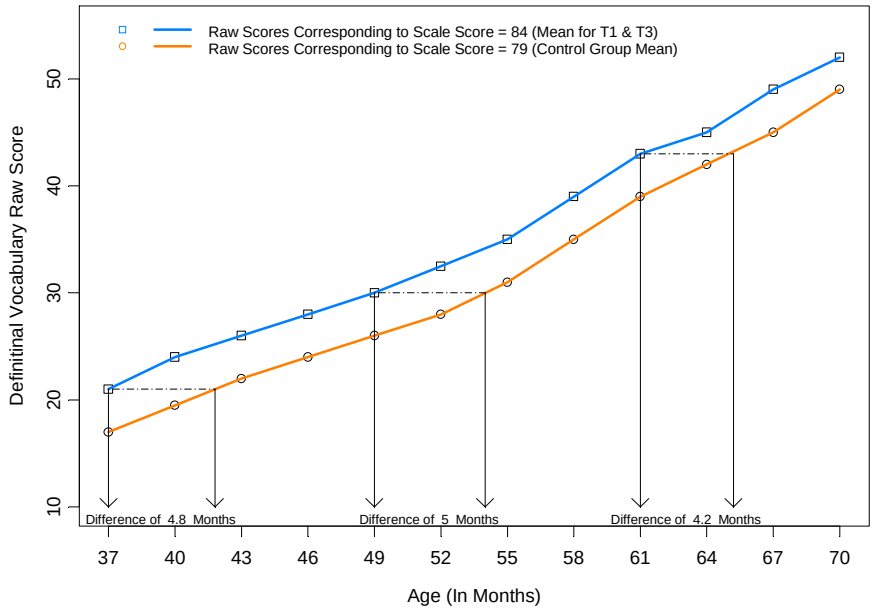
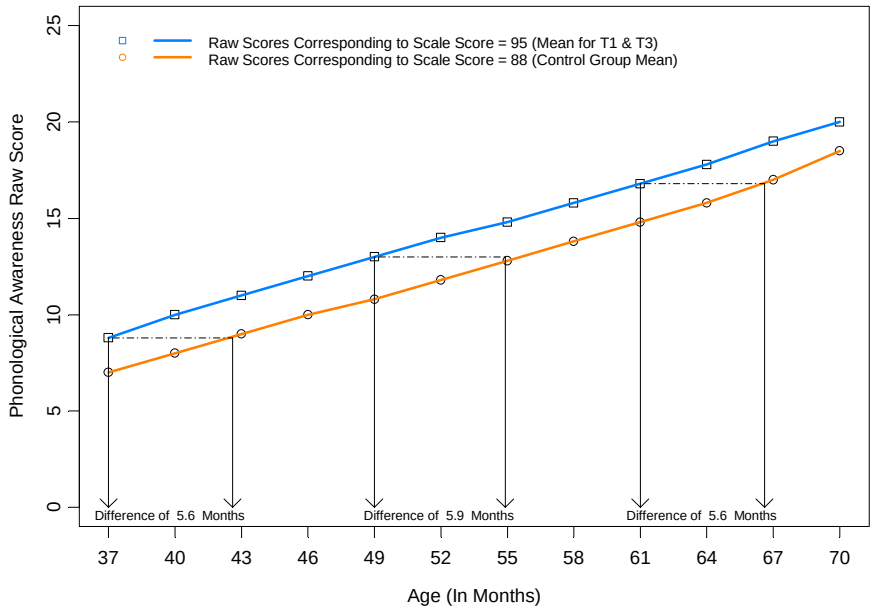


Exhibit 13 Phonological Awareness:
Impact of Treatments 1 and 3 Relative to Growth



Additional Findings

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 14 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[25] and 3% less time in routines, transitions and gross motor play.

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?

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

- 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.[26]

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 (Attachment table A1).

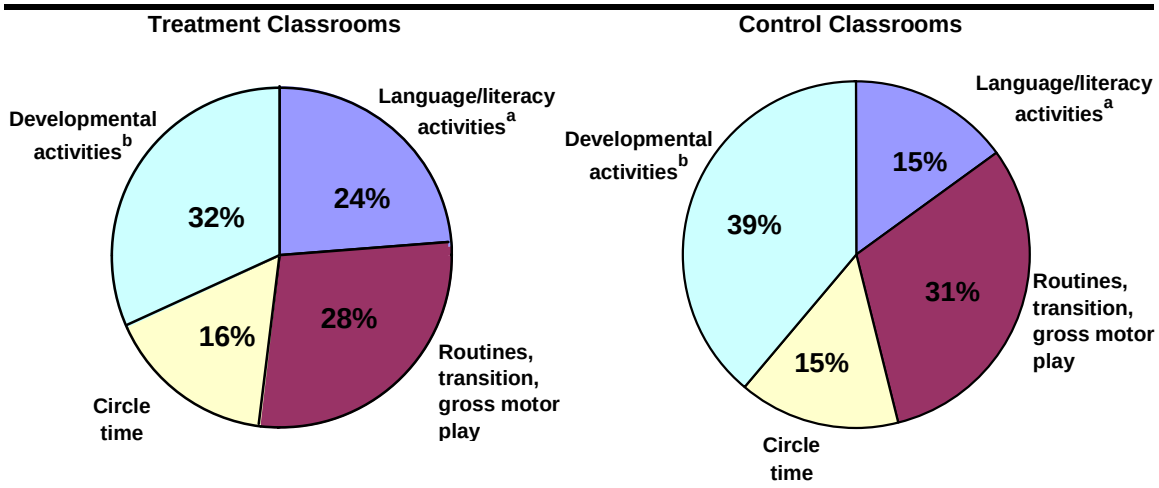
Underlying the second question is the hypothesis that better-educated teachers would be better prepared to grasp and implement a new curriculum, would therefore demonstrate more of the behaviors and interactions that support language and literacy development and would produce greater impacts on children's 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 15). 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 16 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: the impacts of the treatment were similar for children, regardless of the educational background of the teacher.

[26] Descriptions of the models used can be found in Attachment A

Exhibit 14

Children's Activities (OMLIT – Spring 2005)

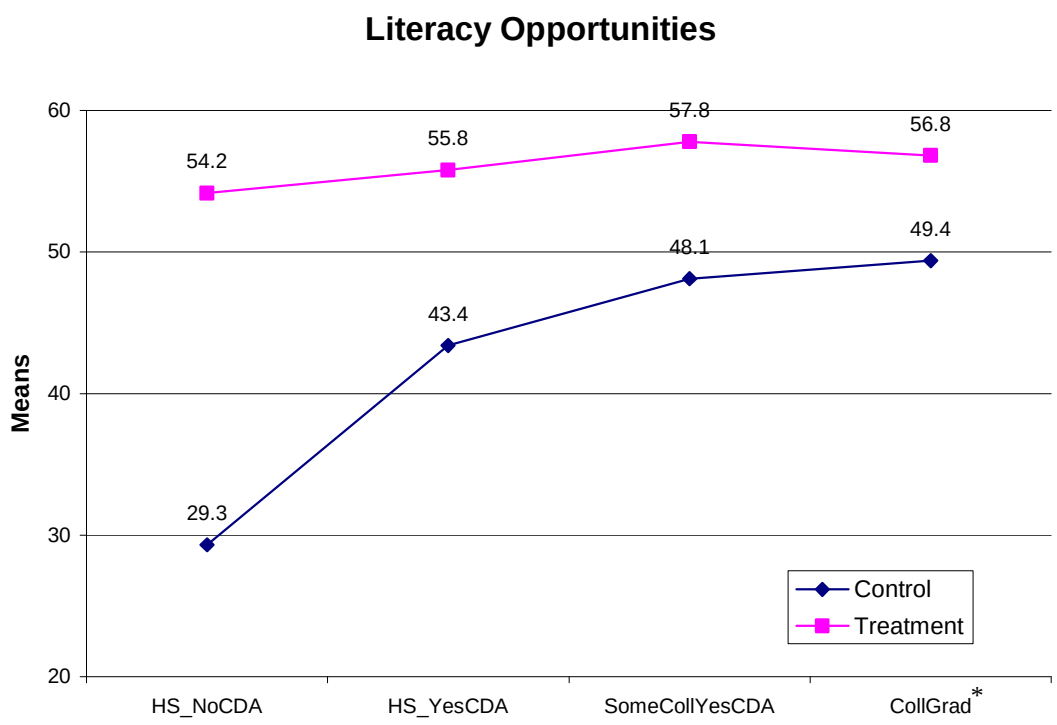


[a] Literacy/language activities include: reading (read-aloud, shared reading, child reading by himself), letters, letter-sound correspondence, writing (emergent tracing/copying, computer language programs).

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

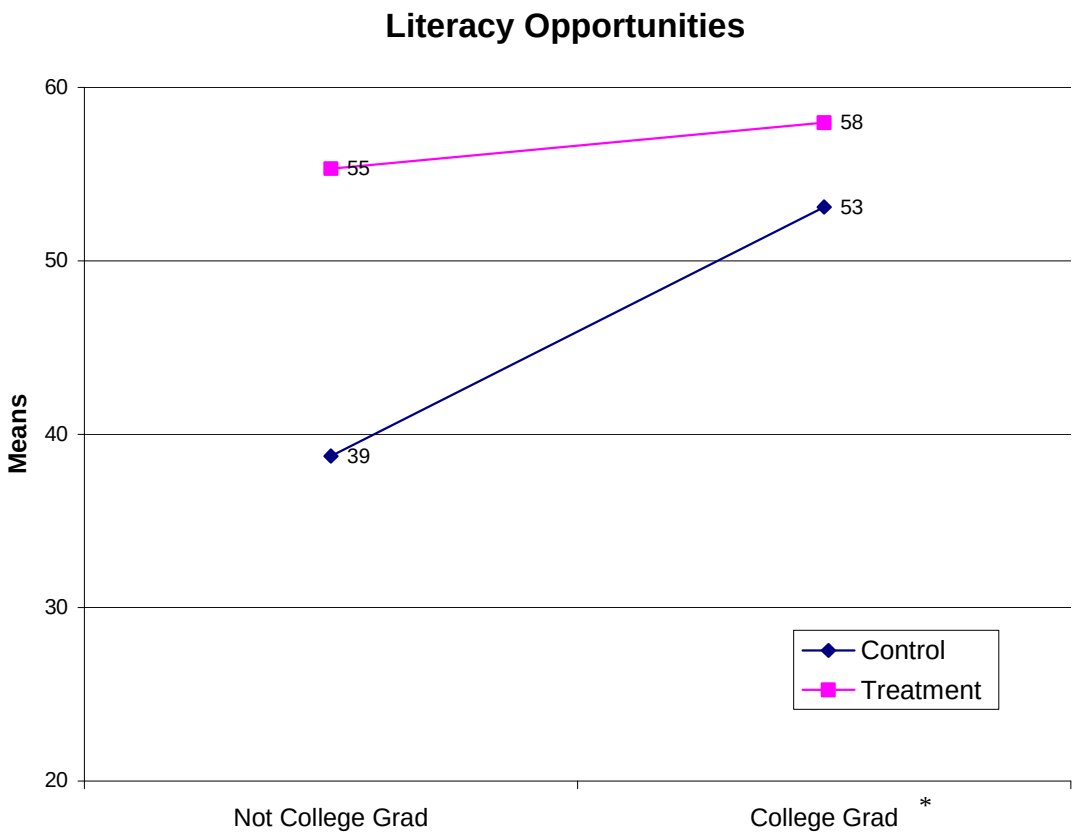
Source: OMLIT Snapshot of Classroom Activity, one day, in-class observations

Exhibit 15 Literacy Opportunities
Treatment-Teacher Education Interaction Effect (Full Sample): P=0.0408



* BA/BS Degree

Exhibit 16 Literacy Opportunities
Treatment-Teacher Education Interaction Effect (Teacher Prefers English): P=0.040



* AA/BA/BS Degree

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. As a result, impacts on children were not affected 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.

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.

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.

Attachment A

Exhibit A1

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 Differ-ence 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
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 A2

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	50.14	(5.62)	50.70	(4.97)	49.81	(4.62)	51.54	(4.26)
Support for Print Knowledge	53.27	(5.62)	53.44	(4.97)	52.89	(4.62)	53.81	(4.26)
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***	

Exhibit A3

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***

Exhibit A4

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

Measure	Control %	RSL %	BELL %	BTL %
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***

Exhibit A5 summarizes results from models for Year 2004 OMLIT construct outcomes, where all three treatment-groups combined were contrasted to 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[27], and year 2003 baseline value of the Arnett "positive, punitive, detached" construct. The models were specified as shown below.

Level-1 Model:

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

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 A5 in the column labeled "Mean Difference T-C". In Exhibit A5, 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 standard deviation shown in the column 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.

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

Exhibit A5

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.

Exhibit A6 summarizes results from models for Year 2004 OMLIT construct outcomes, where each of the treatment groups was contrasted to 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. Impacts of each of three treatments contrasted to control were estimated in models that controlled for year 2003 baseline OMLIT construct measures[28], and year 2003 baseline value of the Arnett "positive, punitive, detached" construct. 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. The values of $\hat{\gamma}_{10}, \hat{\gamma}_{20}, \hat{\gamma}_{30}$ are entered Exhibit A6 in the column labeled "Mean Difference T-C". In Exhibit A6, 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 values of the mean shown in the columns labeled "Treatment Mean (SD)" were calculated as the sum of each of the three treatment effects, $\hat{\gamma}_{10}, \hat{\gamma}_{20}, \hat{\gamma}_{30}$, and the control group mean. The standard deviation shown in those columns were calculated as the standard deviation of OMLIT construct values of each of the three treatment groups. The effect sizes were calculated by dividing the treatment effects, $\hat{\gamma}_{10}, \hat{\gamma}_{20}, \hat{\gamma}_{30}$, by the Year 2004 control group standard deviation. The p-values correspond to two-sided tests of the null hypothesis that the treatment effects are equal to zero.

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

Exhibit A6

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	ns
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	.0343
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	.72
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***			

The results corresponding to classes with Spanish-dominant teachers, shown in the top panel of Exhibit A7 were obtained by fitting the model described for Exhibit A5, to the subset of data consisting of classes with Spanish-dominant teachers. Similarly, the results for English-dominant teachers were obtained by fitting the model described in Exhibit A5 to the subset of classes with English-dominant teachers. Effect sizes were calculated by dividing the estimated treatment effect by the full sample Year 2004 control group standard deviation.

Exhibit A7

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	.2570
Literacy Activities					
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 (centers/classrooms)	***n = 28***	***n = 58***			

Exhibit A8 summarizes results from models for Year 2005 OMLIT construct outcomes, where all three treatment-groups combined were contrasted to the control group. The analytic model corresponding to Exhibit A8 is the same model as previously described from Exhibit A5. The only difference is that Exhibit A8 results are from models that were fit to the data from 2005. Therefore, the outcome variable is

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

and all other model terms are as specified previously. The effect size is calculated as the impact divided by the Year 2004 control group standard deviation.

Exhibit A8

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 A9 summarizes results from models for Year 2005 OMLIT construct outcomes, where each of the treatment groups was contrasted to the control group. The analytic model corresponding to Exhibit A9 is the same model as previously described from Exhibit A6. The only difference is that Exhibit A9 results are from models that were fit to the data from 2005. Therefore, the outcome variable is $Y_{(2005)jk}$ = OMLIT construct from year 2005 observation of classroom *j* nested in block *k*. and all other model terms are as specified previously. The effect size is calculated as the impact divided by the Year 2004 control group standard deviation.

Exhibit A9

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***			

The results corresponding to classes with Spanish-dominant teachers, shown in the top panel of Exhibit A10 were obtained by fitting the model described for Exhibit A8, to the subset of data consisting of classes with Spanish-dominant teachers. Similarly, the results for English-dominant teachers were obtained by fitting the model described in Exhibit A8 to the subset of classes with English-dominant teachers. Effect sizes were calculated by dividing the estimated treatment effect by the full sample Year 2004 control group standard deviation.

Exhibit A10

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***			

Exhibit A11 summarizes results from models for Year 2005 child-level outcomes (TOPEL scores), where all three treatment-groups combined were contrasted to the control group. The data were analyzed 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 a random intercept terms for classes and blocks. Treatment impacts (any of the three treatment groups contrasted to control) were estimated in models that controlled for child's age, sex, and language spoken at home, and for classroom-level mean LapD *Cognitive Total* scores obtained from measurements taken in the fall of 2003 or the fall of 2004. The models were specified as shown 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-2 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 Exhibit A11 in the column labeled "Mean Difference T-C". In Exhibit A11, 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 standard deviation shown in the column 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.

Exhibit A11

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	0.22	0.017
Phonological Awareness	88.74	(16.19)	93.28	(15.95)	4.54	0.28	0.003
Print Knowledge	95.89	(15.31)	102.82	(14.66)	6.93	0.45	0.000
Early Literacy Index	84.93	(16.32)	91.12	(16.70)	6.19	0.38	0.000
Sample Size (children)	***n = 509***		***n = 1014***				

Exhibit A12 summarizes results from models for Year 2005 child-level outcomes (TOPEL scores), where each of the treatment groups were contrasted to the control group. The data were analyzed in same model as specified for Exhibit A11, 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 (representing the contrast of all three treatments combined to control) that was utilized for the Exhibit A11 models.

Other than the differences shown below, all other model terms were identical to those used in the model for Exhibit A11:

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.

Exhibit A12

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	0.28	0.017
Phonological Awareness	88.74	(16.19)	94.36	(16.13)	5.62	0.35	0.003
Print Knowledge	95.89	(15.31)	105.83	(13.03)	9.94	0.65	0.000
Early Literacy Index	84.93	(16.32)	93.20	(15.77)	8.27	0.51	0.000
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	0.07	0.577
Phonological Awareness	88.74	(16.19)	89.31	(15.53)	0.57	0.04	0.767
Print Knowledge	95.89	(15.31)	97.01	(15.45)	1.11	0.07	0.565
Early Literacy Index	84.93	(16.32)	85.93	(16.93)	0.99	0.06	0.637
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	0.31	0.009
Phonological Awareness	88.74	(16.19)	95.82	(15.63)	7.08	0.44	0.000
Print Knowledge	95.89	(15.31)	105.13	(13.63)	9.24	0.60	0.000
Early Literacy Index	84.93	(16.32)	93.81	(16.12)	8.88	0.54	0.000
Sample Size (children)	***n = 509***		***n = 354***				

Exhibit A13 summarizes results from models for Year 2005 child-level outcomes (TOPEL scores), where treatment-groups 1 and 3 combined were contrasted to the control group. The results shown in Exhibit A13 are from the same model as specified for Exhibit A11, except that Treatment Group 2 data are omitted from the analysis. Thus, the treatment dummy becomes a contrast of the combined effect of Treatments 1 and 3, contrasted to control.

Other than the differences shown below, all other model terms were identical to those used in the model for Exhibit A11:

Level-2 Model:

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

where,

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

Exhibit A13

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	0.30	0.001
Phonological Awareness	88.74	(16.19)	95.09	(15.86)	6.35	0.39	0.000
Print Knowledge	95.89	(15.31)	105.51	(13.38)	9.62	0.63	0.000
Early Literacy Index	84.93	(16.32)	93.59	(15.94)	8.66	0.53	0.000
Sample Size (children)	***n = 509***		***n = 674***				

The results shown in the top panel of Exhibit A14 are from the same model as specified for Exhibit A13, except that only data from classes with Spanish-dominant teachers were included (data from classes with English-dominant teachers were omitted). Similarly, the results in the bottom panel of Exhibit A14 are from the same model as specified for Exhibit A13, except that only data from classes with English-dominant teachers were included (data from classes with Spanish-dominant teachers were omitted). Effect sizes were calculated by dividing the estimated treatment effect by the full sample Year 2005 control group standard deviation.

Exhibit A14

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	0.39	0.007
Phonological Awareness	84.64	(16.35)	93.54	(16.09)	8.90	0.55	0.000
Print Knowledge	92.69	(15.23)	105.79	(13.68)	13.10	0.86	0.000
Early Literacy Index	79.46	(16.81)	91.20	(16.64)	11.75	0.72	0.000
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	0.22	0.069
Phonological Awareness	93.47	(14.67)	97.16	(15.23)	3.69	0.23	0.086
Print Knowledge	99.84	(14.49)	106.13	(12.99)	6.30	0.41	0.001
Early Literacy Index	90.16	(13.99)	95.99	(14.54)	5.84	0.36	0.010
Sample Size (children)	***n = 228***		***n = 342***				

The results shown in Exhibit A15 are from the same model as specified for Exhibit A13, except that data from children from homes where English is the only language spoken are excluded. Effect sizes were calculated by dividing the estimated treatment effect by the full sample Year 2005 control group standard deviation.

Exhibit A15

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	0.31	0.004
Phonological Awareness	87.94	(16.62)	94.56	(16.20)	6.62	0.41	0.000
Print Knowledge	95.09	(15.43)	105.57	(13.32)	10.48	0.68	0.000
Early Literacy Index	83.33	(16.82)	92.62	(16.26)	9.29	0.57	0.000
Sample Size (children)	***n = 404***		***n = 525***				

The results shown in Exhibit A16 are also from the same model as specified for Exhibit A13, except that only data from children from homes where English is the only language spoken are included.

Exhibit A16

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

Subtest	Control Mean (SD)		Treatment Mean (SD)		Mean Difference T-C	Effect Size	P-value
Definitional Vocabulary	86.97	(12.45)	91.52	(12.86)	4.55	0.28	0.034
Phonological Awareness	91.69	(14.20)	96.78	(14.47)	5.08	0.31	0.023
Print Knowledge	98.97	(14.47)	104.11	(13.66)	5.14	0.34	0.023
Early Literacy Index	90.37	(13.17)	96.20	(14.17)	5.83	0.36	0.015
Sample Size (children)	***n = 105***		***n = 149***				

The results summarized in Exhibit A17 were obtained from the following two-level HLM model. The results in all three panels of the exhibit utilized the same model specification, but the results in the middle panel were obtained when the model was fit to data only from classes with Spanish-dominant teachers, and the results in the bottom panel correspond to the subset of data from classes with English-dominant teachers.

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 measures 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 the Exhibit, 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 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 $\beta_1 = 0$.

Exhibit A17

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.

Attachment B

Attachment B

Reliability of the Measures

Classroom Observation Measures: Observation Measures of Language and Literacy Instruction (OMLIT)

The observations focused on literacy instructional processes and environments in the classrooms, specifically on aspects of classroom practice that have been shown in empirical research to support children's language development and acquisition of early literacy skills. The complete battery of observation measures 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).

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).

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.

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

1. 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 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 B-1). The agreement was substantially higher when the definition of agreement was expanded to agreement within a point.

Exhibit B-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 B-2). Average percent agreement on the Quality Indicators also was high (88%).[29]

Exhibit B-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%

2. 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.

[29] 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.

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 B-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.

Exhibit B-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 B-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 B-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 B-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 B-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 B-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 B-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 B-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 B-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

3. 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.

Reliability of the TOPEL

Psychometric Properties

Exhibit B-8 shows the internal consistency reliabilities for subtest of the TOPEl at three ages in the standardization sample.

Exhibit B-8
Internal Consistency Reliabilities on the TOPEL

Subtest	3 years	4 years	5 years
Definitional Vocabulary	.91	.92	.91
Phonological Awareness	.85	.86	.88
Print Knowledge	,89	.94	.95

Note: Reliabilities computed from data from national standardization sample (n = 700)

Training of the TOPEL

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