Evaluation of Child Care Subsidy Strategies

Follow-up Study of the Participants in Project Upgrade in Miami-Dade

Research Brief

Office of Planning, Research and Evaluation (OPRE)
Administration on Children and Families
U.S. Department of Health and Human Services

June 2010

OPRE Report # 2011-36

Evaluation of Child Care Subsidy Strategies

Follow-up Study of the Participants in Project Upgrade in Miami-Dade

Research Brief

Submitted to:
Ivelisse Martinez-Beck
Office of Planning, Research and Evaluation (OPRE)
Administration on Children and Families
U.S. Department of Health and Human Services

Submitted by:
Cristofer S. Price
Jean I. Layzer

Contractor:
Project Director: Ann Collins
Abt Associates
55 Wheeler Street
Cambridge, MA 02138
Contract Number: 233-01-0015

Suggested citation: Administration for Children and Families (2010). Evaluation of Child Care Subsidy Strategies: Follow-up Study of Participants in Project Upgrade in Miami-Dade. Cristofer S. Price & Jean I Layzer, Washington, DC: U.S. Department of Health and Human Services.

Abt Associates Inc.

Disclaimer

The work reflected in this publication was performed under Contract Number 233-01-0015 awarded by the U.S. Department of Health and Human Services to Abt Associates. The content of this publication does not necessarily reflect the views or policies of the U.S. Department of Health and Human Services nor does mention of trade names, commercial practices, or organizations imply endorsement by the U.S. government.

Acknowledgements

We would like to acknowledge the active assistance of Dr. Jerome Levitt, Executive Director of the Office of Program Evaluation Miami-Dade County Public Schools, who worked to help identify the Upgrade study children in the Miami-Dade Public Schools administrative records, helped guide us through the research application process, and responded to many requests for additional information. We would also like to acknowledge the assistance of Fred Hicks of The Early Learning Coalition (ELC) of Miami-Dade Monroe Counties, who worked to help identify younger cohort Upgrade Study children who received subsidies to attend child-care centers during the 2005–2006 school year.

Contents

1. Introduction

Project Upgrade was a two-year randomized controlled trial (RCT) to test the effectiveness of three different language/literacy interventions on preschool-age children's language and emergent literacy skills.

The evaluation found significant impacts prior to school entry for two of the three interventions on children's language and emergent literacy skills. For the two interventions where significant impacts were found, there were impacts on both students whose teachers had received the curriculum training in English (English dominant teachers) and students whose teachers received the training in Spanish (Spanish dominant teachers), but the impacts were generally larger for the students with Spanish dominant teachers.

After the experiment ended, we followed children from the original Upgrade study to determine whether the impacts of the interventions observed in the original study persist as children progress through their first, second, and third grade years of elementary school. This research brief describes the design of and findings from the original evaluation and the follow-up study.

2. The Evaluation of Project Upgrade

Project Upgrade was a two-year experimental test of three interventions designed to improve the language and pre-literacy skills of low-income four-year-old children in Miami-Dade County, Florida. The Early Learning Coalition (ELC) of Miami-Dade Monroe Counties partnered with Abt Associates Inc. to conduct the study, providing Child Care and Development Fund (CCDF) quality improvement funds for the interventions. The evaluation (and subsequent follow-up) was funded by the federal government as part of a larger set of evaluations of strategies for the use of child care subsidy monies.

The decision to focus on language and literacy skills reflected the ELC's concern about the poor performance of children receiving child care subsidies on assessments of their language development and was also influenced by three decades of research evidence about the importance of these skills for later reading success, which is itself seen as foundational for learning (Dickinson and Tabors, 2001; Lonigan, Burgess and Anthony, 2000; Whitehurst and Lonigan, 1998; 2001). Over the last decade, there has been growing recognition of the important role that early care and education programs can play in promoting language and pre-literacy skills, especially for low-income and other vulnerable children (National Research Council, 1999; Neumann, Copple and Bredekamp, 2000; Neuman and Roskos, 1998).

Project Upgrade was intended to answer important questions about the possibility of training child care staff, many of whom have limited education beyond high school, to deliver curricula with fidelity, and about the impact of the training and support on teachers' behavior on children's language development and emergent literacy.

The hypotheses that shaped the experiment were that: with adequate training and support, teacher knowledge and attitudes will change; changes in knowledge and attitudes will be reflected, in specific ways, in behavior and interactions with children and in the classroom environment that they create; and changes in behavior and interactions with children, combined with changes in the classroom environment, will result in positive impacts on children's language and emergent literacy skills.

The study's major research questions flowed from these hypotheses and examined two major 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 aspects of the classroom environment, other than teacher language and interactions, that foster early literacy?
- Does training in and ongoing support for preschool language/literacy curricula have positive impacts on children's language development and emergent literacy skills?
- Do the interventions have different effects on teacher and child outcomes?
- Do the interventions have differential effects on teachers whose primary language is not English?
- Do the interventions have differential effects on children in classrooms with teachers whose primary language is not English?
- Do the interventions have differential effects on children whose home language is not English?

The Interventions

As is true for many early childhood interventions tested in recent years, the three interventions tested in Project Upgrade had two components: a professional development component and a curriculum component. To meet the needs of the child care centers, teachers and children in Miami-Dade County, the curriculum developers designed professional development strategies that went considerably farther than the training customarily offered in support of their curricula. It cannot be too strongly emphasized that the results reported here cannot be generalized to the curricula as they are typically used, with minimal training and no ongoing support.

The Professional Development Model

For all three interventions, a single professional development model was agreed upon by the developers, in consultation with the Early Learning Coalition of Miami-Dade County. It reflected a shared understanding of the challenges posed by the child care system in Miami-Dade County, which include: a large number of small centers with director/owners who do not necessarily have an early

childhood education background; many relatively untrained staff; and few classroom materials and resources to support literacy. The model had two important features: a staffing plan with several layers of supervision; and a training plan that featured three sequenced training sessions over an 18-month period, combined with ongoing mentoring and support over the entire period.

The Curriculum Models

The three curricula tested were selected by the Early Learning Coalition after Abt Associates shared with the ELC a systematic and comprehensive review of language/literacy curricula it had conducted. The ELC chose to test two nationally known curricula, *Ready, Set, Leap! (RSL!)* and *Breakthrough to Literacy (BTL)*. The ELC also selected *Building Early Language and Literacy (B.E.L.L.)*, a curriculum not on the Abt list, which was developed by a local academic, Dr. Wendy Cheyney. The three curricula selected differ in instructional approach, breadth of approach, materials provided, intensity and cost, but all three focus on the development of early literacy skills and knowledge. All three include take-home components (books and materials to be used by families with children at home) and tools that teachers could use to assess children's progress in the curriculum.

All three provide some materials in Spanish for children with the aim of motivating reading, regardless of the language. All three meet the Florida Preschool Language and Literacy Learning Standards; two of the three, RSL! and BTL, also meet the state standards for a comprehensive curriculum, since they include math and science concepts.

Overview of the Design

The experiment required a sample size of 162 centers (four-year-old classrooms) to be randomly assigned—36 to each of the three curricula and 54 to the control group (Exhibit 2-1).[1]

Exhibit 2-1: Expected Number of Centers, Classroom Staff, and Children, by Assignment

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

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

The rows in the exhibit show three experimental comparisons: a) a comparison of one of the treatment strands with the control group; b) a comparison of two treatment groups with each other;

[1] An unbalanced design was chosen because of budget considerations that constrained the number of curricula to be tested and the number of centers that could be included in the treatment groups.

and c) a comparison of the average outcome for the combined treatment groups with the average outcome for the control group.

Exhibit 2-2: Projected Minimum Detectable Effects[2] for the Evaluation of Project Upgrade

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

Child care centers in Miami-Dade County were eligible to participate in the study if they served some children whose care was subsidized. They could also serve, if they chose, other children from low-income families. The centers had to have at least one classroom with at least five four-year-olds enrolled at the time of recruitment. They could not be already testing or implementing a literacy curriculum. All children in the selected classrooms were eligible to participate.

The design called for a single classroom to be selected and centers to be grouped by agency affiliation and teacher's dominant language (i.e., the language she preferred to be trained in). For centers with more than one four-year-old classroom serving subsidized children, one classroom was chosen for the experiment. If one classroom had more subsidized children than the other(s), that classroom was selected. If two or more classrooms had the same number of subsidized children, then the one with the most children was chosen. If classrooms were equally large and had the same number of subsidized children, then one classroom was chosen randomly.

The recruitment and eligibility determination processes yielded a total of 300 eligible centers. Ultimately, 165 centers signed agreements to participate and received their assignments. There were no refusals after centers learned their assignments. Over the course of two years, eight centers left the study. Five left because the center was closed or sold to an owner who chose not to participate; only three left because the director decided not to continue with the curriculum to which they were assigned.

Study Measures and Data Collection

The study directly employed three types of measures: a self-administered staff questionnaire to provide information on the educational background and experience of teachers in the Upgrade classrooms; a battery of observation measures, the *Observation Measures of Language and Literacy Instruction* (OMLIT, Goodson et al., 2004), that focuses on the language and literacy environment of

[2] Calculations of minimum detectable effects (MDEs) assumed two-sided hypothesis tests with alpha level $p<0.05$, 80 percent power, and the sample sizes shown in Exhibit 2-1. For impacts on teacher behaviors, MDE calculations assumed that model terms for randomization blocks and baseline observational covariates would account for 15 percent of total variance. For impacts on child outcomes, MDE calculations assumed a class-level intra-class correlation equal to 0.10, and that model terms for blocks and class-level mean child assessment scores at baseline would account for 25 percent of class-level variance. To arrive at these estimates we used a set of measures likely to be used for the study.

and interactions within the preschool classroom, but also captures a wide range of other activities, paired with the *Arnett Caregiver Rating Scale* (Arnett, 1989), a measure 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.

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

The experiment was conducted over a two-year period. Centers were recruited and randomly assigned between August and October 2003. Observers were recruited and trained in September 2003, and retrained in spring 2004 and spring 2005. Baseline observations were conducted before training in the interventions took place, from October to late November 2003. Classrooms were observed in late spring 2004, after approximately six months of implementation of the curricula and again in late spring 2005, after 18 months of implementation. Child assessors were recruited and trained in spring 2005. Outcomes for four-year-olds were measured in late spring 2005, after between two and ten months of potential exposure to the interventions.[3] 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.

Analytic Approach

Impacts on teacher instructional behaviors and classroom environment were analyzed in two-level hierarchical linear models where teachers (at level-1) were nested with randomization blocks (level-2). The impact models included as covariates measures of the instructional behavior or classroom environment measured prior to implementation (2003), and a baseline measure of the Arnett "positive, punitive, detached" construct from 2003.

A combined estimate of the average impact of all three treatments contrasted with control was calculated, as well as estimates of the contrasts of each of the three treatments with the control group. Additionally, since the three treatments were each developed as English language curricula with English language training materials, and each had to be adapted to allow for training in Spanish for teachers who preferred to be trained in Spanish, there was a great deal of interest in examining separate impact estimates for the teachers who were trained in Spanish (Spanish dominant teachers) and teachers who were trained in English (English dominant teachers).

Impacts on measures of child language and emergent literacy were analyzed in three-level hierarchical linear models with children (level-1) nested in centers (level-2), and centers nested in randomization blocks (level-3). To increase the precision of the impact estimates, the models included terms to control for classroom-average Learning Accomplishment Profile-Diagnostic Assessment

[3] 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.

(LAP-D) cognitive total scores, child's age, sex, and home language. As in the approach described above, impacts were estimated for the combined average impact of all three treatments contrasted with control, each of the three treatments contrasted with control, and separate impacts were estimated for children in classes with Spanish dominant and English dominant teachers.[4] Additionally, since the analyses of each of the three treatments contrasted with control showed significant impacts for only two of the three interventions (RSL! and BTL), and the impacts for those two interventions did not differ significantly from one another, estimates were calculated for the combined averaged effect of RSL! and BTL contrasted with control.

A complete description of the analytic strategy used is provided in Appendix A. A more detailed description of the study and findings can be found in the final report from the study at: www.acf.hhs.gov/programs/opre/cc/upgrade_miami-dade/reports.

Results

Key findings are summarized below and in Exhibits 2-3 and 2-4. Impacts are described in terms of effect sizes.

- 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 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, definitional vocabulary, although children in the two treatment groups had significantly higher scores, they still lagged considerably behind the national norms. The impacts represent between four and nine months of developmental growth, depending on the outcome.
- The interventions resulted in a substantial increase in the time spent on language and literacy activities, both teacher-directed and child-initiated. This did not eliminate other important

[4] All children in classes with Spanish dominant teachers had Spanish as their first language.

developmental activities. Rather, time spent on each of the other activities was reduced slightly.

- There was a small but significant relationship between teachers' educational attainment and some aspects of their behavior with children before the interventions. The training and ongoing mentoring provided as an integral part of the interventions eliminated this relationship. That is, as a result of the training and mentoring, less-educated teachers looked remarkably similar to their better-educated counterparts in the extent to which they provided activities that supported literacy. Teachers' educational qualifications did not modify the impacts of the interventions on child outcomes.

Exhibit 2-3: Key Impact Findings: Teacher Instructional Behaviors and Classroom Environment

Domain/Construct (measure)	All Teachers[a] (n = 157) Effect size	Spanish-dominant Teachers[b] (n = 75) Effect size	English-dominant Teachers[c] (n = 82) 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*	.22
Classroom literacy environment (OMLIT, 2005)			
Literacy Resources	.28*	.34	.23
Literacy Activities	.80***	.80***	.77**

[a] Outcomes shown are combined outcomes for all teachers in all three treatment groups (n=104); the reference group for the impact is all teachers in the control group (n = 53).

[b] Outcomes shown are combined outcomes for all Spanish-dominant teachers in all three treatment groups (n=49); the reference group for the impact is Spanish-dominant teachers in the control group (n = 26).

[c] Outcomes shown are combined outcomes for all English-dominant teachers in all three treatment groups (n=55); the reference group for the impact is English-dominant teachers in the control group (n = 27).

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

Exhibit 2-4: Key Impact Findings: Child Language and Emergent Literacy (TOPEL, Spring 2005)[a]

	All Children (n = 1,183)	Children in Classrooms with Spanish-dominant Teachers (n =613)	Children in Classrooms with English-dominant Teachers (n = 570)
	Effect Size	Effect Size	Effect size
Definitional Vocabulary	.30***	.39**	.22
Phonological Awareness	.39***	.55***	.23
Print Knowledge	.63***	.86***	.41**
Early Literacy Index	.53***	.72***	.36**

[a] Outcomes shown are combined outcomes for the two interventions that showed significant impacts (RSL! and BTL). 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 in the final report on the impacts of the interventions (Layzer et al., 2009). Control sample sizes in the three columns are n = 509, n = 281, and n = 228, respectively.

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

3. The Follow-Up Study

To determine whether the interventions that had produced significant outcomes at the end of preschool had any lasting positive effects on their early school performance, we followed children from the original study into the Miami-Dade Public Schools (MDPS). For the original study, child outcomes were assessed in spring 2005. The following fall (fall 2005) approximately three-fourths of the children entered kindergarten, and about a quarter had to wait an additional year before entering kindergarten in fall 2006, because they were too young to meet the age-cut-off for public kindergarten. For school years 2007–2008 and 2008–2009 MDPS conducted assessments of first and second grade students in the spring of each year[5] using the SAT-10 assessment battery. MDPS regularly assesses third grade students using the FCAT. Both types of assessments produce scale-score measures of math and reading achievement, as well as subtest scores on component skills. We focused our analyses on the summary reading and math achievement scale scores. We obtained follow-up achievement data from spring assessments conducted during the 2007/2008 and 2008/2009 school years.

For the younger cohort of children, we used these data to estimate the impacts of the preschool literacy programs on reading and math achievement measured in the spring of their first and second grade years. For the older cohort of children, we estimated impacts of the preschool literacy programs on reading and math achievement measured in the spring of their second and third grade years.

Of the 1,535 children assessed in the original Florida Upgrade study, we obtained first, second, and/or third grade follow-up measurements on 1,137 children (74 percent). We also obtained follow-up measurements on 127 children who were in the study centers in the original randomized design, but

[5] This testing schedule was in place for a brief period and has been discontinued.

who were not present at the time of the initial child assessments. The reading and math achievement scores of these 127 children were included in the analyses for the current study.

Research Questions

The research questions for the Follow-up Study were driven by the findings from the original Florida Upgrade Study. Since the original study had shown significant impacts of the RSL and BTL interventions on child outcomes, and the impacts did not differ significantly from one another, we asked:

- What is the impact of preschool exposure to the RSL and BTL interventions on reading and math achievement in first, second, and third grades?[6]

In the original study we observed that the effect sizes of the estimated impacts of RSL and BTL on each the four measures of child language and emergent literacy were nearly twice as large for children in classes with Spanish dominant teachers compared with children in classes with English dominant teachers. We therefore followed up on each of these subgroups and asked:

- For students of Spanish dominant teachers,[7] what is the impact of the RSL and BTL interventions on reading and math achievement in first, second, and third grades?
- For students of English dominant teachers, what is the impact of the RSL and BTL interventions on reading and math achievement in first, second, and third grades?

In addition to the research questions above, we conducted some exploratory analyses to explore the characteristics of children in the follow-up study (Appendix B.1), to compare children who were in the follow-up study with those lost at follow-up (Appendix B.2) and to compare children retained in second grade with children who continued on to third grade (Appendix B.3). Since there were no significant impacts of the B.E.L.L. intervention in the original Project Upgrade study, we did not expect to see impacts of that intervention on the achievement scores of first, second, or third grade students. To test our expectation of no impact in these later years, we fitted impact models to the data. Additionally, we fitted impact models to estimate impacts separately for students who received RSL and BTL interventions. The results of these analyses are shown in Appendix B.4.

Creating the Analytic Sample

In the 164 centers that participated in the original evaluation, parent permissions were obtained for 1,719 children to participate in the study, which meant that the study had permission to assess the children in the child care setting and to follow them as they progressed through school. Of those with permissions, 1,535 were assessed as part of the original Project Upgrade study, 127 were absent on

---

[6] As in the original study, estimates of the average impact of RSL and BTL combined contrasted with the control group are provided to address this question. This approach results in increased power to detect effects, relative to estimation of separate impacts for each intervention. We also provide separate impact estimates for each of the interventions.

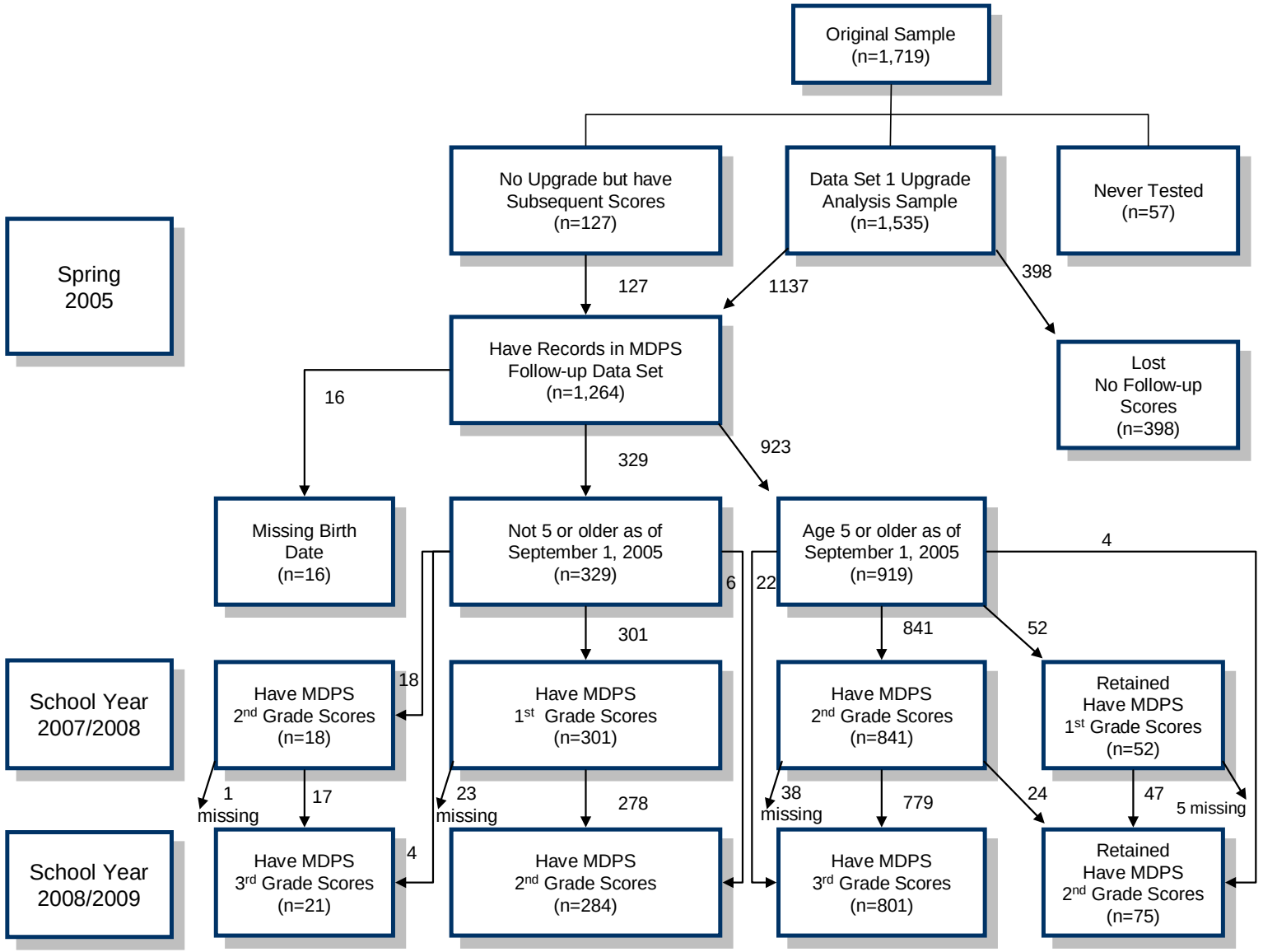
[7] Ninety-three percent of students with Spanish dominant teachers spoke Spanish at home. Fifty-five percent of students with English dominant teachers spoke Spanish at home.

testing day but were subsequently identified and tested as part of the follow-up study, and 57 were never tested (Exhibit 3-1).

Of the 1,719 children with permissions, 1,264 (74 percent) were subsequently identified in the Miami-Dade Public School (MDPS) records. About three-quarters of the children identified in the MDPS records were age five years or older as of September 1, 2005, and were therefore eligible to enter kindergarten in the Miami-Dade public schools in the 2005–2006 school year. For the majority of these children, we obtained second grade SAT-10 scores from assessments conducted in the spring of the 2007–2008 school year (n=841), and third grade FCAT scores from assessments conducted in the spring of the 2008–2009 school year (n=801). Some of the children in this group, however, either did not enter kindergarten in fall 2005 (i.e., waited an extra year to enter), or entered and were retained for a second year of kindergarten, first, or second grade and therefore do not have test scores from the same assessments and years as the majority of children in this group. For example, 56 children in this age group have first grade SAT-10 scores from the spring of the 2007–2008 school year instead of the second grade scores that the majority of children in this age cohort had. See the right-hand side of Exhibit 3-1 for details on the data available for this older cohort of children.

About one-quarter of the children identified in the MDPS records had not reached their fifth birthday on September 1, 2005, and were therefore ineligible to enter kindergarten in the Miami-Dade public schools at that time. The majority of these children did not enter kindergarten until September of 2006, and for most children in this younger cohort, we have first grade SAT-10 scores from assessments conducted in the spring of the 2007–2008 school year (n=301), and second grade SAT-10 scores from assessments conducted in the spring of the 2008–2009 school year (n=284). A very small number of children evidently either entered kindergarten in spring 2005, before their fifth birthday, or had incorrect birthdate information in our records. This group of younger cohort children has second grade scores from the 2007–2008 school year (n=18) instead of the first grade scores that are typical of this age group, and third grade scores from the 2008–2009 school year (n=21) instead of the second grade scores that are typical of this age group. See the left-hand side of Exhibit 3-1 for details on the data available for this younger cohort of children.

Exhibit 3-1: Sample Flow Diagram



Analytic Approach

To address the first research question, *"What is the impact of the RSL and BTL interventions on reading and math achievement in first, second, and third grades?"* we contrasted first, second, and third grade achievement scores of the children who had experienced the RSL and BTL interventions with the scores of children who had been in the control centers. Four separate impact analyses were conducted corresponding to the outcome data for the younger cohort who were assessed in first and second grades,[8] and outcome data for the older cohort, who were assessed in second and third grades. The analytic impact models used were of the same form as the models used in the original Project Upgrade study. They were three-level hierarchical linear models where students (level-1) were nested within preschool centers (level-2), and centers were nested within randomization blocks (level-3). A treatment indicator took the value "1" for students who had been in centers where RSL or BTL was implemented, and took the value "0" if the student had been in a control center. As in the original models, to increase the precision of the impact estimates, the models included terms to control for classroom-average Learning Accomplishment Profile-Diagnostic Assessment (LAP-D) cognitive total scores, child's age, sex, and home language. For details of model specifications, see Appendix D. The four separate subsets of data used in these analyses were formed by processes that are not completely known, and unlikely to be completely at random. For example, the factors that determined whether a child's test scores were out of sequence with the typical pattern for his/her age group (e.g., retained or moved ahead) are unknown. And the factors that determined whether a child entered the MDPS system and was tested, as opposed to not appearing in the MDPS records, are unknown. While there are limited relevant demographic data with which to compare the children who were found in the MDPS records with children who were lost at follow-up, we did conduct several analyses which provide information about the representativeness of the analytic subsets. First, for each subset we asked, were there treatment impacts on the TOPEL outcomes measured in the original Project Upgrade study? If there were no preschool impacts on a particular subset, we would not expect to see impacts persisting into subsequent years. We also describe the age distributions in the younger and older cohorts. And finally, for the students who were assessed in the original Project Upgrade study, we conducted analyses to determine if there were differences in impacts on the preschool TOPEL assessments between students who were measured at follow-up and those who were lost at follow-up.

Summary of Results

Results of analyses used to address the primary research questions indicate that, for the younger cohort, the impacts of preschool treatment with RSL and BTL interventions persisted into elementary school (Exhibit 3-2). The impacts were greater for students with Spanish dominant teachers, and were larger for first grade achievement scores than for scores in second grade.

For the older cohort, there was no evidence of positive impact of treatment for the students with English dominant teachers, nor were there significant impacts on the combined group of students with Spanish and English dominant teachers. For students with Spanish dominant teachers the estimated effect sizes of impact on second grade achievement were greater than 0.20, which is a size that is often in the realm of what is considered to be educationally meaningful but, with relatively small

[8] For the younger cohort, second grade test scores included data from those that were in second grade in the 2008–2009 school year, and those that had moved ahead and were in second grade in the 2007–2008 school year.

sample sizes, the test failed to reject the null hypothesis of no impact at the conventional $p<0.05$ criterion. In third grade, the estimated impact on reading scores of students with Spanish dominant teachers was positive (effect size 0.12) but not statistically different from zero.

Two potential explanations for the bigger treatment impacts on the younger cohorts are that the interventions may have been more beneficial if experienced at a younger age; or that many of the younger cohort of students may have remained in the same classrooms and received additional exposure to the interventions; or a combination of both. Examination of the impacts of the interventions on the preschool TOPEL assessment provides some support for the first hypothesis. Exhibit 3-3 shows that, for each of the four subgroups examined (two cohorts, two grades), there were significant impacts on the TOPEL language and emergent literacy assessments conducted in spring 2005 when all of the students were in the four-year-old classrooms. The pattern of findings there is similar to the pattern displayed in Exhibit 3-2, in that the point estimates of the impacts were larger for the younger cohort.

In order to investigate the second potential explanation for larger impact in the younger cohort (i.e., that many of the younger cohort may have received additional exposure to the interventions), we used administrative records from the child care subsidy program from the Early Learning Coalition of Miami-Dade /Monroe to identify children from the younger cohort who received a subsidy to remain in the center they attended during the original study period (September 2004–May 2005). We note that this measure is a very rough proxy for the proportion of the younger children who received an additional treatment (or control) instruction during the 2005–2006 school year because:

- Not all children received subsidies, and therefore some children who attended for an additional year do not appear on the subsidy rolls; and
- For children who did receive a subsidy, we don't know if they were placed in a class with the same teachers they had during the study year.

Of the 329 children in the younger cohort, we were able to identify 160 (48.6 percent) as having received a subsidy to attend the same center the 2005–2006 school year as they attended during the original study period (the 2004–2005 school year). Additionally, we identified one child who had attended a B.E.L.L. center in 2004–2005 but received a subsidy to attend a BTL center in 2005–2006, one child who had attended an RSL center in the study year and was subsidized to attend a B.E.L.L. center the following year, and two children who were in control centers during the study year but were subsidized to attend a B.E.L.L. center in 2005–2006. This evidence does suggest that a non-trivial proportion of younger cohort children may have received additional intervention in the year subsequent to the study year.

In order to give an indication of the magnitude of treatment effects, the impact estimates in Exhibits 3-2 and 3-3 are presented in standardized effect size units. Impacts in these units are often compared with the rule-of-thumb guidelines attributed to Cohen (1988) that suggest that standardized effect sizes of 0.20, 0.50, and 0.80 correspond to "small," "medium" and "large" effects, respectively. An additional perspective on the size and meaning of the impact estimates can be gained by comparing the effect sizes with national benchmarks. Hill et al. (2008) calculated the annual gain averaged over seven nationally normed reading achievement tests, and converted

Exhibit 3-2: Summary of Impacts at Follow-up

Grade / School Year of Assessment	Cohort[a]		BTL&RSL vs Control (Effect Size)		Spanish Dominant BTL&RSL vs Control (Effect Size)	English Dominant BTL&RSL vs Control (Effect Size)
1st Grade / 2007-08	Younger	Reading	0.36 *		0.55 *	0.16
		Math	0.46 **		0.66 **	0.19
2nd Grade / 2008-09 (also includes 2nd graders 2007-08).[9]	Younger	Reading	0.25		0.38 ~	0.08
		Math	0.25 ~		0.31	0.16
2nd Grade / 2007-08	Older	Reading	0.09		0.24 ~	-0.11
		Math	0.07		0.22 ~	-0.08
3rd Grade / 2008-09	Older	Reading	0.07		0.12	0.02
		Math	-0.01		0.18	-0.21

~ if p<0.10, * if p<0.05, ** if p<0.01.

[a] Younger cohort children had not reached their fifth birthday by September 1, 2005 and were not eligible to enter kindergarten that year. Older cohort children were 5 years or older as of September 1, 2005 and were eligible to enter kindergarten that year.

[9] The analysis included 18 children whose birthdays as recorded would have put them in the younger cohort. We believe that they are actually older children whose birthdates were entered incorrectly into the database (or possibly written incorrectly on the permission slip. Nevertheless, we adopted the more conservative approach and included them with the younger cohort.

Exhibit 3-3: Summary of Impacts on Preschool TOPEL Scores for the Four Follow-up Subgroups

Grade / School Year of Assessment	Cohort[a]		BTL&RSL vs Control (Effect Size)	Spanish Dominant BTL&RSL vs Control (Effect Size)	English Dominant BTL&RSL vs Control (Effect Size)
1st Grade / 2007-08	Younger	Definitional Vocabulary	0.17	0.04	0.28
		Phonological Awareness	0.62 **	0.64 **	0.57 *
		Print Knowledge	0.87 **	0.99 **	0.88 **
		Early Literacy Index	0.65 **	0.64 *	0.66 *
2nd Grade / 2008-09 (also includes 2nd graders 2007–2008 who were held back)	Younger	Def. Vocab.	0.09	0.08	0.22
		Phono. Aware.	0.61 **	0.70 **	0.58 *
		Print Knowledge	0.80 **	0.90 **	0.85 **
		Early Lit. Index	0.55 **	0.61 **	0.62 *
2nd Grade / 2007-08	Older	Def. Vocab.	0.27 *	0.30 ~	0.23
		Phono. Aware.	0.28 **	0.39 *	0.16
		Print Knowledge	0.54 **	0.70 **	0.36 **
		Early Lit. Index	0.47 **	0.58 **	0.34 *
3rd Grade / 2008-09	Older	Def. Vocab.	0.26 *	0.26	0.23
		Phono. Aware.	0.26 *	0.40 *	0.08
		Print Knowledge	0.55 **	0.70 **	0.39 **
		Early Lit. Index	0.47 **	0.57 **	0.31 *

~ if p<0.10, * if p<0.05, ** if p<0.01.

[a] Younger cohort children had not reached their fifth birthday by September 1, 2005 and were not eligible to enter kindergarten that year. Older cohort children were 5 years or older as of September 1, 2005 and were eligible to enter kindergarten that year.

the average gains into effect size units. They report that the average annual gains for the transitions from grades K to 1, 1 to 2, 2 to 3, and 3 to 4 are 1.52, 0.97, 0.60, and 0.36 standardized effect size units, respectively. Thus, the reading impact estimate for the full group of younger cohort students tested in the spring of first grade (effect size = 0.37) can be thought of as a little over a third of the maturational growth expected over the first grade year on the national benchmark. For math, they report annual gains averaged over six nationally normed achievement tests. For K to 1, 1 to 2, 2 to 3, and 3 to 4 they are 1.14, 1.03, 0.89, and 0.52 standardized effect size units, respectively. Thus, the math impact estimate for the full group of younger cohort students tested in the spring of first grade (effect size = 0.45) can be thought of as close to half of the maturational growth expected over the first grade year on the national benchmark.

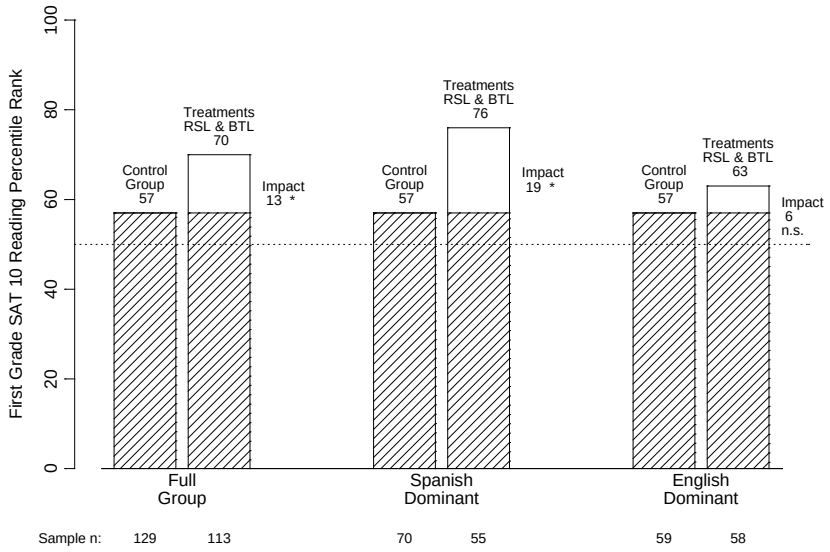
To provide an additional indication of the size of the treatment effects, we found the percentile rank corresponding to the control group mean SAT-10 score, and the percentile rank corresponding to the impact model-adjusted treatment group mean on the SAT-10 score, and examined the difference in the percentile ranks. For example, for the full group of students in the younger cohort who were tested in the first grade, the control group mean score on the SAT-10 reading test was 562. This corresponds to a percentile rank of 57. This indicates that the control group mean for this group was at a level that was higher than the scores that 57 percent of a national sample of first graders would be expected to attain. The treatment impact was 18 scale score points, making the model-adjusted mean for the treatment group equal to 580. This corresponds to a percentile rank of 70. Thus, the interventions boosted the treatment group members' mean from the 57th to the 70th percentile, or 13 percentile points (see Exhibits 3-4 to 3-6).

Discussion

The fact that the children in the "four-year-old" study classrooms in 2005 spanned a wide age range—from 36 months to over six years of age—resulted in two distinct cohorts of children, who entered kindergarten a year apart. This, combined with the variation in the MDPS assessment strategy, drove us to investigate the effects of the Project Upgrade interventions separately for children who experienced them at a younger age (and possibly for a longer period of time) and for older children who experienced at most 10 months of the interventions.

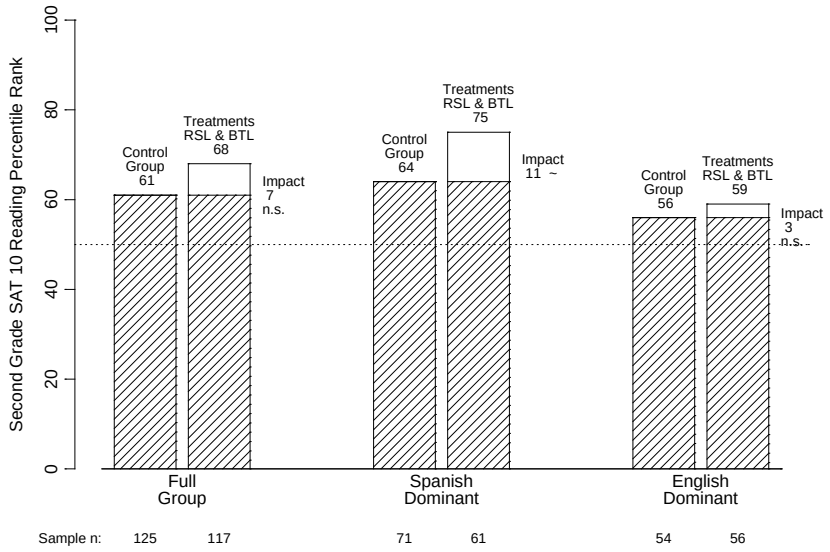
The finding that, for the former group of children, the impact of the interventions persisted through the early years of school is intriguing. We cannot be sure whether earlier exposure or continued exposure to the intervention with teachers or a combination of both factors produced the effect. The fact that the impact of the interventions was greater for the younger cohort at the end of the preschool year, on three of the four outcomes measured, offers some support for the hypothesis that, by itself, earlier exposure was a factor. It proved harder to determine whether continued exposure was also a contributing factor. Data obtained from the ELC's subsidy records showed that about half of the younger cohort remained in the same center for all or part of an additional year. However, this may overestimate or underestimate continued exposure, for two reasons. First, we do not have information on the children who were not receiving subsidies in 2005–2006. Second, we cannot be sure that the children who remained in the centers remained in the same classrooms. Closing these two gaps in our information would have required investigation of records from individual child care centers.

Exhibit 3-4: Effect Sizes Converted to Percentile Rank Differences Younger Cohort, Grade 1 SAT-10 Reading Test



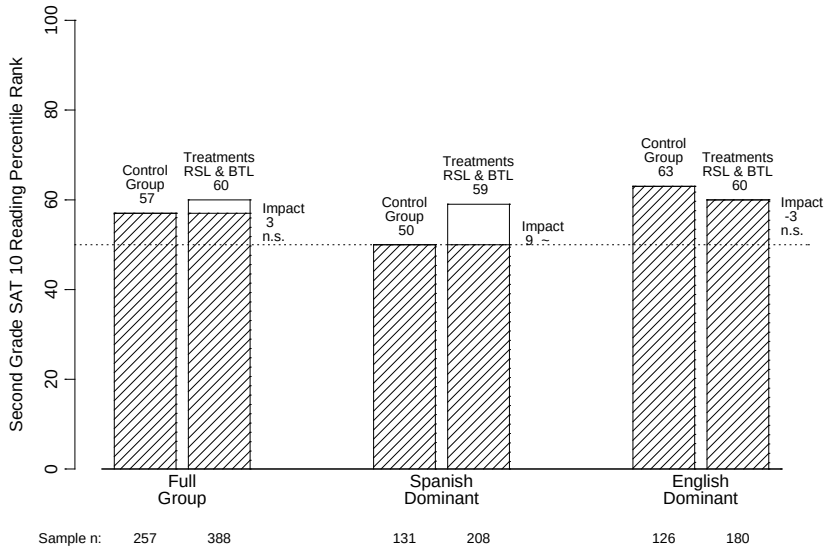
Impact ~ p<0.10, * p<0.05, ** p<0.01

Exhibit 3-5: Effect Sizes Converted to Percentile Rank Differences Younger Cohort, Grade 2 SAT-10 Reading Test



Impact ~ p<0.10, * p<0.05, ** p<0.01

Exhibit 3-6. Effect Sizes Converted to Percentile Rank Differences
Older Cohort, Grade 2 SAT-10 Reading Test



Impact ~ p<0.10, * p<0.05, ** p<0.01

The absence of enduring impacts for the older cohort of children is disappointing but hardly surprising, since it echoes similar findings from other interventions. We hoped to be able to obtain first grade SAT-10 assessments for this cohort, to see if we could learn more about when the effect of interventions dissipated; however, MDPS did not administer the test in the 2006–2007 school year.

Given that there was no immediate impact of the B.E.L.L intervention we had not expected to find that impacts would emerge in the follow-up study. Those expectations were confirmed in analyses that estimated impact at follow-up for each of the three interventions (Appendix C). In general, there were no significant differences between the impacts of RSL and BTL programs at follow-up. However, RSL appears to have significantly reduced the number of children who were held back a year in the early grades of school (see Appendix B.3, Exhibit B-3.1).

Finally, the duplication of our original finding that the interventions were most effective for children in classrooms where Spanish was the dominant language (since the teacher spoke Spanish as a first language and the children in the classrooms were primarily Spanish speakers) suggests that the right kind of intervention early in life and before school begins could help these children, in particular, enter school better prepared.

Appendix A: Analytic Strategy for the Evaluation of Project Upgrade

Introduction

This description of the analysis strategy for Project Upgrade begins with a discussion of the size of the analytic samples used to estimate program impacts on child outcomes and teacher behavior and the literacy environment. Subsequent sections describe how the outcome measures were created, the analytical models for estimating impacts on teacher behavior, the literacy environment, and child outcomes, the analytical approaches to subgroup analysis, and finally, the analytical models used for non-experimental analyses.

Analytic Samples

Analyses of impacts on teacher behavior and the literacy environment were based on data collected in the time-frame spanning fall 2003 through spring 2005. Baseline data were collected in fall 2003, prior to implementation of the experimental treatments. Data collected in spring 2004 and spring 2005 represent about 6 months and 18 months of implementation, respectively. The measurements used to estimate impacts on child outcomes were collected in spring 2005.

Baseline data on teacher behavior and the literacy environment were collected on 165 classes nested within 20 randomization blocks. Within randomization blocks, centers were randomly assigned to each of the three treatment groups, or to the control group. Data were obtained from one class per center. Consequently, the numbers of classes and centers are always identical and the terms "class" and "center" are used interchangeably throughout this discussion. Over the two years of the study, seven centers were lost to attrition, resulting in analysis samples composed of 161 classes with data from year 2004, and 157 classes with measurements from year 2005. Exhibit A-1 summarizes the number of classes in the analytic samples in each treatment group for each data collection year.

Exhibit A-1: Sizes of Analysis Samples of Classes

Treatment Group	2003 (Baseline)	2004 (1 Year Post Implementation)	2005 (2 Years Post Implementation)
RSL	38	37	36
B.E.L.L.	36	34	33
BTL	36	36	35
Control	55	54	53
Total:	165	161	157

Impacts on child outcomes were estimated using data from 1,535 children nested in 154 classes. These impact estimates correspond to children who were tested using the English language version of the child assessment instrument. Exhibit A-2 shows information on the number of children per treatment group included in the analytic data set. In 2005, there were three classes in which classroom observations were made using the OMLIT instruments, but for which no child outcome measures were obtained. Enrollment in the study classrooms was lower in spring 2005 than it had been in

earlier years,[10] so all children present in the classroom who had been in the classroom for at least two months, and whose parents had given permission for the child's assessment, were tested.

Exhibit A-2: Size of 2005 Child Outcome Analysis Sample

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

Creation of Analysis Variables

Teacher Behavior and Literacy Environment Outcome Measures

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

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

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

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

Exhibit A-3: Teaching Behaviors and Environmental Supports in OMLIT Constructs	
OMLIT Construct	**Specific Teaching Behaviors or Environmental Supports**
Support for oral language development	Reading aloud: • Time, # books • proportion of read alouds with different supports for comprehension of text (5 types) • Proportion of read alouds with open-ended questions • Quality of open-ended questions, vocabulary supports, post-reading extensions Literacy activities: • Time on oral language activities • Proportion of oral language activities with small groups • Quality of teacher/child discussion • Overall rating of oral language support • Frequency of oral language activities • Quality of oral language activities
Support for print knowledge (letters, letter-sound correspondence, writing, concepts of print)	Reading aloud: • Proportion of read-alouds with discussion of print concepts • Classroom activities • Time in activities with text, letters • Time in activities with writing (copying, emergent) • Proportion text, writing activities in small groups • Proportion activities with print involved Literacy activities: • Time on print knowledge activities • Proportion of print knowledge activities with small groups • Time on emergent writing activities • Time on copying/tracing activities • Proportion of writing activity in small groups • Proportion of print knowledge activities with small groups • Overall rating of print knowledge support • Frequency of writing activities • Quality of writing activities • Frequency of print knowledge activities • Quality of print knowledge activities
Support for phonological awareness	Reading aloud: • Proportion of read alouds with discussion of sounds Literacy activities: • Time on sounds • Proportion of activities on sounds with small groups • Overall rating of quality of support for learning sounds: • Frequency of activities on sounds • Quality of activities on sounds
Support for print motivation	Reading aloud: • Proportion of read alouds with support for print motivation • Number of RAPs • Number of minutes of reading aloud Literacy activities: • Time on activities involving print motivation • Proportion of activities on print motivation with small groups

Exhibit A-3: Teaching Behaviors and Environmental Supports in OMLIT Constructs	
OMLIT Construct	**Specific Teaching Behaviors or Environmental Supports**
Literacy resources in classroom	Adequacy of: • environmental print • text materials • writing resources • rich, integrated theme • literacy manipulatives • integration of print in other centers
Literacy activities in classroom	• teacher presents information/reads text • teacher writing • focused oral language activity • child(ren) reading/shared reading • child(ren) writing

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

Exhibit A-4: Reliability of OMLIT Constructs : Internal Consistency and Inter-Rater Reliability

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

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

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

A scale score for each of the six outcome constructs was created for each class for each year by summing the relevant standardized items, as previously described. The final step involved re-scaling

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

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

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

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

Exhibit A-5: Correlation Among 2004 OMLIT Scores

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

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

Child Outcome Measures

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

The procedure for converting raw TOPEL scores into standardized scores is straightforward given the child's chronological age and "Raw Scores to Percentile Ranks and Standard Scores" conversion tables provided in the *Test of Preschool Early Literacy Examiner's Manual*.[11] However, the conversion of the Pre-CTOPP test results into scores that are as nearly equivalent as possible to the raw TOPEL scores requires some explanation.

All of the test items on the TOPEL assessment instrument were administered as part of the Pre-CTOPP instrument. However, the Pre-CTOPP instrument included items that are not administered in the TOPEL. Furthermore, there were several differences between the two instruments in the order that items were administered. And finally, the following important difference between the two instruments in the administration instructions had to be considered. In the administration of the Pre-CTOPP, all items within each subtest were administered to a child, regardless of the number of items (s)he answered correctly. The administration instructions for the TOPEL are such that, when a child gives incorrect responses to three items in a row, no additional items on the subtest are to be administered, and all remaining items are to be scored as zeros (incorrect).

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

Measures Used as Covariates or as Descriptors of the Sample

The *Arnett Caregiver Rating Scale* (Arnett, 1989) was completed for the lead teacher in each classroom at baseline (fall 2003) and each follow-up data collection point (spring 2004 and spring 2005). The instrument produces ratings on the caregiver's emotional tone, discipline style,

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

supervision of and interest in children, and encouragement and independence. Scores were produced for three subscales, Positive Affect, Not Punitive, and Engaged (opposite of detached), and a total scale was created from the three subscales. Using the same process as described previously for the OMLIT scales, the scores on each subscale and the total score were re-scaled so that the 2004 control group had a mean of 50 and a standard deviation of 10. Scores for treatment groups or the control group from other years can be interpreted relative to the 2004 control group mean. The three subscale scores from 2003 were used to test for baseline equivalence among treatment and control groups. The 2003 Arnett total score was used as a covariate in models used to estimate treatment impacts on teacher behavior and literacy environment (OMLIT outcomes).

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

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

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

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

Analysis Methods

Baseline Balance Tests

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

- The *Cognitive Total*, *Language Total,* and *Fine Motor Total* subscales of the LAP-D
- The following measures of teacher behavior derived from the OMLIT—*Support for Oral Language*, *Support for Print Knowledge,* and *Literacy Resources*
- The Arnett subscale measures of *Positive Affect, Not Punitive,* and *Engaged*

- ❖ Proportion of teachers preferring training in Spanish
- ❖ Teacher education level

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

Level-1 Model:

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

Level-2 Model:

$$\beta_{0k} = \gamma_{00} + u_k$$
$$\beta_{1k} = \gamma_{10}$$
$$r_{jk} \sim N(0, \sigma^2)$$
$$u_k \sim N(0, \tau_{00})$$

where

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

The parameter estimate $\hat{\gamma}_{10}$ from the model above is the estimated difference between treatment and control groups at baseline. The effect size was calculated by dividing the treatment-control difference, $\hat{\gamma}_{10}$, by the Year 2004 control group standard deviation.[12] The p-value corresponds to a two-sided test of the null hypothesis that the treatment effect is equal to zero.

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

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

Estimation of Impacts on Teacher Behavior and Instructional Practices

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

[12] The OMLIT measures were scaled such that the 2004 control group standard deviation was equal to 10. Effect sizes for 2003, 2004, and 2005 OMLIT measures were all calculated relative to the 2004 control group standard deviation

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

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,[13] and year 2003 baseline value of the Arnett "positive, not punitive, not detached" construct. The models were specified as shown below.

Level-1 Model:

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

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 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.

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

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

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

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

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

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

Level-1 Model:

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

Level-2 Model:

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

where

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

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

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

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

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

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

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

Estimation of Impacts on Child Outcomes

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

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

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

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

Level-1 Model:

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

Level-2 Model:

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

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

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

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

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

Level-3 Model:

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

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

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

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

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

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

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

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

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

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

where

$Y_{(2005)ijk}$	=	TOPEL outcome measure from spring of 2005 for student *i*, nested in classroom *j* nested in block *k*.
Age_{ijk}	=	Age at time of testing of student *i*, nested in classroom *j* nested in block *k*.
$SexMale_{ijk}$	=	1 if student *i*, nested in classroom *j* nested in block *k* is male; 0 if female
$HomeLang1_{ijk}$	=	1 if home language of student *i*, nested in classroom *j* nested in block *k* is English only; 0 if HomeLang2=1 or if home language is a mix of English and Spanish, a mix of English and some other language, or if some other language is the primary language in the home
$HomeLang2_{ijk}$	=	1 if home language of student *i*, nested in classroom *j* nested in block *k* is Spanish only or a mix of English and Spanish; 0 if HomeLang1=1 or if home language is a mix of English and Spanish, a mix of English and some other language, or if some other language is the primary language in the home
Trt_{jk}	=	1 if classroom *j* nested in block *k* was in Treatment Groups 1, 2, or 3;
	=	0 if control group.
$MeanLapD\_CT_{jk}$	=	Class-level mean LAP-D Cognitive Total Score of class *j* nested in block *k* , calculated from tests administered in fall 2003 and fall of 2004.

The parameter estimate $\hat{\gamma}_{010}$ from the model above is the estimated treatment effect. The effect size was calculated by dividing the treatment effect, $\hat{\gamma}_{010}$, by the Year 2005 control group standard deviation. The p-value corresponds to a two-sided test of the null hypothesis that the treatment effect is equal to zero.

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

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

Level-2 Model:

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

where,

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

$Trt2_{jk}$ = 1 if classroom *j* nested in block *k* was in Treatment Group 2; = 0 else.
$Trt3_{jk}$ = 1 if classroom *j* nested in block *k* was in Treatment Group 3; = 0 else.

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

Level-2 Model:

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

where

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

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

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

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

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

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

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

Level 1

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

Level 2

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

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

Appendix B. Follow-up Study—Other Analyses

B.1 Descriptive Statistics on Each Subgroup

For each of the analysis data sets, Exhibits B-1.1–B-1.5 show means, standard deviations, minimums and maximums of the children's age (at the time of the TOPEL assessments, and as of September 1, 2005), and relevant test scores. The exhibits also show the proportions of children in each data set that had Spanish dominant teachers, English dominant teachers, that were male, that were in each of the three treatment groups or control group, and the proportions whose home language was English, Spanish (or Spanish and English), and other. The results show, for example, that the mean age of children at the time they were assessed in the original Project Upgrade Study in spring 2005 was 5.0 years, 52 percent of students had Spanish dominant teachers, 75 percent spoke Spanish at home, and 51 percent were male (Exhibit B-1.1). The mean age, as of September 1, 2005, of students in the younger cohort was 4.7 years (Exhibits B-1.2 and B-1.3), while the mean age of students in the older cohort was 5.5 years (Exhibits B-1.4 and B-1.5).

There were no significant differences among the treatment groups in the means of children's ages at the time of the Topel assessment (Exhibit B-1.6). Nor were there significant differences among the treatment groups in the proportions of children that were in the younger or older Cohorts (Exhibit B-1.7).

Exhibit B-1.1: Upgrade Analysis Data set

Variable	Label	N	Miss	Mean	Std Dev	Minimum	Maximum
Age[a]	child's age at time of Topel Assessment (Project Upgrade)	1535	0	5.04	0.49	3.05	6.54
Age_Sep1_2005	Child's age as of Sept 1, 2005	1512	23	5.28	0.50	3.26	6.96
TrnSpan	= 1 if child's teacher was Spanish dominant	1535	0	0.52	0.50	0	1
TrnEng	= 1 if child's teacher was English dominant	1535	0	0.48	0.50	0	1
SexMale	=1 if child is male	1535	0	0.51	0.50	0	1
RSL	=1 if in RSL treatment group	1535	0	0.21	0.41	0	1
B.E.L.L.	=1 if in B.E.L.L. treatment group	1535	0	0.23	0.42	0	1
BTL	=1 if in BTL treatment group	1535	0	0.23	0.42	0	1
Control	=1 if in Control group	1535	0	0.33	0.47	0	1
Home_Eng	=1 if home language English	1535	0	0.20	0.40	0	1
Home_Span	=1 if home language Spanish or Span/Eng	1535	0	0.75	0.43	0	1
Home_Other	=1 if home language Other	1535	0	0.05	0.22	0	1
DV_Stndzd	Definitional Vocab Standardized	1402	133	80.63	17.41	55	120
PA_Stndzd	Phonological Awareness Standardized	1477	58	91.59	16.14	55	132
PK_Stndzd	Print Knowledge Standardized	1521	14	100.27	15.21	62	142
B_Stndzd	Topel Total Standardized	1365	170	88.64	16.77	47	138

[a] The age variable includes n=23 regression imputed values of ages for individuals with missing birth dates.

Exhibit B-1.2: Young Cohort, 1st Grade SAT-10

Variable	Label	N	Miss	Mean	Std Dev	Minimum	Maximum
Age	child's age at time of Topel Assessment (Project Upgrade)	301	0	4.58	0.37	3.76	5.70
Age_Sep1_2005	Child's age as of Sept 1, 2005	301	0	4.71	0.23	4.01	4.99
TrnSpan	= 1 if child's teacher was Spanish dominant	301	0	0.51	0.50	0	1
TrnEng	= 1 if child's teacher was English dominant	301	0	0.49	0.50	0	1
SexMale	=1 if child is male	301	0	0.51	0.50	0	1
RSL	=1 if in RSL treatment group	301	0	0.20	0.40	0	1
B.E.L.L.	=1 if in B.E.L.L. treatment group	301	0	0.20	0.40	0	1
BTL	=1 if in BTL treatment group	301	0	0.18	0.38	0	1
Control	=1 if in Control group	301	0	0.43	0.50	0	1
Home_Eng	=1 if home language English	301	0	0.26	0.44	0	1
Home_Span	=1 if home language Spanish or Span/Eng	301	0	0.68	0.47	0	1
Home_Other	=1 if home language Other	301	0	0.05	0.22	0	1
DV_Stndzd	Definitional Vocab Standardized	242	59	79.83	18.41	55	115
PA_Stndzd	Phonological Awareness Standardized	256	45	89.66	15.11	57	126
PK_Stndzd	Print Knowledge Standardized	265	36	98.15	15.58	73	135
B_Stndzd	Topel Total Standardized	231	70	87.26	16.82	54	125
Yn1_ReadScore	1st Grade SAT10 Reading	301	0	570.44	50.36	454	667
Yn1_SatMathScore	1st Grade SAT10 Math	301	0	556.23	41.55	459	671

Exhibit B-1.3: Young Cohort, 2nd Grade SAT-10

Variable	Label	N	Miss	Mean	Std Dev	Minimum	Maximum
Age	child's age at time of Topel Assessment (Project Upgrade)	302	0	4.60	0.39	3.40	5.73
Age_Sep1_2005	Child's age as of Sept 1, 2005	302	0	4.72	0.23	3.70	4.99
TrnSpan	= 1 if child's teacher was Spanish dominant	302	0	0.55	0.50	0	1
TrnEng	= 1 if child's teacher was English dominant	302	0	0.45	0.50	0	1
SexMale	=1 if child is male	302	0	0.50	0.50	0	1
RSL	=1 if in RSL treatment group	302	0	0.21	0.40	0	1
B.E.L.L.	=1 if in B.E.L.L. treatment group	302	0	0.20	0.40	0	1
BTL	=1 if in BTL treatment group	302	0	0.18	0.39	0	1
Control	=1 if in Control group	302	0	0.41	0.49	0	1
Home_Eng	=1 if home language English	302	0	0.25	0.43	0	1
Home_Span	=1 if home language Spanish or Span/Eng	302	0	0.71	0.46	0	1
Home_Other	=1 if home language Other	302	0	0.05	0.22	0	1
DV_Stndzd	Definitional Vocab Standardized	240	62	80.21	18.61	55	115
PA_Stndzd	Phonological Awareness Standardized	257	45	90.18	15.09	57	126
PK_Stndzd	Print Knowledge Standardized	267	35	99.52	15.69	73	136
B_Stndzd	Topel Total Standardized	232	70	88.12	16.86	54	125
Yn2_ReadScore	2nd Grade SAT10 Reading	284	18	613.62	39.29	513	729
Yn2_SatMathScore	2nd Grade SAT10 Math	284	18	597.17	46.22	484	716

Exhibit B-1.4: Older Cohort, 2nd Grade SAT-10

Variable	Label	N	Miss	Mean	Std Dev	Minimum	Maximum
Age	child's age at time of Topel Assessment (Project Upgrade)	841	0	5.37	0.39	3.88	6.68
Age_Sep1_2005	Child's age as of Sept 1, 2005	841	0	5.53	0.29	5.00	6.96
TrnSpan	= 1 if child's teacher was Spanish dominant	841	0	0.54	0.50	0	1
TrnEng	= 1 if child's teacher was English dominant	841	0	0.46	0.50	0	1
SexMale	=1 if child is male	841	0	0.50	0.50	0	1
RSL	=1 if in RSL treatment group	841	0	0.22	0.41	0	1
B.E.L.L.	=1 if in B.E.L.L. treatment group	841	0	0.23	0.42	0	1
BTL	=1 if in BTL treatment group	841	0	0.25	0.43	0	1
Control	=1 if in Control group	841	0	0.31	0.46	0	1
Home_Eng	=1 if home language English	841	0	0.16	0.37	0	1
Home_Span	=1 if home language Spanish or Span/Eng	841	0	0.79	0.41	0	1
Home_Other	=1 if home language Other	841	0	0.05	0.21	0	1
DV_Stndzd	Definitional Vocab Standardized	709	132	80.34	16.75	55	114
PA_Stndzd	Phonological Awareness Standardized	745	96	91.74	15.78	55	126
PK_Stndzd	Print Knowledge Standardized	761	80	101.29	14.25	62	124
B_Stndzd	Topel Total Standardized	696	145	88.63	15.76	47	123
Ol2_ReadScore	2nd Grade SAT10 Reading	841	0	606.67	39.65	476	729
Ol2_SatMathScore	2nd Grade SAT10 Math	840	1	588.35	39.29	484	716

Exhibit B-1.5: Older Cohort, 3rd Grade FCAT

Variable	Label	N	Miss	Mean	Std Dev	Minimum	Maximum
Age	child's age at time of Topel Assessment (Project Upgrade)	801	0	5.37	0.39	4.57	6.68
Age_Sep1_2005	Child's age as of Sept 1, 2005	801	0	5.53	0.29	5.00	6.96
TrnSpan	= 1 if child's teacher was Spanish dominant	801	0	0.55	0.50	0	1
TrnEng	= 1 if child's teacher was English dominant	801	0	0.45	0.50	0	1
SexMale	=1 if child is male	801	0	0.49	0.50	0	1
RSL	=1 if in RSL treatment group	801	0	0.22	0.42	0	1
B.E.L.L.	=1 if in B.E.L.L. treatment group	801	0	0.23	0.42	0	1
BTL	=1 if in BTL treatment group	801	0	0.24	0.43	0	1
Control	=1 if in Control group	801	0	0.31	0.46	0	1
Home_Eng	=1 if home language English	801	0	0.15	0.36	0	1
Home_Span	=1 if home language Spanish or Span/Eng	801	0	0.80	0.40	0	1
Home_Other	=1 if home language Other	801	0	0.04	0.20	0	1
DV_Stndzd	Definitional Vocab Standardized	680	121	80.49	16.69	55	114
PA_Stndzd	Phonological Awareness Standardized	715	86	92.02	15.91	55	126
PK_Stndzd	Print Knowledge Standardized	729	72	101.84	14.08	62	124
B_Stndzd	Topel Total Standardized	669	132	88.97	15.67	47	123
Ol3_FcatMathScore	3rd Grade FCAT Reading	801	0	350.43	58.85	100	500
Ol3_ReadScore	3rd Grade FCAT Math	800	1	317.04	51.52	100	500

Exhibit B-1.6: Age at Time of Topel Assessment (Spring 2005) by Treatment Groups

Intervention:	RSL Mean	B.E.L.L. Mean	BTL Mean	Control Mean	p-value[a]
Age	5.04	5.05	5.13	4.98	0.07

Data from 1,535 children analyzed in Project Upgrade impact analyses

[a] P-value is from 3 degree-of-freedom F-test of whether mean ages differ among treatment groups. The test is from a three-level hierarchical linear model where students (level-1) are nested in centers (level-2) and centers are nested in randomization blocks (level-3). In this model the dependent variable is age and the independent variables are 3 dummy variables that represent the four treatment groups.

Exhibit B-1.7: Percentage of Children in Each Treatment Group that were in the Young and Older Cohorts

Intervention:	RSL %	B.E.L.L. %	BTL %	Control %	p-value[a]
In Young Cohort	26.6	24.1	21.6	34.3	0.11
In Older Cohort	73.4	75.9	78.4	65.7	

Data from 1,248 children in MDPS data that were in classified as belonging to young or older cohorts (see Exhibit 3.1)

[a] P-value is from 3 degree-of-freedom F-test of whether the proportion in the older cohort varies among treatment groups. The test is from a three-level hierarchical linear model where students (level-1) are nested in centers (level-2) and centers are nested in randomization blocks (level-3). In this model the dependent variable is age and the independent variables are 3 dummy variables that represent the four treatment groups.

B.2. TOPEL Scores of Children in the Follow-up Sample, and Lost at Follow-up

The flow chart in Exhibit 3.1 shows that there were 398 children that were in the original Project Upgrade analysis data set, but who were not found in the MDPS records. In this section we explore the characteristics of these children that were "lost at follow-up." The analysis is focused on the 1,535 children that were in the original Project Upgrade analysis data set, and contrasts the characteristics of the 398 that were lost at follow-up to the 1,137 that were measured at follow-up. The results summarized in Exhibits B-2.1 – B-2.5 indicate that the lost at follow-up group:

- was younger
- was less likely to speak Spanish at home
- had higher TOPEL scores on three of the four TOPEL assessments

But there were no significant differences between the lost and measured at follow-up groups on:

- intervention group status
- treatment impacts
- proportions that were male

Exhibit B-2.1: Lost at Follow-up by Intervention

Intervention:	RSL		B.E.L.L.		BTL		Control		
Lost at Follow-up	n	(%)	n	%	n	(%)	n	(%)	n
No	245	(76.6)	256	(74.0)	263	(74.1)	373	(72.6)	1,137
Yes	75	(24.4)	90	(26.0)	92	(25.9)	141	(27.4)	398
	320	(100)	346	(100)	355	(100)	514	(100)	1,535

Chi-square test of independence between lost at follow-up and treatment group: p=0.65.

Exhibit B-2.2: Lost at Follow-up by Language Spoken at Home

Home Language:	English Only		Spanish Only or Spanish & English		Other		
Lost at Follow-up	n	(%)	n	%	n	(%)	n
No	205	(68.1)	882	(76.4)	50	(63.3)	1,137
Yes	96	(31.9)	273	(23.6)	29	(36.7)	398
	301	(100)	1,155	(100)	79	(100)	1,535

Chi-square test of independence between retention and home language: p=0.001

Exhibit B-2.3: Lost at Follow-up by Sex of Child

Gender:	Female		Male		
Lost at Follow-up	n	(%)	n	%	n
No	558	(73.6)	579	(74.5)	1,137
Yes	200	(26.4)	198	(25.5)	398
	758	(100)	777	(100)	1,535

Chi-square test of independence between retention and gender: p=0.69

Exhibit B-2.4: Age and TOPEL Scores for Children that were and were not Lost at Follow-up

Lost at Follow-up	No (n=1,137)		Yes (n=398)		t-test p-value
Measure	Mean	(s.d.)	Mean	(s.d.)	
Age	5.1	(0.45)	5.0	(0.58)	0.0001
TOPEL Score: Definitional Vocabulary	79.7	(17.4)	83.2	(17.3)	0.0011
TOPEL Score: Phonological Awareness	90.7	(15.9)	94.2	(16.6)	0.0002
TOPEL Score: Print Knowledge	100.0	(15.0)	100.9	(15.6)	0.34
TOPEL Score: Early Literacy Index	87.7	(16.5)	91.2	(17.4)	0.007

Exhibit B-2.5: Summary of Impacts on Preschool TOPEL Scores for Lost at Follow-up and Not Lost at Follow-up

	BTL&RSL vs Control (Effect Size)	"Lost" by Treatment Interaction Test (p-value)	Lost at Follow-up BTL&RSL vs Control (Effect Size)	Measured at Follow-up BTL&RSL vs Control (Effect Size)
Definitional Vocabulary	0.30 **	n.s.	0.39 **	0.26 **
Phonological Awareness	0.39 **	n.s.	0.48 **	0.38 **
Print Knowledge	0.63 **	n.s.	0.64 **	0.62 **
Early Literacy Index	0.53 **	n.s.	0.59 **	0.50 **

~ if p<0.10, * if p<0.05, ** if p<0.01.

B.3. Children Retained in Grade

The flow chart shown in Exhibit 3.1 shows that there were 80 older cohort children that were a grade below their peers at the times of the 2007–2008 and 2008–2009 school year spring assessments. That is, while most of their peers were in second and third grades for those assessments, these 80 children were in first and second grade. These 80 children either entered kindergarten a year later than their peers, or repeated kindergarten, first, or second grades. Many or most of these children presumably fell a year behind their age-mates due to issues related to academic achievement, (e.g. they weren't ready to enter kindergarten, or they were retained in grade). Some may have had a delayed entry into kindergarten for non-academic reasons (e.g., parent believed that it is better to be an older child in a class, rather than a younger child[14]). In this section we describe what we know about these 80 children (we refer to these as the "retained children". Specifically, we compared them to the remaining 839 older cohort children that were in second grade in 2007/2008 and/or in third grade in 2008/2009 (we refer to this latter group as the "not-retained children"). First we ask, was the intervention received (RSL, B.E.L.L., BTL, or no intervention control group) related to retention? We also ask, was the language spoken in the home related to retention, and were boys more or less likely to be retained than girls? We also compared the ages, the second grade SAT-10 reading and math assessment scores from the 2007–2008 school year, and the TOPEL assessment scores collected at the end of the intervention year in the original Project Upgrade study of the retained and not-retained children where not retained.

The results indicate that:

- There were significant differences among the treatment groups in the proportion of children that were retained.
  - Fewer children that had been in the RSL intervention were retained as compared to children in the other treatment conditions (Exhibit B-3.1).
- Males were more likely to be retained than females (Exhibit B-3.2).

[14] Six of the 80 children had August birthdays, and thus were less than a month older than the 5-year-old cut-off for eligibility to enter kindergarten.

- ❖ Language spoken at home was not related to retention (Exhibit B-3.3).
- ❖ Retained children were an average of 44 days younger than the not-retained children (Exhibit B-3.4).
- ❖ SAT-10 reading and math scores from the 2007–2008 school year were lower for retained than not-retained children (Exhibit B-3.4).
- ❖ TOPEL scores from spring 2005, at the end of the original Project Upgrade intervention year, were lower for retained than not-retained children (Exhibit B-3.4).

Exhibit B-3.1: Retention by Intervention

Intervention:	RSL		B.E.L.L.		BTL		Control		Total
Retained[a]	n	(%)	n	%	n	(%)	n	(%)	n
No	186	(97.9)	190	(89.2)	206	(90.8)	257	(88.9)	839
Yes	4	(2.1)	23	(10.8)	21	(9.2)	32	(11.1)	80
	190	(100)	213	(100)	227	(100)	289	(100)	919

[a] "Retained" = "Yes" if child was one grade level below peers. See text for details.
Chi-square test of independence between retention and treatment group: p=0.003

Exhibit B-3.2: Retention in Second Grade by Sex of Child

Gender:	Female		Male		Total
Retained[a]	n	(%)	n	%	n
No	424	(93.8)	415	(88.9)	839
Yes	28	(6.2)	52	(11.1)	80
	452	(100)	467	(100)	919

[a] "Retained" = "Yes" if child was one grade level below peers. See text for details.
Chi-square test of independence between retention and home language: p=0.008

Exhibit B-3.3: Retention in Second Grade by Language Spoken at Home

Home Language:	English Only		Spanish Only or Spanish & English		Other		Total
Retained[a]	n	(%)	n	%	n	(%)	n
No	133	(91.1)	667	(91.1)	39	(95.1)	839
Yes	13	(8.9)	65	(8.9)	2	(4.9)	80
	146	(100)	732	(100)	41	(100)	919

[a] "Retained" = "Yes" if child was one grade level below peers. See text for details.

Chi-square test of independence between retention and home language: p=0.67

Exhibit B-3.4: Age and Second Grade Test Scores for Children that were and were not Retained

Retained[a]	No (n=839)		Yes (n=80)		t-test p-value
Measure	Mean	(s.d.)	Mean	(s.d.)	
Age (as of September 1, 2005)	5.53	(0.30)	5.41	(0.26)	0.0004
2007-08 2nd Grade SAT-10 Reading Score[b]	608.7	(38.2)	538.5	(22.8)	<0.0001
2007-08 2nd Grade SAT-10 Math Score[c]	589.9	(38.5)	534.34	(26.8)	<0.0001
Spring 2005 Preschool TOPEL Definitional Vocabulary[d]	80.5	(16.8)	68.2	(15.1)	<0.0001
Spring 2005 Preschool TOPEL Phonological Awareness[e]	91.9	(15.9)	78.9	(13.3)	<0.0001
Spring 2005 Preschool TOPEL Print Knowledge[f]	101.8	(14.1)	83.7	(11.8)	<0.0001
Spring 2005 Preschool TOPEL Early Literacy Index[g]	88.9	(15.8)	71.1	(13.7)	<0.0001

Note; The difference between the mean ages of the retained and not retained children is 0.12 years which is equivalent to 44 days.

[a] "Retained" = "Yes" if child was one grade level below peers. See text for details.

[b] Retained = "No", n = 817, retained = "Yes", n = 24.

[c] Retained = "No", n = 816, retained = "Yes", n = 24.

[d] Retained = "No", n = 714, retained = "Yes", n = 50.

[e] Retained = "No", n = 749, retained = "Yes", n = 59.

[f] Retained = "No", n = 764, retained = "Yes", n = 62.

[g] Retained = "No", n = 701, retained = "Yes", n = 48.

Appendix C: Follow-up Study – Impacts for Each of the Three Treatments

Exhibit C.1. Summary of Impacts at Follow-up

Grade / School Year of Assessment	Cohort[a]		RSL vs Control (Effect Size)	B.E.L.L. vs Control (Effect Size)	BTL vs Control (Effect Size)
1st Grade / 2007-08	Younger	Reading	0.23	0.20	0.50 **
		Math	0.26	0.00	0.64 **
2nd Grade / 2008-09 (also includes 2nd graders 2007–2008)	Younger	Reading	0.11	0.13	0.44 *
		Math	0.23	0.15	0.29 ~
2nd Grade / 2007-08	Older	Reading	0.08	0.00	0.09
		Math	0.05	-0.05	0.10
3rd Grade / 2008-09	Older	Reading	0.06	0.01	0.10
		Math	-0.07	0.04	0.05

~ if p<0.10, * if p<0.05, ** if p<0.01 .

[a] Younger cohort children had not reached their fifth birthday by September 1, 2005 and were not eligible to enter kindergarten that year. Older cohort children were 5 years or older as of September 1, 2005 and were eligible to enter kindergarten that year.

Appendix D. Follow-up Study - Model Specifications

Original and follow-up impacts were estimated in three-level hierarchical linear models where students (level-1) were nested in classrooms (level-2), and classes were nested in randomization blocks (level-3). The models included a random intercept terms for classes and blocks. Treatment impacts (RSL and BTL treatment groups combined contrasted to control) were estimated in models that controlled for child's age, sex, and language spoken at home, and for classroom-level mean LAP-D Cognitive Total scores obtained from measurements taken in fall 2003 or fall 2004. The models were specified as shown below.

Level-1 Model:

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

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

Yijk	=	is a TOPEL outcome or a SAT10 achievement outcome, or a FCAT achievement outcome for student i, nested in classroom j nested in block k.
Ageijk	=	Age at time of testing of student i, nested in classroom j nested in block k.
SexMaleijk	=	1 if student i, nested in classroom j nested in block k is male; 0 if female
HomeLang1ijk	=	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
HomeLang2ijk	=	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

Trtjk	=	1 if classroom j nested in block k was in Treatment Groups 1 or 3 (RSL or
	=	BTL);
		0 if control group.
MeanLapD_CTjk	=	Class-level mean LAP-D Cognitive Total Score of class j nested in block k , calculated from tests administered in fall 2003 and fall of 2004.

The parameter estimate $\hat{\gamma}_{010}$ from the model above is the estimated treatment effect. The effect size was calculated by dividing the treatment effect, $\hat{\gamma}_{010}$, by the 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.

References

Arnett, J. (1989). Caregivers in day care centers: Does training matter? Developmental Psychology, 10, 541-552.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (second ed.). Hillsdale, NJ: Lawrence Erlbaum.

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

Goodson, B.D., Layzer, C., Smith, W.C., and Rimdzius, T. (2004). Observation Measures of Language and Literacy Instruction (OMLIT). Cambridge, MA: Abt Associates Inc.

Hill, C.J., Bloom, H.S., Black, A.R., Lipsey, M.W., (2008). Empirical Benchmarks for Interpreting Effect Sizes in Research Child Development Perspectives, Volume 2, Number 3, Pages 172-177.

Layzer, J. I., Layzer, C. J., Goodson, B.D., Price, C. (2009). Evaluation of Child Care Subsidy Strategies - Findings from Project Upgrade. U.S. Dept. of Health and Human Services Administration for Children and Families Child Care Bureau. Washington, DC.

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.

Lonigan, C.J., Wagner, R.K., Torgesen, J.K., and Rashotte, C.A. (2002). *The Preschool Comprehensive Test of Phonological and Print Processing.* Tallahassee, FL: Florida State University.

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.