

U.S. Department of Education

Washington, D.C. 20202-5335



IES ANNUAL PERFORMANCE REPORT

CFDA # 84.305G

PR/Award # R305G040145

Budget Period # 1

Report Type: Q3 Performance

****Table of Contents****

Forms

1. Grant Performance Report (ED 524B) Project Status Chart - Section A - 1	e1
2. Grant Performance Report (ED 524B) Project Status Chart - Section A - 2	e11
3. Grant Performance Report (ED 524B) Project Status Chart - Section A - 3	e17
4. Grant Performance Report (ED 524B) Project Status Chart - Section B & C	e20
<i>hurtig-sectionb</i>	e21
<i>hurtig-IRB</i>	e22
5. Grant Performance Report Cover Sheet (ED 524B) - Revised 2008	e28
<i>Hurtig executive summary</i>	e30

Narratives

1. Project Narrative - (Tables...)	e33
<i>hurtig-tables</i>	e34
2. Project Narrative - (Charts...)	e54
<i>Hurtig-charts</i>	e55
3. Project Narrative - (Program Specific Requirements...)	e67
<i>Hurtig-narrative</i>	e68

This report was generated using the PDF functionality. The PDF functionality automatically numbers the pages in this report. Some pages/sections of this report may contain 2 sets of page numbers, one set created by the applicant and the other set created by e-Report's PDF functionality. Page numbers created by the e-Report PDF functionality will be preceded by the letter e (for example, e1, e2, e3, etc.).



**U.S. Department of Education
Grant Performance Report (ED 524B)
Project Status Chart**

PR/Award #: **R305G040145**

SECTION A - Project Objectives Information and Related Performance Measures Data (See Instructions. Use as many pages as necessary.)

1 . Project Objective Check if this is a status update for the previous budget period.
Update Impact Analyses with a reclassification of students based on ELL status

. Performance Measure	Measure Type	Quantitative Data					
		Target			Actual Performance Data		
		Raw Number	Ratio	%	Raw Number	Ratio	%
	PROJ		/			/	

Explanation of Progress (Include Qualitative Data and Data Collection Information)

This report updates the report of June 2008 that presented findings on the impacts of BTL on kindergarten students and teachers at the end of the second year of implementation (2006-07). Findings are updated based on the reclassification of 230 students' ELL status (from non-ELL to ELL), which was recently corrected in our data files. The findings in this report are based on a sample of 43 schools (22 BTL and 21 control), 134 kindergarten classrooms where BTL had been in place for one or two years, and two successive cohorts of kindergarten students. The full student sample consists of 3,107 Kindergarten students with spring scores (1,788 BTL and 1,319 C). We begin with a brief summary of the findings. We then present the research questions that guided these analyses and more detailed explanation of methods and results.

Summary

- Scattered findings on teachers/classrooms after 1 year, disappear at end of 2nd year
- No child effects overall or for ELL students

Research Questions

These analyses address three research questions about impacts of BTL on kindergarten students, based on a combined

sample across the two cohorts:

1. What are the impacts of BTL on the language development and early reading comprehension of children at the end of kindergarten?
2. Do the impacts of BTL differ by students? ELL status?
3. Do kindergarten children who have been exposed to BTL for two years (preschool and kindergarten) outperform children who have been exposed to BTL for one year (kindergarten only)?

A fourth research question looks at impacts on teaching practices for teachers separately after one versus two years of implementing BTL:

4. What are the impacts of BTL on kindergarten teachers? instruction and classroom environment at the end of the first year of implementation? At the end of the second year?

At this point in the study, pretest, first, and second year outcome data have been collected on two successive cohorts of kindergartners and their classrooms. Some of the students in BTL kindergarten classrooms were also in BTL preschool classrooms; therefore, we can address the third research question listed above (although note caveats below). When the first cohort of kindergartners was studied, their teachers were in their first year of implementation, and as the second cohort entered their kindergarten year the teachers were in their second year of implementation. The analyses in this memo include data for both cohorts of students, and both years of implementation for teachers.

Data Collection

Below we describe the number of subjects in the sample for these two years of data collection and the measures used.

Sample

Two cohorts of classrooms and students comprise the study sample. Table 1 displays the numbers of schools, classrooms, and students included in the study sample. During the 2005-06 school year, baseline observation and assessment data were collected in the fall from 108 kindergarten classrooms and 1,156 students. Follow-up data were collected in the spring on 1,574 students. In the 2006-07 school year, baseline observation data were collected in the (26) kindergarten classrooms that were new to the study, having replaced classrooms that dropped from the study after Year 2. Baseline assessment data were also collected in the fall in 108 kindergarten classrooms on 1,261 students. Follow-up data were collected on 1,533 students in the spring. Across both cohorts, impacts are estimated on data from 134 classrooms and 3,107 students. For the purposes of analysis, teachers who joined the study in the second year of implementation were included in the sample of teachers who had implemented the curriculum for one year (in both treatment and control groups). Thus, the sample for 2-year teachers is much smaller than that for 1-year teachers (because although they were in the same classrooms in which 1-year teachers had been, the replacement teachers had only been in the study for one year). Attrition rates for both cohorts of children are presented in Appendix A.

Notes:

a Four potential schools were recruited to replace the schools that dropped out in the first year. One signed an agreement to participate and was randomly assigned (coin flip) to the treatment group. No additional schools were recruited.

b One school has only kindergarten classes (no pre-K).

c Seventeen teachers in schools assigned to the treatment condition left their schools or changed grade level and were replaced by 15 teachers (2 teachers were not replaced); 9 teachers in schools assigned to the control group left their schools or changed grade level and were replaced by 11 teachers. Since assignment was at the school level, when a teacher joins the study in a particular school, she is automatically assigned to the same status (treatment or control) as the other teachers at that school.

Measures

Classroom measures

The measures used for classroom observation were the same as those used in the pre-K classrooms: the QUEST, CLOC, RAP, and Arnett were administered at baseline (fall of the teacher's first year implementing BTL) and the full OMLIT (Snapshot, CLIP, RAP, QUILL, and CLOC) plus the Arnett were administered at follow-up in the spring of the teachers' first and second years in the study (first year only for those teachers who were new to the study in the 2006-07 school year). (Please see Table 2a below.) Descriptions of all of these measures can be found in the reports for the previous years of the study.

Child outcome measures

Each cohort of children was given two tests at baseline (fall) and a battery of four tests at follow-up (spring). All of the tests are individually administered, and all are described below:

Peabody Picture Vocabulary Test / Third Edition (PPVT-III; Dunn & Dunn, 1997)

This assessment measures receptive vocabulary. It was administered at baseline, in English. It has been normed, and is standardized to have a mean of 100 and a standard deviation of 15.

Preschool Comprehensive Test of Phonological and Print Processing (Pre-CTOPPP, Lonigan, Wagner, Torgesen, & Rashotte, 2002), Print Knowledge subtest: This subtest measures early knowledge about written language conventions and form as well as alphabet letters. The test requires children to identify examples of aspects of print, identify letters and written words, point to specific letters, name specific letters, say the sounds associated with specific letters, and identify letters associated with specific sounds. The test was used to provide some parity with earlier tests (this test would be administered three times to a child tested in pre-K and kindergarten) once at the start of each year and once in spring of the pre-K year. The Pre-CTOPPP has now been normed and standardized and is known as the Test of Preschool Early Literacy (TOPEL; Lonigan, Wagner, Torgesen, & Rashotte, 2006). An algorithm was used to convert scores from the Pre-CTOPPP

to TOPEL scores to be standardized.

Expressive One-Word Picture Vocabulary Test / Third Edition (EOWPVT; Brownell, 2000)

This test is a standardized, norm-referenced measure of an individual's English speaking vocabulary. Standard scores are set to have a mean of 100 and standard deviation of 15.

Woodcock Reading Mastery Test / Revised/Normative Update (WRMT-R/NU; Woodcock, 1998)

This test battery measures several important aspects of reading ability. We used three subtests from this battery: Letter Identification (at kindergarten only), Word Identification, and Word Attack. While the Letter Identification subtest is part of the Readiness cluster, the Word Identification and Word Attack subtests combined form the Basic Skills cluster. Each subtest is described briefly below. This test has also been normed.

Letter Identification

This test measures the child's ability to identify uppercase and lowercase letters. The letter forms presented include roman, italic, bold type, serif and sans serif type styles, cursive characters and special type styles. The child is shown the letter and asked to provide the name of the letter.

Word Identification

This test measures the child's ability to identify isolated words that appear in large type on the stimulus pages. To get credit, the child must produce a natural reading of the word within 5 seconds.

Word Attack

This test requires the child to read either nonsense words or words with extremely low frequency. Nearly all phonemes in the English language are represented in at least one of their major spelling patterns in the items. The test measures the child's ability to apply phonic and structural analysis skills in order to pronounce words with which s/he may be unfamiliar.

Findings

Descriptive Statistics for Study Schools

The final analytic sample for the kindergarten analyses consists of 43 schools. As can be seen in Table 2c, the schools serve largely low-income students of color; on average, 89.3% of the students attending these schools are categorized as coming from low-income families, and 85.1% of the students are minorities. The schools in the sample are large; the average school enrollment is 791, with a range of 139 to 1969 students. Mobility rates are also high in this sample of schools, averaging 23.7% and ranging from a low of 5.7% to a high of 56.1%.

Baseline Equivalence of School, Classroom, Teacher and Student Characteristics

The process of random assignment is intended to ensure that the BTL group and control group are statistically equivalent. At

the same time, with smaller samples, it is possible that random assignment resulted in two groups that were different on one or more characteristics that could be related to study outcomes. The purposes of analyzing the baseline data collected before the implementation of BTL are to both describe the study sample and identify where the random assignment procedure did not yield equivalent groups. Variables on which the schools, classrooms, teachers, or students differ were then considered as likely candidates for inclusion as covariates in our impact models.

School Characteristics

As seen in the first panel of Table 3, BTL and control schools are the same in terms of the percentage of students from low income families, the total enrollment, and mobility rate. However, BTL schools serve more minority students, on average, than control schools. The .42 of a standard deviation difference is statistically significant at the .05 level. This indicates that it could be an important covariate in the impact models.

Differences in Classroom Characteristics and Baseline Measures

Reported baseline differences here are estimated using a 2-level HLM model (classrooms nested within schools with school-level random error terms). This model employs the sample stratifiers (north, Spanish, north and Spanish interaction) and an observation year indicator as covariates.

The second panel of Table 3 summarizes teacher characteristics, for which there are not statistically significant differences, on average, across BTL and control classrooms. The third panel of Table 3 summarizes classroom characteristics for both the BTL and control classrooms. Prior to randomization, schools were stratified by whether or not instruction was likely to be delivered in Spanish or another non-English language. The baseline equivalence models indicate that, on average, BTL and control schools did not differ any of the 6 classroom and teacher characteristics tested, including percentages of non-English and English speaking students and classroom size.

In terms of classroom interactions at baseline (fourth and fifth panels of Table 3), on the QUEST and the Arnett scores, there were no statistically significant baseline differences between the two groups on any of the QUEST or Arnett measures at p less than or equal to .05 level. However, there is one QUEST subscale - supporting social and emotional development - for which the effect size could be considered substantially meaningful even though the difference is not statistically significant. We do not have a hypothesis for what kind of difference this would make in the outcomes. (This analysis was based on 111 classrooms observed in Fall 2005 and 25 classrooms that were introduced to the study in Fall 2006.)

Differences in Student Measures

As seen in the sixth panel of Table 3, there were no differences in student characteristics, on average, between children in BTL and control schools. Further, students in BTL and control schools on average did not differ significantly on their scores on either the PPVT or the Print Knowledge subscale of the Pre-CTOPP.

The following analyses summarize impacts on kindergarten teachers' instructional behaviors, their classrooms, and their students' early literacy outcomes.

Method

Impacts reported here are estimated through a 2-level HLM model (classrooms nested within schools) which uses sample stratifiers as well as an observation year indicator and percentage of minority students at the school level, which was significantly different at treatment and control schools at baseline. For the Arnett measures, baseline values on these variables are used as covariates. For the model for a typical outcome, see Appendix C.

Classroom Outcomes

Tables 4 and 5 summarize the impact of BTL on OMLIT and Arnett scores after one and two year(s) of implementation in kindergarten classrooms. Analyses of OMLIT measures after one year of implementation are based on observations of 131 classrooms, 104 of which were observed in Spring 2006. The remaining 27 classrooms were those of teachers who replaced teachers who left the study; these teachers were introduced to the study in the 2006-2007 school year and observed in Spring 2007. Arnett analyses are based on the same sample of 131 classrooms. These analyses employ school-level baseline measures as covariates. Analyses of OMLIT and ARNETT measures at the end of two years of BTL implementation are based on observations of 80 classrooms, all of which were observed in Spring 2007.

Teachers after one year of implementing BTL.

As seen in Table 4a, there is a substantial and statistically significant impact of BTL on kindergarten teachers' oral language instruction and on print motivation at the end of one year of implementation. This suggests that teachers are providing students with more opportunities to participate in classroom and instructional activities aimed at improving oral language skills, especially through book-focused activities, a key emphasis of the BTL curriculum. Examples of activities to develop children's oral language include discussions ("sharing," book-related discussions, discussions aimed at building children's vocabulary and concept knowledge), questioning as part of a shared book reading (?dialogic reading?), and conversations involving an adult and one or more children focusing on topics other than management.

Impact estimates on two other OMLIT measures are also statistically significant: proportion of time spent in literacy-related activities, and proportion of time spent in computer activities. Teachers in BTL classrooms were encouraged by coaches to weave literacy throughout the day, and both BTL training and coaching provided BTL teachers with explicit explanation and practice in designing activities to reinforce vocabulary and concepts throughout the day. Thus, it is not surprising that observers documented more literacy activities in BTL classrooms. The finding that BTL students spent 4% of their time, on average, using the computer, whereas their counterparts in control schools spent 1% of their time working with computers, is not surprising, given that individualized software instruction is another key component of BTL. Computer-based activities in BTL classrooms are all interactive and include electronic book reading (designed to simulate lap reading); as well as activities focused on letter and word knowledge, phonological sensitivity, vocabulary and concept knowledge, and auditory

comprehension skills.

Finally, effect sizes of impacts on three OMLIT measures (print knowledge, ELL students, and literacy resources) are all larger than 0.2, which could be considered substantially meaningful, even though these impacts are not statistically significant.

Teachers after two years of implementing BTL.

As seen in Table 4b, nearly all of the significant impacts shown at the end of one year of BTL vanished at the end of the second year. The only remaining statistically significant impact is the proportion of time spent in computer activities (ES = 4.588, p less than or equal to 0.0001), which is dramatic but not surprising given that many control group classrooms did not have or use computers in the classroom.

Another impact is large and nearly statistically significant—the approach to working with ELL students. The effect is more than three-quarters of a standard deviation in favor of the control group. This is a troubling trend, but we do not have a hypothesis regarding why implementation of BTL might have affected the classrooms in this way or what else was going on in the district that might have had a positive impact on the way teachers in the control group work with ELLs.

BTL classrooms may also have spent more time in literacy activities (ES = 0.534), but the impact is not statistically significant ($p = .062$).

One reason for the failure to reach statistical significance is that the year-two teacher group is so much smaller—much smaller than the study design called for (thus, the second year analysis is under-powered to detect moderately-sized effects). The one-year group included 131 classrooms, while the two-year group included only 80, so the standard errors are much larger in the two-year analysis, and several sizeable impacts do not attain statistical significance.

Impacts on Arnett scores.

Impact estimates on Arnett measures are presented in Tables 5a and 5b. Although none of these estimates are statistically significant, effect sizes of two measures (positive and permissive) in the analysis of classrooms of one-year teachers and of three measures (positive, detachment, and the Arnett standardized composite) in the analysis of two-year teachers are larger than 0.25. We do not have a hypothesis about what the effect of these differences might be on the implementation of the curriculum or on classroom outcomes nor of why implementation of the curriculum might affect scores on the Arnett subscales.

Student Measures

Overall Impacts

Method

Table 6 presents the impact of BTL on student test scores; first separately for each cohort and then across two cohorts. The estimates reported here are estimated through a three-level HLM model that represents the nested structure of the data (students nested within classrooms, which are nested within schools). This model employs a-priori selected covariates (sample stratification indicators, cohort indicator, school-level pretest, and percentage of minority students at the school level). For the models used for each outcome, see Appendix C. A number of other covariates (gender, age, ELL status, and percentage of ELL students at the classroom level) were tested for inclusion using the backwards elimination technique.

Results and Discussion

As seen in the first panel of Table 6, overall there were no statistically significant differences between the BTL group and control group on any of the four tests of early literacy skills at the end of kindergarten across both cohorts. Analyses by cohort, presented in the second and third panels, indicate the same pattern as the overall sample, with no significant differences across BTL and control students for either cohort. These results provide support for the pooling of the two cohorts in analyses. It is also worth noting that not only is there no statistically significant impact of BTL, but there is also no impact of any substantial magnitude that would indicate a trend towards an impact of the curriculum.

We are considering further analyses to explain why the curriculum might not have produced an impact. Our hypotheses fall into three main types: insufficient implementation in the treatment classrooms, high quality instruction in the counterfactual, and insufficient time to make a substantial impact on outcomes.

Insufficient implementation.

The first group includes the fact that teachers did not receive as much support from coaches as the developers suggested. In other instances of implementation of this curriculum, coaches visited teachers every two weeks or no less than monthly, whereas in this implementation, some teachers received five coaching visits during the year, others received fewer. A second implementation problem was that the interactive software component of the curriculum was unevenly implemented; some teachers did it very well, others did not adjust the settings of the computers (rendering them ineffective - students would continue endlessly at the same level), and that technical difficulties with computers prevented teachers from using them to the fullest extent possible. A third problem with the implementation is that the core books for BTL are intended to be used as just one book to be used in a unit, and the developers expected that teachers would supplement these books with other trade books. To the extent that this was not done consistently, an extremely limited vocabulary would have been presented to the children. This would explain why, even with more time spent on book-related activities and discussions, if those activities are based on reduced-language books, then the impact on children's vocabulary development will be smaller.

High quality instruction in control classrooms.

Our observations do not lend support to this hypothesis. We observed teachers in the treatment group classrooms spending

more time doing activities that should have led to better early literacy outcomes for children.

Insufficient time spent on high-quality instruction to make an impact.

Finally, it is possible that although the difference in key instructional behaviors was statistically significant, it was still not substantial enough to produce a statistically significant impact in children's outcomes. It is possible that if the teachers in the treatment classrooms spend 6% more time on an activity, for example, than teachers in the control classrooms, and if the time observed is the only time in the day that those activities occur (i.e., if what we observed represented 100% of the literacy instruction in the day), then 6% would be 9 minutes. It is possible that 9 additional minutes each day may not be sufficient to produce a statistically significant impact.

Differences in impact by language of child

Table 7 presents the results of subgroup analyses based on ELL status. Once the sample is divided into ELL and non-ELL students, the number of schools drops for the ELL group from 43 to 29 because there are 14 schools in the sample that have no ELL students. No statistically significant differences are found between BTL and control students in the ELL or non-ELL subgroup.

Exposure

In Table 8, we present test score comparisons of BTL students who were exposed to the program one, two, and three years. It is very important to keep in mind that these analyses deviate from the experimental setting as these groups were not determined randomly. In addition, since we do not have enrollment histories on either group of students (BTL or control), our findings are limited by the lack of information about students' school experiences prior to being in our sample. These results, therefore, should be interpreted with caution.

For each outcome measure (Expressive Vocabulary, Word Attack, Word Identification, Letter Identification), we conducted three initial comparisons:

- 1 year BTL exposure versus 1 year control (n = 2,340)
- 2 years BTL exposure versus 2 years control (n = 710), and
- 3 years BTL exposure versus 3 years control (n = 57).

In addition, because the 3 year exposure sample is extremely small, we also combined the 2 and 3 year exposure groups and compared their outcomes.

As shown in Table 8, there were no statistically significant differences between BTL and control students with 1 or 2 years of exposure. For those students with 3 years of exposure, there was a statistically significant difference between BTL and control students in Expressive Vocabulary and Letter ID (7.934 and 6.792 points, respectively) in favor of BTL students. However, once we combined the 2- and 3-year exposure groups, these differences were no longer statistically significant.





**U.S. Department of Education
Grant Performance Report (ED 524B)
Project Status Chart**

PR/Award #: **R305G040145**

SECTION A - Project Objectives Information and Related Performance Measures Data (See Instructions. Use as many pages as necessary.)

2 . Project Objective Check if this is a status update for the previous budget period.
 Test impacts of BTL by examining differences in performance on three language outcomes (expressive vocabulary, decoding, and word reading), at the end of Kindergarten, Grade 1, and Grade 2

. Performance Measure	Measure Type	Quantitative Data					
		Target			Actual Performance Data		
		Raw Number	Ratio	%	Raw Number	Ratio	%
	PROJ		/			/	

Explanation of Progress (Include Qualitative Data and Data Collection Information)

Summary
 In this analysis, we tested impacts of BTL by examining differences in performance on three language outcomes (expressive vocabulary, decoding, and word reading), at the end of Kindergarten, Grade 1, and Grade 2. There were no statistically significant differences between Treatment and Control at the end of Kindergarten, first, or second grade for either cohort. We also tested impacts on the growth in children's abilities from Kindergarten to Grades 1 and 2. Compared with children in BTL classrooms, children in Control classrooms, on average, showed significantly more growth from Kindergarten to Grade 1 in both decoding and word reading, although the difference in growth was small. The amount of growth between Grade 1 and Grade 2 could only be tested for the first cohort of children in the sample, and there was no significant difference between the growth of children in BTL and Control group classrooms. Once data become available, we will test whether there is any impact of the treatment on children's growth from Grade 1 to Grade 2 for cohort 2 and for the two cohorts combined.

We examined impacts as a function of English Language Learner (ELL) status by conducting separate, parallel analyses for

ELL and non-ELL subgroups on the same outcomes. For the ELL children, there was a statistically significant impact favoring BTL over Control on one outcome (decoding) at the end of Kindergarten, although this difference did not persist to the end of Grade 1 (or Grade 2). For the non-ELL subgroup, there was no impact on decoding at the end of Kindergarten, Grade 1, or Grade 2. There were no impacts on expressive vocabulary or word reading at the end of Kindergarten, first, or second grade for either the ELL or non-ELL subgroup.

Note that because we tested so many hypotheses in this analysis, there is a strong possibility that the few impacts detected likely occurred by chance rather than as a result of the intervention.

Measuring Changes in Impacts on Students Over Time: By Cohort and by ELL Subgroup

The overarching research question for these analyses is as follows: Is there any long-term effect of BTL (exposure) in Kindergarten on student language and literacy outcomes at the end of first or second grade? To understand this, we examine student growth on language and literacy outcomes from Kindergarten through first and second grade. The data for this analysis are standardized test scores through the end of second grade for the first cohort of Kindergarten students in our sample and through the end of first grade for the second cohort of Kindergarten students. In this analysis, therefore, we can examine change for cohort 1 from the spring of Kindergarten to the spring of second grade and from the spring of Kindergarten to the spring of first grade for cohort 2.

A second research question asks about differential impacts for ELL versus non-ELL children.

As shown in Table 9, the total sample consists of 7,382 students (4,557 ELL students and 2,825 non-ELL students across all three grades) in 43 schools (22 assigned to the Treatment condition, and 21 assigned to the Control condition). A detailed breakdown by cohort, grade, ELL status, and treatment status is presented.

The analysis is based on a difference-in-differences approach, in which we test to see whether there are statistically significant differences between BTL and Control students at each time point (spring of Kindergarten, spring of Grade 1, and spring of Grade 2) and then test whether the differences at each time point are different across time points (and whether these differences-in-differences are statistically significant). All analyses control for Kindergarten pre-test scores at the school level. Separate models are run for three outcomes: Expressive One-Word Picture Vocabulary Test (EOWPVT), WRMT: Word Identification (Word ID), WRMT: Word Attack. Scores are Normal Curve Equivalents for EOWPVT, and W-scores for Word ID and Word Attack.

We use a hierarchical model to conduct these analyses because our data are nested. In this case, the data have three levels of nesting: time is nested within individual students who are, in turn, nested within schools. See Appendix D for the three-level hierarchical linear growth model that is used in analyses.

Because we have different numbers of test points for the two cohorts of students, the results of these analyses are reported separately by cohort. For completeness, we present results from analyses that estimate pooled BTL-Control group differences across the two cohorts for the sample overall and for ELL and non-ELL subgroups in Appendix E. In general, these pooled cohort differences display patterns very similar to those of the by-cohort differences presented below. Once we have data at the end of second grade for cohort 2, we will test for differences across the cohorts from Kindergarten to Grade 2. If none are found, we will combine the cohorts for all analyses. If there are differences, we will continue to present the cohort results separately.

Results

Descriptive statistics for the sample are presented by measure in Table 10. Findings for each outcome are presented by cohort and then by ELL subgroup by cohort. Findings are presented graphically in the figures; estimates used to create these graphs are presented in Tables 3-5.

Trends in the estimated effect of BTL exposure on each outcome, by cohort and for ELL and non-ELL subgroups by cohort, are summarized below.

Caveat on Interpretation

When interpreting these results, it is important to keep the "multiple comparisons" or "multiple hypothesis testing" issue in mind. Multiple hypothesis testing is a problem because as the number of tests conducted increase, the probability of making a Type I error (i.e. finding a difference when in fact there is none) increases. More specifically, when 20 hypothesis tests are performed in a setting where there are no true differences between two conditions at the usual $p < 0.05$ significance level, we expect to have one false significant difference (i.e., one difference by chance). Note that we conducted 81 tests when examining the BTL-control differences (excluding the exploratory analyses), and one would expect to observe 4 significant differences (two favoring the BTL group and two favoring the control group) by chance alone (when there were no "true differences"). Hence, we suggest that the three statistically significant findings presented below should be interpreted with caution as they may well just be due to chance.

Expressive Vocabulary

At the three time points indicated by the deltas in the graphs (Kindergarten, Grade 1, and Grade 2), for cohort 1, the difference between the average expressive vocabulary (EOWPVT) test scores of BTL and Control students is 0.7, 0.8, and -0.3 points in Kindergarten, first, and second grade respectively (Figure 1A). None of these differences is statistically significant at the p less than or equal to 0.05 level. When we tested for "differences in differences," we found that there were no statistically significant differences in the amount of growth for BTL versus Control between Kindergarten and Grade 1, Grade 1 and Grade 2, or between Kindergarten and Grade 2. The same results hold for cohort 2 (Figure 1B), with no statistically significant differences between BTL and Control students at either Kindergarten or first grade and no statistically significant difference between the BTL versus Control differences between Kindergarten and Grade 1. For both cohorts, BTL and Control groups are performing below the national average of 50 NCE at all time points.

For the comparison between the ELL and non-ELL subgroups by cohort (see Figures 1C and 1D), there were no statistically significant differences found between BTL and Control groups for either ELL or non-ELL students at the end of K, Grade 1 or Grade 2 for Expressive Vocabulary. There were also no statistically significant "differences in differences" between grades within either subgroup for either cohort. For both cohorts, BTL and Control groups within ELL and non-ELL subgroups are performing below the national average of 50 NCE at all time points.

Decoding

As indicated in Figure 2A, the difference in the average word attack test scores (WRMT-R: Word Attack) of BTL and Control students is 2.1, -0.02, and 0.3 points in Kindergarten, first, and second grade respectively. None of these differences is statistically significant at the p less than or equal to 0.05 level. When we tested for "differences in differences," we found that there were no statistically significant differences between the BTL versus Control differences from Kindergarten to Grade 1, from Grade 1 to Grade 2, or from Kindergarten to Grade 2. Both BTL and Control groups are performing above the national norm at the end of Kindergarten and Grade 1, but then fall below the norm by the end of Grade 2. For cohort 2 (Figure 2B), BTL students outperformed Control students in Kindergarten by 1.9 W-score points, while Control students outscored BTL students in first grade by 1.5 W-score points. However, the difference between BTL and Control students is not statistically significant at either Kindergarten or first grade. On the other hand, we do find a statistically significant "difference in differences." That is, the difference in the BTL versus Control differences between Kindergarten and Grade 1 is 3.4 W-score points (effect size = 0.21) and it is statistically significant at the $p = 0.03$ level, most likely because there was a difference in Kindergarten in favor of the BTL group (a positive number) that then became a difference in first grade in favor of the Control group (a negative number), in other words, the Treatment group not only failed to maintain their lead from Kindergarten but actually fell behind the Control group at Grade 1, so their relative growth rate was negative. Results from the assessments to be conducted in Spring 2009 will reveal whether the pattern of cohort 1 is repeated in cohort 2. Both BTL and Control groups in cohort 2 were performing above the national norm at the end of Kindergarten and Grade 1.

Figures 2C and 2D present results for ELL and non-ELL subgroups for cohorts 1 and 2. For cohort 1, we find the same results as above, i.e., no statistically significant differences between BTL and Control groups within ELL or non-ELL subgroups at the end of K, Grade 1 or Grade 2 and no statistically significant "difference in differences" between grades. Both BTL and Control groups within each subgroup are performing above the national norm at the end of Kindergarten and Grade 1, but then fall below the norm by the end of Grade 2. For cohort 2, we find a slightly different pattern from that described above, i.e., there is a statistically significant difference in favor of the Treatment group at the end of Kindergarten within the ELL subgroup (4.5 W score points; effect size = 0.31; $p < 0.05$), although this difference does not persist until the end of Grade 1. For non-ELLs, there are no statistically significant differences between BTL and Control groups at the end of Kindergarten or first grade. There is not a statistically significant "difference in differences" between Kindergarten and Grade 1 for either ELL or non-ELL subgroups. Both BTL and Control groups within ELL and non-ELL subgroups are performing above the national norm at the end of Kindergarten and Grade 1.

Word Reading

As indicated in Figure 3A, the differences in the average word identification test scores (WRMT-R: Word Identification) of BTL and Control students are 2.8, -3.4, and -0.4 points in Kindergarten, first, and second grade respectively. None of these differences is statistically significant at the p less than or equal to 0.05 level. When we tested for "differences in differences," we found that there was a statistically significant difference in differences between BTL versus Control scores from Kindergarten to Grade 1 (6.2 W-score points, effect size = 0.20, $p = 0.02$). Once again, this is likely because although BTL students outperformed Control students in Kindergarten, Control students outperformed BTL students in first grade. By the end of Grade 2, BTL students had almost caught up to Control students. Both BTL and Control groups are performing above the national norm at the end of Kindergarten and Grade 1, but then fall below the norm by the end of Grade 2. For cohort 2 (Figure 3B), there were no statistically significant differences between BTL and Control students at either Kindergarten or first grade and no statistically significant difference between the BTL versus Control differences between Kindergarten and Grade 1. Both BTL and Control groups are performing above the national norm at the end of Kindergarten and Grade 1.

Similar results were obtained for tests of impacts of BTL for the ELL and non-ELL subgroups by cohort (see Figures 3C and 3D). In both cohorts, no statistically significant differences were found between BTL and Control groups within ELL or non-ELL subgroups at the end of K, Grade 1 or Grade 2 for word reading. There were also no statistically significant "differences in differences" between grades within either subgroup for either cohort. For cohort 1, both BTL and Control groups within ELL and non-ELL subgroups are performing above the national norm at the end of Kindergarten and Grade 1, but then fall below the national norm by the end of Grade 2. For cohort 2, both BTL and Control groups within ELL and non-ELL subgroups are performing above the national norm at the end of Kindergarten and Grade 1.

Exploratory Analyses of ELL versus Non-ELL Patterns of Change

An examination of Figures 2C-2D and 3C-3D indicates that, for two of the outcomes - Word Attack and Word ID - patterns of change for each outcome are similar for the ELL and non-ELL subgroups and across BTL and Control groups within each cohort. With or without BTL, ELLs and non-ELLs in our sample are performing similarly on these two outcomes. This is an interesting finding in and of itself. In addition, at the last data collection point, the difference between study children's scores and the national norm is not statistically significant on both outcomes, irrespective of treatment or ELL status. (The last data collection point is Grade 2 for cohort 1 and Grade 1 for cohort 2.)

When we tested Word Attack outcomes for cohort 1, we found that although there were statistically significant differences between ELL and non-ELL children in the BTL group and statistically significant differences between ELL and non-ELL children in the Control group at the end of Kindergarten, these differences no longer existed at the end of Grade 1 or Grade 2. This means that although ELL children in both BTL and Control groups in cohort 1 are starting out significantly lower than the non-ELL children in their same treatment group in Kindergarten, they catch up with the non-ELL children by the end of Grade 1 and perform similarly to the non-ELL children through the end of Grade 2. For cohort 2, there are no statistically significant differences between ELL and non-ELL children in the BTL group or between ELL and non-ELL

children in the Control group at the end of Kindergarten or Grade 1.

For Word ID, we found a similar pattern for cohort 1, with statistically significant differences between ELL and non-ELL children within both the BTL and Control groups at the end of Kindergarten, but no statistically significant differences at the end of Grade 1 or Grade 2. For cohort 2, we found a statistically significant difference at the end of Kindergarten between ELL and non-ELL children in the Control group, but not in the BTL group. As with Word Attack, however, this difference is no longer statistically significant at the end of Grade 1 within either group (BTL or Control).

Interestingly, a very different pattern is evident when we examine the expressive vocabulary outcome (Figures 1C and 1D). Especially in cohort 1, it appears that ELLs and non-ELLs, independent of their treatment status, are performing differently. While both ELL and non-ELL subgroups are performing below the national norm (NCE = 50), the non-ELL subgroup appears to be approaching the norm from Kindergarten to Grade 2, while the ELL subgroup's performance remains flat, even losing a little ground from Kindergarten to Grade 2. As a result, the gap between the two groups widens over time. For cohort 2, the pattern is similar but there is less of a gap between the ELL and non-ELL children.

For cohort 1, we found statistically significant differences between ELL and non-ELL children in the BTL group and statistically significant differences between ELL and non-ELL children in the Control group at the end of Kindergarten, Grade 1 and Grade 2. This means that ELL children in both BTL and Control groups are starting out significantly lower than the non-ELL children in their same treatment group in Kindergarten and remain at a significantly lower level through Grade 2. For cohort 2, the same pattern held at the end of Kindergarten and Grade 1. Spring 2009 data from cohort 2 will show whether cohort 2 data continue to follow the same pattern as cohort 1 data.

In general, these findings indicate a shortfall in ELL children's vocabulary knowledge compared with non-ELL children's vocabulary knowledge (whether in BTL or Control), whereas ELL and non-ELL children differed only in Kindergarten, if at all, in their performance on tests of decoding and word reading, with ELL children catching up to non-ELL children in Grade 1 (cohorts 1 and 2) and Grade 2 (cohort 1). In addition, on tests of decoding and word reading, both ELL and non-ELL children's scores were close to the grade level norm, while in vocabulary, both ELL and non-ELL children performed statistically significantly below the national norm, with only non-ELL children making some progress toward it by the end of Grade 2.



**U.S. Department of Education
Grant Performance Report (ED 524B)
Project Status Chart**

PR/Award #: **R305G040145**

SECTION A - Project Objectives Information and Related Performance Measures Data (See Instructions. Use as many pages as necessary.)

3 . Project Objective Check if this is a status update for the previous budget period.
Implement an analysis of the writing samples from BTL and Non-BTL classrooms (Grant Supplement Award)

. Performance Measure	Measure Type	Quantitative Data					
		Target			Actual Performance Data		
		Raw Number	Ratio	%	Raw Number	Ratio	%
	PROJ		/			/	

Explanation of Progress (Include Qualitative Data and Data Collection Information)

ClimbWrite: An analysis of elicited writing in young children.

The Breakthrough to Literacy (BTL) curriculum that is being assessed in the large randomized cluster study in the Chicago public schools includes writing as an essential classroom practice. When implemented with fidelity, the BTL curriculum integrates listening, reading, writing and speaking, around a focus book in whole group, small group and individual instruction. Repeated opportunities for exposure to language and cognitive activities within the context of familiar daily practices set the stage for building both the deep and surface structures of language.

The Chicago Study (CLIMBERS) offers us the opportunity to examine a large sample of writing at the Pre-K and Kindergarten levels to more precisely examine the relationship of early reading and writing. At the time we submitted the original grant we did not have a tool to systematically examine the writing of children in Pre-K and Kindergarten, and so did not include an analysis of children's writing and its relationship to other emerging literacy skills. We are now in a position to be able to also systematically examine the impact of the BTL curriculum on the emergent writing skills of children. The supplement has allowed us to more fully assess the impact of BTL as it is implemented on a large scale in Chicago by

allowing us to look at both reading and writing.

Utilizing the Picture prompts developed for this project we collected a pilot sample of writing samples during the spring of 2008. Teachers asked kindergarten students to complete a writing task using picture prompts and standard instructions. We examined the writing samples that kindergarteners in the Chicago and West Des Moines public schools produced in response to this task. Writing samples from Chicago were collected from classrooms identified by Abt as classrooms (BTL and non-BTL classrooms) in which teacher were observed engaging the children in significant writing activities (based on OMLIT: Quill observation scores). The West Des Moines samples came from classrooms that were using the BTL curriculum. Each sample was scanned into a specialized database developed at the University of Iowa. The samples were then coded utilizing a computer assisted coding scheme.

Research Questions

In this preliminary analysis, we examined the children's overall productivity and spelling strategies to address two research questions:

1. How do students vary in the length of written text, prevalence of incorrectly spelled words, and spelling strategies at the end of kindergarten?
2. How does students' writing change during the second half of the kindergarten year, in terms of the length of written text, prevalence of incorrectly spelled words, and spelling strategies?

RQ1. Writing and Spelling at the End of Kindergarten

We examined descriptive statistics and bivariate correlations for total word count, incorrect spelling, and spelling strategies in kindergarten writing collected at the end of the kindergarten year. The sample included 165 students in 9 classrooms in Chicago and West Des Moines public schools. We estimated a two-level model to test for differences in kindergarten writing outcomes between students and classrooms. We included a school district indicator variable in the Level-2 model to test for school district differences in writing outcomes. For each writing outcome, we estimated the model as described in the Narrative section.

Findings indicated that students from the West Des Moines sample differed substantially from students from the Chicago sample (See Table 14). The average writing sample length among Chicago students was 11.7 words, while the average length among West Des Moines students was 30.1 words ($p < .0001$). Approximately 11% of the variation in the length of students' writing was between classrooms ($p < .10$), indicating that children demonstrated more productivity in their writing in some Chicago classrooms than in other Chicago classrooms, and children's writing was more productive in some West Des Moines classrooms than in others. By writing longer text, students in West Des Moines used more invented spelling (average of 11.7 incorrectly spelled words in West Des Moines) than in Chicago (average of 4.1 incorrectly spelled words; $p < .0001$), which provided them more opportunity to employ spelling strategies reflecting correct consonant and vowel sounds. Only 5% of the students in Chicago produced written text of 21 or more words and a high incidence of invented spelling; however, 53% of students in the West Des Moines sample produced this type of written text. The site and

classroom variations in students' writing can be attributed to variations in the level of implementation of writing opportunities in the site and classroom curriculum.

RQ2. Growth in Writing and Spelling Strategies during Kindergarten

Students completed the writing task at three time points (February, April, and May) in the West Des Moines sample (71 students in 3 classrooms). We estimated the following linear growth model to investigate change in students' writing productivity, incorrect spelling, and spelling strategies over the second half of kindergarten. This model is fully described in the Narrative section.

Findings indicated that students' writing productivity and incidence of invented spelling increased over the second half of the kindergarten year (See Table 15). In February, the average length of written text was 16 to 18 words in two of the West Des Moines classrooms and was 27 words in a third classroom. Students' writing increased in length by an average of 4.7 words at each subsequent time point that writing was elicited ($p < .001$). In writing longer texts, students had greater opportunity for employing their invented spelling strategies; the number of incorrectly spelled words in students' writing increased by an average of 1.4 words at each time point ($p < .001$). In particular, students employed advanced spelling strategies (i.e., correct consonant and vowel sounds in misspelled multisyllabic words) with an increasing proportion of incorrectly spelled words over the second half of kindergarten. On average, in February, students employed advanced spelling strategies in 15% of their incorrectly spelled words, and this percentage increased by an average of 5 percentage points at each subsequent time point ($p < .001$). By the end of kindergarten, students were using advanced spontaneous spelling strategies in 25% of their incorrectly spelled words, suggesting that students were using more multi-syllabic words and more advanced spelling strategies over the second half of kindergarten.

A complete set of writing samples were obtained at three time points during the 2008-2009 academic year (Fall, Winter, Spring) from the kindergarten classrooms in a sub-sample of the Chicago public schools in order to apply the analytic techniques we developed to look at growth.



**U.S. Department of Education
Grant Performance Report (ED 524B)
Project Status Chart**

PR/Award #: **R305G040145**

SECTION B - Budget Information (See Instructions. Use as many pages as necessary.)

Title : hurtig-sectionb

File : H:\HURTIG\CLIMBERS\year5progressreport\hurtig-sectionb.doc

SECTION C - Additional Information (See Instructions. Use as many pages as necessary.)

Title : hurtig-IRB

File : M:\hurtiglab\climbers 2009 annual report\climbers- IRB- 08-09 approvals\hurtig-irb.pdf



U.S. Department of Education
Grant Performance Report (ED 524B)
Project Status Chart

OMB No. 1890 - 0004
Expiration: 10-31-2007

PR/Award
#:R305G040145

SECTION B - Budget Information *(See Instructions. Use as many pages as necessary.)*

Project Period: 06/01/2008 – 05/31/2009

Summary of Year 5 Expenses

Project personnel / fringe	213,517
Supplies	3,482
Travel	5,656
Other / miscellaneous	1,365
Other / tuition fees for grad students	16,558
Subcontract to Abt Associates	291,017
F & A costs	58,245
TOTAL EXPENDED YEAR 5	589,840



Abt Associates Inc.

Institutional Review Board Notice of Approval

Principal Investigator/Project Director: Carolyn Layzer

Project Title: CLIMBERS

Sponsor Agency: ED/IES

Abt IRB #: 0036

Protocol Approval Date: June 16, 2009

Review Type: Expedited

Type of Approval: Continuing Review

Please note the following requirements:

Problems or adverse reactions: If any problems in treatment of human subjects or unexpected adverse reactions occur as a result of this study, you must notify the IRB Chairperson or IRB Administrator immediately.

Consent forms: In the event the approved study includes procedures for written informed consent, you only may use consent forms that bear the Abt Associates Inc. IRB approval stamp.

Changes in protocol, study design, or study materials: If there are changes in procedures, the study design, or study materials (e.g., survey instruments, consent forms), you must submit these materials for IRB review and approval before they are implemented.

Renewal: You are required to apply for renewal of approval at least annually for as long as the study is active. Your next review date should be on or before **June 15, 2010**.

IRB Administrator
Tammy Kolbe

Date
June 11, 2009

Cc:



Human Subjects Office

340 Medicine Administration Building
Iowa City, Iowa 52242-1101
319-335-6564 Fax 319-335-7310
irb@uiowa.edu
<http://research.uiowa.edu/hso>

IRB ID #: 200611707

To: Richard Hurtig

From: IRB-01 DHHS Registration # IRB00000099,
Univ of Iowa, DHHS Federalwide Assurance # FWA00003007

Re: Analysis of children's writing samples.

Protocol Number:

Protocol Version:

Protocol Date:

Amendment Number/Date(s):

Approval Date: 10/05/08

Next IRB Approval Due Before: 10/05/09

Type of Application:

- New Project
- Continuing Review
- Modification

Type of Application Review:

- Full Board:
- Meeting Date:
- Expedited
- Exempt

Approved for Populations:

- Children
- Prisoners
- Pregnant Women, Fetuses, Neonates

Source of Support: US Department of Education

Investigational New Drug/Biologic Name:

Investigational New Drug/Biologic Number:

Name of Sponsor who holds IND:

Investigational Device Name:

Investigational Device Number:

Sponsor who holds IDE:

This approval has been electronically signed by IRB Chair:
Martha Jones, CIP, MA
10/05/08 2128

OFFICE OF THE VICE PRESIDENT
FOR RESEARCH

IRB Approval: IRB approval indicates that this project meets the regulatory requirements for the protection of human subjects. IRB approval does not absolve the principal investigator from complying with other institutional, collegiate, or departmental policies or procedures.

Agency Notification: If this is a New Project or Continuing Review application and the project is funded by an external government or non-profit agency, the original HHS 310 form, "Protection of Human Subjects Assurance Identification/IRB Certification/Declaration of Exemption," has been forwarded to the UI Division of Sponsored Programs, 100 Gilmore Hall, for appropriate action. You will receive a signed copy from Sponsored Programs.

Recruitment/Consent: Your IRB application has been approved for recruitment of subjects not to exceed the number indicated on your application form. If you are using written informed consent, the IRB-approved and stamped Informed Consent Document(s) are attached. Please make copies from the attached "masters" for subjects to sign when agreeing to participate. The original signed Informed Consent Document should be placed in your research files. A copy of the Informed Consent Document should be given to the subject. (A copy of the *signed* Informed Consent Document should be given to the subject if your Consent contains a HIPAA authorization section.) If hospital/clinic patients are being enrolled, a copy of the signed Informed Consent Document should be placed in the subject's chart, unless a Record of Consent form was approved by the IRB.

Continuing Review: Federal regulations require that the IRB re-approve research projects at intervals appropriate to the degree of risk, but no less than once per year. This process is called "continuing review." Continuing review for non-exempt research is required to occur as long as the research remains active for long-term follow-up of research subjects, even when the research is permanently closed to enrollment of new subjects and all subjects have completed all research-related interventions and to occur when the remaining research activities are limited to collection of private identifiable information. Your project "expires" at 12:01 AM on the date indicated on the preceding page ("Next IRB Approval Due on or Before"). You must obtain your next IRB approval of this project on or before that expiration date. You are responsible for submitting a Continuing Review application in sufficient time for approval before the expiration date, however the HSO will send a reminder notice approximately 60 and 30 days prior to the expiration date.

Modifications: Any change in this research project or materials must be submitted on a Modification application to the IRB for prior review and approval, except when a change is necessary to eliminate apparent immediate hazards to subjects. The investigator is required to promptly notify the IRB of any changes made without IRB approval to eliminate apparent immediate hazards to subjects using the Modification/Update Form. Modifications requiring the prior review and approval of the IRB include but are not limited to: changing the protocol or study procedures, changing investigators or funding sources, changing the Informed Consent Document, increasing the anticipated total number of subjects from what was originally approved, or adding any new materials (e.g., letters to subjects, ads, questionnaires).

Unanticipated Problems Involving Risks: You must promptly report to the IRB any serious and/or unexpected adverse experience, as defined in the UI Investigator's Guide, and any other unanticipated problems involving risks to subjects or others. The Reportable Events Form (REF) should be used for reporting to the IRB.

Audits/Record-Keeping: Your research records may be audited at any time during or after the implementation of your project. Federal and University policies require that all research records be maintained for a period of three (3) years following the close of the research project. For research that involves drugs or devices seeking FDA approval, the research records must be kept for a period of three years after the FDA has taken final action on the marketing application.

Additional Information: Complete information regarding research involving human subjects at The University of Iowa is available in the "Investigator's Guide to Human Subjects Research." Research investigators are expected to comply with these policies and procedures, and to be familiar with the University's Federalwide Assurance, the Belmont Report, 45CFR46, and other applicable regulations prior to conducting the research. These documents and IRB application and related forms are available on the Human Subjects Office website or are available by calling 335-6564.



Human Subjects Office

340 Medicine Administration Building
Iowa City, Iowa 52242-1101
319-335-6564 Fax 319-335-7310
irb@uiowa.edu
<http://research.uiowa.edu/hso>

IRB ID #: 200712719
To: Richard Hurtig
From: IRB-02 DHHS Registration # IRB00000100,
Univ of Iowa, DHHS Federalwide Assurance # FWA00003007
Re: Analysis of Children's Writing Elicited with Picture Prompts

Approval Date: 10/31/08

Next IRB Approval Due Before: 10/31/09

Type of Application:

- New Project
- Continuing Review
- Modification

Type of Application Review:

- Full Board:
Meeting Date:
- Expedited

- Exempt

Approved for Populations:

- Children
- Prisoners
- Pregnant Women, Fetuses, Neonates

Source of Support: US Department of Education

This approval has been electronically signed by IRB Chair:
Jerry Suls, PHD
10/31/08 1135

OFFICE OF THE VICE PRESIDENT
FOR RESEARCH

IRB Approval: IRB approval indicates that this project meets the regulatory requirements for the protection of human subjects. IRB approval does not absolve the principal investigator from complying with other institutional, collegiate, or departmental policies or procedures.

Agency Notification: If this is a New Project or Continuing Review application and the project is funded by an external government or non-profit agency, the original HHS 310 form, "Protection of Human Subjects Assurance Identification/IRB Certification/Declaration of Exemption," has been forwarded to the UI Division of Sponsored Programs, 100 Gilmore Hall, for appropriate action. You will receive a signed copy from Sponsored Programs.

Recruitment/Consent: Your IRB application has been approved for recruitment of subjects not to exceed the number indicated on your application form. If you are using written informed consent, the IRB-approved and stamped Informed Consent Document(s) are attached. Please make copies from the attached "masters" for subjects to sign when agreeing to participate. The original signed Informed Consent Document should be placed in your research files. A copy of the Informed Consent Document should be given to the subject. (A copy of the *signed* Informed Consent Document should be given to the subject if your Consent contains a HIPAA authorization section.) If hospital/clinic patients are being enrolled, a copy of the signed Informed Consent Document should be placed in the subject's chart, unless a Record of Consent form was approved by the IRB.

Continuing Review: Federal regulations require that the IRB re-approve research projects at intervals appropriate to the degree of risk, but no less than once per year. This process is called "continuing review." Continuing review for non-exempt research is required to occur as long as the research remains active for long-term follow-up of research subjects, even when the research is permanently closed to enrollment of new subjects and all subjects have completed all research-related interventions and to occur when the remaining research activities are limited to collection of private identifiable information. Your project "expires" at 12:01 AM on the date indicated on the preceding page ("Next IRB Approval Due on or Before"). You must obtain your next IRB approval of this project on or before that expiration date. You are responsible for submitting a Continuing Review application in sufficient time for approval before the expiration date, however the HSO will send a reminder notice approximately 60 and 30 days prior to the expiration date.

Modifications: Any change in this research project or materials must be submitted on a Modification application to the IRB for prior review and approval, except when a change is necessary to eliminate apparent immediate hazards to subjects. The investigator is required to promptly notify the IRB of any changes made without IRB approval to eliminate apparent immediate hazards to subjects using the Modification/Update Form. Modifications requiring the prior review and approval of the IRB include but are not limited to: changing the protocol or study procedures, changing investigators or funding sources, changing the Informed Consent Document, increasing the anticipated total number of subjects from what was originally approved, or adding any new materials (e.g., letters to subjects, ads, questionnaires).

Unanticipated Problems Involving Risks: You must promptly report to the IRB any serious and/or unexpected adverse experience, as defined in the UI Investigator's Guide, and any other unanticipated problems involving risks to subjects or others. The Reportable Events Form (REF) should be used for reporting to the IRB.

Audits/Record-Keeping: Your research records may be audited at any time during or after the implementation of your project. Federal and University policies require that all research records be maintained for a period of three (3) years following the close of the research project. For research that involves drugs or devices seeking FDA approval, the research records must be kept for a period of three years after the FDA has taken final action on the marketing application.

Additional Information: Complete information regarding research involving human subjects at The University of Iowa is available in the "Investigator's Guide to Human Subjects Research." Research investigators are expected to comply with these policies and procedures, and to be familiar with the University's Federalwide Assurance, the Belmont Report, 45CFR46, and other applicable regulations prior to conducting the research. These documents and IRB application and related forms are available on the Human Subjects Office website or are available by calling 335-6564.



CHICAGO PUBLIC SCHOOLS

Office of Research, Evaluation, and Accountability
125 South Clark Street, 11th floor Chicago, Illinois 60603
Telephone: 773/553-2320
Fax: 773/553-2436

March 05, 2009

Carolyn Layzer
Abt Associates Inc.
55 Wheeler St
Cambridge, MA 02138

Dear Ms. Layzer:

Thank you for your interest in conducting research in The Chicago Public Schools. The Research Review Board of the Office of Research, Evaluation, and Accountability has reviewed your proposal for research entitled Breakthrough to Literacy in Chicago Public Schools and has approved your request to conduct research. Although your study has been approved, school principals have final authority over activities that are allowed to take place in the school. If data collection continues beyond a year from this approval, please complete the Modification & Continuing Review Process Checklist.

Upon completion of the research study, a copy of the final report or summary of the results must be provided to the Research Review Board. The Board reserves the right to use the information in the research report or summary for planning, solicitation of grants and staff development.

Please note that your study has been assigned Project ID #189. If you have any questions, please contact Michelle Acker on my staff at 773-553-2452. If you need additional clarification after contacting Ms. Acker feel free to contact me at 773-553-2497.

Sincerely,

A handwritten signature in black ink that reads "Bret Feranchak". The signature is written in a cursive, flowing style.

Bret Feranchak
Director of Program Evaluation and Applied Research
Chair, Research Review Board
Office of Research, Evaluation, and Accountability

- a. Are you claiming indirect costs under this grant? Yes
 No
- b. If yes, do you have an Indirect Cost Rate Agreement approved by the Federal government? Yes
 No
- c. If yes, provide the following information:
 Period Covered by the Indirect Cost Rate Agreement: From: 7/1/2007 To: 6/30/2010 (mm/dd/yyyy)
 Approving Federal agency: ED Other (Please specify): NIH
 Type of Rate (For Final Performance Reports Only): Provisional Final Other (Please specify):
- d. For Restricted Rate Programs (check one) -- Are you using a restricted indirect cost rate that :
- Is included in your approved Indirect Cost Rate Agreement?
 Complies with 34 CFR 76.564(c)(2)?

Human Subjects (Annual Institutional Review Board (IRB) Certification) (See instructions.)

10. Is the annual certification of Institutional Review Board (IRB) approval attached? Yes
 No N/A

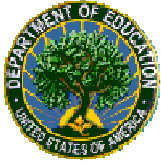
Performance Measures Status and Certification (See instructions.)

11. Performance Measures Status
- a. Are complete data on performance measures for the current budget period included in the Project Status Chart? Yes No
- b. If no, when will the data be available and submitted to the Department? (mm/dd/yyyy)
12. To the best of my knowledge and belief, all data in this performance report are true and correct and the report fully discloses all known weaknesses concerning the accuracy, reliability, and completeness of the data.

Name of Authorized Representative: Jordan Cohen	Title: Vice President Research
Signature:	Date:

Grant Performance Report (ED 524B) Executive Summary Attachment:

Title : Hurtig executive summary
 File : M:\hurtiglab\climbers 2009 annual report\hurtig-executive-summary.doc



U.S. Department of Education
Grant Performance Report (ED 524B)
Executive Summary

OMB No. 1890 - 0004
Expiration: 10-31-2007

PR/Award #: (Please
Enter)

R305G040145

Executive Summary

The purpose of this study is to provide evidence of the effectiveness of a reading intervention, *Breakthrough to Literacy*, taken to scale in the Chicago Public Schools (CPS). The *Breakthrough to Literacy* (BTL) model is driven by the four best predictors of reading achievement: oral language, oral comprehension, and vocabulary development; alphabet knowledge; phonological/phonemic awareness; and concepts of print. It incorporates professional development and in-classroom support for the teacher; classroom materials, manipulatives, and print materials for children; individualized software instruction; home connection materials; instructional print guides for teachers; and a powerful diagnostic and management software tool for the teacher.

This project is designed as a randomized cluster design study investigating whether BTL has an impact on both teacher and student outcomes. In this design, randomization to either a BTL condition or a comparison (“as-is”) condition occurred at the school level. Therefore, all studied classrooms and students within a school are in the same condition. The design called for two cohorts of students in pre-kindergarten who would receive two years of early literacy intervention. The outcome data, whether measured on classrooms or students, are clustered, or nested within school. All analyses reported account for this nesting by using hierarchical linear models (HLM), which estimate treatment impacts while accounting for the fact that the units being measured are not completely independent.

Forty-four schools were recruited and randomly assigned to either implement BTL in their preschool classrooms, or to a comparison group who would continue to implement their current preschool curricula. Prior to randomization, the schools were stratified with regard to their geographic location and whether or not the instruction was likely to be offered partly or entirely in Spanish (or another language). This created four strata or blocks of schools within which schools were randomly assigned to either implement BTL or remain in the “as-is” comparison condition. The study followed two consecutive cohorts of students in order to have enough statistical power to detect small impacts on students and moderate impacts on teachers.

The randomized cluster design allows us to estimate the impact of BTL on students’ literacy skills, teacher instruction, and classroom characteristics. The randomization took place at the school level, such that 22 schools were randomly assigned to implement the BTL curriculum and 22 were in the control condition. We use hierarchical linear modeling (HLM) to estimate impacts on students while taking into

account the nesting of students within schools, background characteristics of students and schools, and baseline data. To estimate impacts on teachers and classrooms, we use OLS regression. Like HLM, this method allows us to estimate the impacts of BTL on teachers and classrooms, while controlling for background characteristics and baseline data. However, HLM is not necessary for these analyses since data are not nested.

In the first year of the study, we collected baseline data on both students and classrooms. We used the above mentioned techniques to confirm that the random assignment process did indeed yield two groups that at the outset were equivalent on the outcome measures of interest. These baseline data serve as the covariates in the impact models that were fit on the follow-up data for the first cohort of pre-kindergarten classrooms.

During the second year of the project a second cohort of pre-kindergarten children were enrolled in the study and the first cohort of kindergarten classes were enrolled. BTL staff continued to provide teacher training and classroom support for the pre-kindergarten and kindergarten classrooms implementing the BTL curriculum. The Abt assessment team conducted both classroom observations and individual child assessments in both implementation and control schools.

During the third year of the project a second cohort of kindergarten children were enrolled in the study. BTL staff again provided teacher training and classroom support for the classrooms implementing the BTL curriculum. The Abt assessment team conducted both classroom observations and individual child assessments in both implementation and control schools. The impact analyses for the Pre-Kindergarten cohorts were completed.

During the fourth year of the project the Abt assessment team continued to follow both BTL and control classroom children to obtain assessments as they completed the first grade. The first impact analyses of the Kindergarten cohorts were completed and the analysis of the first cohort's first grade assessments was completed.

With supplemental funds we began a systematic examination of writing samples collected from the BTL classrooms. Writing samples from BTL classrooms were scanned and coded using a computer assisted coding system developed by the University of Iowa project staff. Preliminary analyses of these coded samples have been completed. This included a comparison of growth in writing skills with data from the classroom observations performed by the Abt team. In order to compare writing in BTL and non-BTL classrooms, a set of curriculum neutral writing probes were developed and a preliminary set of writing samples were obtained from a subset of BTL and non-BTL classrooms.

During the fifth year of the project the Abt assessment team continued to follow both BTL and control classroom children to obtain assessments as they completed the second grade. Further analysis of impact of pre-Kindergarten and Kindergarten cohorts

were conducted to look for differential effects in English Language Learners. Longitudinal analyses tested impacts of BTL by examining differences in performance on three language outcomes (expressive vocabulary, decoding, and word reading), at the end of Kindergarten, Grade 1, and Grade 2. These analyses also examined the differences between ELL and non-ELL subgroups. Working with Abt we have also conducted exploratory analyses of the utilization of the individualized instruction component of the BTL curriculum to gain an additional perspective of the fidelity of the implementation of BTL.

Analysis of writing samples continued. Samples from BTL and non-BTL using the curriculum neutral prompts that were collected during the spring term of year four (2007-2008) were scanned and coded. Analyses focused on determining how students vary in the length of written text, prevalence of incorrectly spelled words, and spelling strategies at the end of kindergarten. A second set of samples were collected over the 2008-2009 at three time points in order to be able to examine growth in the children's writing over the course of the year in Kindergarten. Abt conducted classroom observations to allow relating our growth analyses base on the writing samples to classroom practices.

Project Narrative

Tables

Attachment 1:

Title: **hurtig-tables** Pages: **0** Uploaded File: **M:\hurtiglab\climbers 2009 annual report\hurtig-tables.pdf**

Table 1: Sample for Kindergarten Data Collection

Study Years 2 and 3					
	Schools assigned to BTL	Classrooms implementing BTL ^a	Schools assigned to Control Group	Classrooms in Control Group	Totals
Total schools originally recruited	22		22		44
Schools that withdrew—Year 1	1		1		1
Total number of participating schools at end of Year 1	21		21		42
Total schools enrolled for Year 2	1		0 ^a		
Total number of participating schools at end of Years 2 and 3 ^b	22		21		43
K classrooms recruited, cohort 1		62		46	108
Total number of K children assessed, cohort 1 (AY05-06)—baseline (Fall '05)		681		475	1156
Total number of K children assessed, cohort 1 (AY05-06)—follow-up (Spr. '06)		905		669	1574
K classrooms ^c , cohort 2		60		48	108
Total number of K children assessed, cohort 2 (AY06-07)—baseline (Fall '06)		713		548	1261
Total number of K children assessed, cohort 2 (AY06-07)—follow-up (Spr. '07)		883		650	1533

Table 2a: Schedule of Classroom Observation Measures

Measure	Kindergarten (fall—1 st year)	Kindergarten (spring—1 st year)	Kindergarten (spring—2 nd year)
QUEST (general classroom quality)	✓		
OMLIT-RAP (reading aloud)	✓	✓	✓
OMLIT-CLOC (materials inventory)	✓	✓	✓
Arnett (teacher affect and responsiveness)	✓	✓	✓
OMLIT-Snapshot (activities, grouping, and integration of literacy)		✓	✓
OMLIT-CLIP (literacy instruction)		✓	✓
OMLIT-QUILL (quality rating of literacy instruction)		✓	✓

Table 2b: Schedule of Child Assessment Measures

Test	Kindergarten (fall)	Kindergarten (spring)
PPVT	✓	
Pre-CTOPPP: Print Knowledge	✓	
EOWPVT		✓
WRMT-R: Letter Identification		✓
WRMT-R: Word Identification		✓
WRMT-R: Word Attack		✓

Table 2c: Descriptive Statistics for Study Schools at
Baseline
(analytic sample, n=43)

	Mean	(S.D.)
Total Enrollment	791	(369)
Percent Low Income	89.3	(9.2)
Percent Minority	85.1	(22.6)
Mobility Rate	23.7	(9.9)

Data obtained from Chicago Public Schools website (2004).

Table 3: Estimated Differences at Baseline Between BTL and Control Kindergarten Students and Classrooms on Key Descriptive Characteristics

	BTL Schools	Control Schools	Difference Estimate	Standard Error	Effect size	Statistical Significance
School Characteristics (n=43)						
% Low Income	91.54	86.51	5.03	2.68	0.47	0.069
Total Enrollment	797.55	755.29	42.27	95.47	0.16	0.661
% Minority	90.20	78.36	11.84	5.82	0.42	0.049 *
Mobility Rate	23.56	23.96	-0.40	3.16	-0.05	0.900
Teacher Characteristics (n=109)						
Years worked in current position	10.01	7.84	2.17	1.79	0.35	0.234
% with at least master's degree	41.11	37.50	3.61	10.07	0.07	0.722
% with a teaching certificate	94.48	85.42	9.07	6.11	0.25	0.147
% who speak a language other than English at home	38.10	29.17	8.93	9.41	0.19	0.349
Classroom Characteristics (n=134)						
Class size (enrolled)	23.28	22.82	0.45	0.97	0.09	0.643
% Enrolled Children--English Speaking	59.77	51.17	8.60	8.06	0.20	0.293
% Enrolled Children--non-English Speaking	40.19	48.85	-8.66	8.05	-0.20	0.289
Classroom Measures (n=137)						
<i>Quest</i>						
Child Pre (something missing in this label?)	2.44	2.43	0.01	0.05	0.03	0.854
Caring and Responding	2.27	2.26	0.01	0.07	0.03	0.877
Positive Guidance	2.24	2.23	0.01	0.07	0.03	0.879
Supervision	2.93	2.92	0.01	0.04	0.03	0.846
Does No Harm	2.81	2.82	0.00	0.06	-0.01	0.955
Social and Emotional Development	1.83	1.72	0.11	0.07	0.35	0.131
Play	1.38	1.29	0.09	0.10	0.21	0.336
Cognitive Development	1.77	1.71	0.06	0.06	0.19	0.350
Instructional Style	2.15	2.11	0.04	0.09	0.08	0.683
Learning Activities and Opportunities	1.60	1.53	0.07	0.08	0.20	0.378

Language Development	2.01	1.98	0.04	0.07	0.11	0.604
Television and Computers	2.88	2.83	0.05	0.04	0.20	0.267
Arnett						
Positive	2.97	2.87	0.10	0.10	0.15	0.325
Punitive (higher is less desirable)	3.50	3.41	0.09	0.10	0.12	0.365
Detached (higher is less desirable)	3.54	3.58	-0.04	0.09	-0.06	0.675
Permissive	2.61	2.51	0.10	0.08	0.20	0.219
Promotion of Child Independence	2.73	2.74	-0.01	0.15	-0.02	0.940
Arnett Standardized Composite (I don't think this is meaningful descriptively)	51.22	50.47	0.75	1.44	0.08	0.604
Student Characteristics and Measures (n=2417)						
Age	72.10	72.45	-0.35	0.29	-0.09	0.236
PPVT	54.60	54.80	-0.20	1.98	-0.01	0.921
Pre-CTOPP (Print Knowledge)	24.34	25.53	-1.20	1.09	-0.13	0.278

Notes: 1 The student sample includes only students tested in the fall. Data on *age* are missing for 120 students and scores on *PPVT* are missing for 23 students; those students are thus excluded from the sample for these analyses.

2 PPVT and Pre-CTOPP are presented as raw scores.

3 Table reads: The average estimated percent low-income is 91.54% in BTL schools and 86.51% in Control schools. The difference of 5.03 percentage points is not statistically significant at the $p \leq .05$ level.

Table 4a: Impact Estimates for OMLIT constructs of the kindergarten classrooms implementing BTL for one year

	BTL Schools	Control Schools	Impact	Standard Error of the Impact	Effect size of Impact	Statistical Significance of Impact	
OMLIT Measures - Year 1							
Oral Language	52.279	44.573	7.706	1.865	0.735	0.0002	*
Print Knowledge	49.862	48.313	1.549	1.539	0.250	0.321	
Phonological Awareness	48.587	47.452	1.136	2.282	0.100	0.622	
Print Motivation	50.828	44.823	6.005	1.826	0.641	0.002	*
ELL_Students	54.797	49.395	5.401	4.209	0.298	0.213	
Literacy Resources	50.234	45.902	4.333	2.397	0.374	0.079	
Literacy Activities	50.427	47.054	3.373	1.621	0.469	0.045	*
Proportion of time spent in computer act.	0.043	0.005	0.038	0.007	1.651	<.0001	*
Writing subscale	0.855	0.824	0.031	0.060	0.093	0.611	
Proportion of time spent in routines	0.308	0.328	-0.020	0.025	-0.142	0.428	
Proportion of time in emergent writing act.	0.058	0.080	-0.022	0.022	-0.215	0.313	
Opportunities/Materials to engage in writing	1.570	1.709	-0.140	0.174	-0.178	0.428	
Proportion of time in writing activities	0.120	0.139	-0.020	0.024	-0.156	0.429	

Notes:

1 Sample size is 131 classrooms (except for the "ELL_Students" outcome which is available for 86 classrooms). Of these, 104 were observed in Spring 2006 and 27 were observed in Spring 2007.

2 Table reads: The average estimated score on the Oral Language construct of the OMLIT is 52.279 points for BTL teachers and 44.573 points for control teachers. The difference of 7.706 points is statistically significant at the $p \leq .05$ level.

Table 4b: Impact Estimates for OMLIT outcomes of the classrooms implementing BTL for two years

OMLIT Measures – Year 2

Oral Language

Print Knowledge

Phonological Awareness

Print Motivation

ELL_Students

Literacy Resources

Literacy Activities

Proportion of time spent in computer activities

Writing subscale

Proportion of time spent in routines

Proportion of time in emergent writing act.

Opportunities/Materials to engage in writing

Proportion of time in writing activities

Notes:

Sample size is 80 classrooms (except for the "ELL_Students" outcome which is available for 58 classrooms). All of these classrooms were observed in Spring 2007. Impacts estimates are from a 2-level HLM models (classrooms nested within schools with school level random error terms) which employs a-priori selected covariates (north, spanish, north*spanish, and percent minority). These covariates are centered.

Table 5a: Impact Estimates for Arnett subscales of the Kindergarten classrooms implementing BTL for one year

	BTL Schools	Control Schools	Impact	Standard Error of the Impact	Effect size of Impact	Statistical Significance of Impact
Arnett Measures - Year 1						
Positive	3.019	2.847	0.172	0.122	0.241	0.168
Punitive	3.366	3.294	0.072	0.152	0.096	0.638
Detach	3.507	3.605	-0.098	0.106	-0.153	0.365
Permiss.	2.802	2.662	0.140	0.099	0.263	0.168
Promotion of Child Independence	3.003	2.964	0.039	0.135	0.060	0.771
Arnett Standardized Composite	50.826	50.392	0.433	1.688	0.049	0.799

Notes:

1 Sample size is 131 classrooms. Of these, 104 were observed in Spring 2006 and 27 were observed in Spring 2007.

2 Table reads: The average estimated score on the Arnett *Positive* subscale is 3.019 for BTL teachers and 2.847 for control teachers. The difference of .172 points is not statistically significant at the $p \leq .05$ level.

Table 5b: Impact Estimates for Arnett subscales of the Kindergarten classrooms implementing BTL for two years

	BTL Schools	Control Schools	Impact	Standard Error of the Impact	Effect size of Impact	Statistical Significance of Impact
Arnett Measures - Year 2						
Positive	3.156	3.364	-0.208	0.182	-0.384	0.263
Punitive	3.475	3.571	-0.096	0.176	-0.159	0.586
Detach	3.513	3.660	-0.147	0.130	-0.337	0.268
Permiss.	2.969	2.889	0.080	0.134	0.168	0.555
Promotion of Child Independence	3.220	3.299	-0.079	0.181	-0.110	0.667
Arnett Standardized Composite	50.216	52.373	-2.157	2.134	-0.321	0.319

Notes:

1 Sample size is 80 classrooms. All of these were observed in Spring 2007.

2 Table reads: The average estimated score on the Arnett *Positive* subscale is 3.156 for BTL teachers and 3.364 for control teachers. The difference of -0.208 points is not statistically significant at the $p \leq .05$ level.

Table 6: Estimated Differences Between Kindergarten Students in BTL and Control Classrooms

	BTL Schools	Control Schools	Impact Estimate	Standard Error of Impact	Effect size of Impact	Statistical Significance of Impact
Cohorts 1 and 2						
Student Measures						
Expressive One Word Vocabulary	80.44	79.75	0.69	1.53	0.04	0.654
Word Attack	100.85	99.57	1.28	0.87	0.11	0.149
Word ID	99.83	99.92	-0.09	1.44	-0.01	0.950
Letter ID	98.82	97.45	1.37	1.19	0.10	0.257
Cohort 1 Only						
Student Measures						
Expressive One Word Vocabulary	78.29	77.49	0.80	1.52	0.05	0.603
Word Attack	99.96	99.00	0.96	1.04	0.08	0.364
Word ID	100.21	98.62	1.59	1.65	0.10	0.343
Letter ID	98.76	97.11	1.64	1.35	0.12	0.231
Cohort 2 Only						
Student Measures						
Expressive One Word Vocabulary	81.95	81.59	0.36	1.97	0.02	0.857
Word Attack	102.00	100.38	1.62	1.19	0.14	0.183
Word ID	100.52	100.87	-0.35	1.82	-0.02	0.846
Letter ID	98.99	97.86	1.13	1.43	0.09	0.434

Notes: 1 Overall student impacts are based on 43 schools (22 BTL and 21 Control), 134 classrooms (76 BTL and 58 Control), and 3107 students (1788 BTL and 1319 Control) from two cohorts.

2 Cohort 1 student impacts are based on 43 schools (22 BTL and 21 Control), 132 classrooms (76 BTL and 58 Control), and 3107 students (1788 BTL and 1319 Control) from two cohorts.

3 Cohort 2 student impacts are based on 43 schools (22 BTL and 21 Control), 132 classrooms (76 BTL and 58 Control), and 3107 students (1788 BTL and 1319 Control) from two cohorts.

4 All outcome measures are reported in standardized scores with a mean of 100 and a standard deviation of 15.

5 Table reads: The average estimated score on Expressive One Word Vocabulary is 80.44 points for students in BTL schools and 79.75 for students in Control schools. The difference of 0.69 points is not statistically significant at the $p \leq .05$ level.

**Table 7: Estimated Differences Between Kindergarten Students in BTL and Control Classrooms:
ELL and non-ELL Subgroups**

	BTL Schools	Control Schools	Impact Estimate	Standard Error of Impact	Effect size of Impact	Statistical Significance of Impact
ELL Subgroup						
Student Measures						
Expressive One Word Vocabulary	76.5695	76.311558	0.2579	2.143	0.0156	0.9053
Word Attack	100.1560	98.585427	1.5706	0.9991	0.1383	0.1302
Word ID	98.4372	98.659548	-0.2223	1.8744	0.0142	0.9067
Letter ID	96.8666	96.081658	0.7849	1.7698	0.0591	0.6617
Non-ELL Subgroup						
Student Measures						
Expressive One Word Vocabulary	85.9467	87.816445	-1.8697	1.3749	0.1132	0.1821
Word Attack	101.9290	101.92543	0.0036	1.0634	0.0003	0.9973
Word ID	102.5618	105.061186	-2.4994	1.503	0.1593	0.1048
Letter ID	102.0429	102.491396	-0.4485	0.7954	0.0338	0.5762

Notes: 1 The ELL subgroup includes 29 schools (14 BTL and 15 Control) and 1,830 students (1,034 BTL and 796 Control).

2 The non-ELL subgroup includes 43 schools (22 BTL and 21 Control) and 1,277 students (754 BTL and 523 Control).

3 Table reads: The average estimated score on Expressive One Word Vocabulary is 76.56 points for ELL students in BTL schools and 76.31 for ELL students in Control schools. The difference of 0.25 points is not statistically significant at the $p \leq .05$ level.

Table 8: Exploratory Analysis of BTL Exposure versus Control Exposure (1, 2, 3 and 2 or 3 years)

Contrast	Expressive Vocabulary (n = 3107)	Word Attack (n = 3106)	Word ID (n = 3106)	Letter ID (n = 3107)
1 yr BTL vs. 1 yr Control	1.421 (1.55)	1.304 (0.90)	-0.089 (1.46)	1.342 (1.21)
2 yrs BTL vs. 2 yrs Control	-1.359 (1.78)	1.453 (1.11)	0.318 (1.69)	1.103 (1.39)
3 yrs BTL vs. 3 yrs Control	7.934* (3.90)	2.523 (2.86)	0.103 (3.81)	6.792* (3.07)
2 or 3 yrs BTL vs. 2 or 3 yrs Control	-0.644 (1.76)	1.541 (1.09)	0.325 (1.66)	1.546 (1.37)

Notes:

1 Standard errors are in parentheses below the point values.

2 * $p \leq .05$

3 Table reads: The estimated difference between BTL and control students with 1 year of exposure on the Expressive One Word Picture Vocabulary test is 1.421 points, in favor of BTL students. The difference is not statistically significant at the $p \leq .05$ level.

Table 9. Longitudinal Sample by Cohort, Grade, ELL Status, and Treatment Status

	Kindergarten				Grade 1				Grade 2					
	ELL		Non-ELL		ELL		Non-ELL		ELL		non-ELL			
	BTL	Ctrl	BTL	Ctrl	BTL	Ctrl	BTL	Ctrl	BTL	Ctrl	BTL	Ctrl		
Cohort 1 children	561	441	502	329	497	376	310	193	431	337	237	130		
Cohort 2 children	619	457	399	291	491	347	250	184	NA ^a	NA	NA	NA		
Total by Treatment & Grade	1,180	898	901	620	988	723	560	377	431	337	237	130		
Total by Treatment, Grade & ELL status	2,078		1,521		1,711		937		768		367			
Total by Grade	3,599				2,648				1,135					
Total Sample	7,382													
All Grades	ELL students				4,557				Non-ELL students				2,825	

^a These data will be available at the end of the 2008-09 school year.

Note: This table includes all students with either a pre-test or post-test score or both.

Table 10: Descriptive Statistics^{1,2}

		1 st Cohort				2 nd Cohort			
		Mean	Std. Dev.	Min.	Max.	Mean	Std. Dev.	Min.	Max.
Kindergarten									
Expressive Vocabulary	Raw Score	43.71	16.91	0.00	98.00	45.31	17.04	0.00	103.00
	Std. Score	79.34	15.96	54.00	146.00	81.05	16.60	0.00	146.00
	NCE	23.30	19.24	1	99	25.53	20.11	1	99
Word Attack	Raw Score	2.20	4.42	0.00	34.00	2.70	4.51	0.00	39.00
	Std. Score	99.24	10.83	0.00	135.00	101.00	10.98	0.00	136.00
	W-Score	449.38	14.14	440	551	451.71	14.86	440	522
Word Identification	Raw Score	7.74	10.84	0.00	62.00	8.52	11.27	0.00	81.00
	Std. Score	99.12	15.52	0.00	148.00	100.33	15.45	48.00	155.00
	W-Score	370.73	28.92	340	481	372.99	29.50	340	515
Age		6.12	0.31	5.36	7.39	6.12	0.31	5.28	7.34
Test Date		-	-	02/27/2006	05/30/2006	-	-	02/10/2007	06/07/2007
First Grade									
Expressive Vocabulary	Raw Score	53.72	17.60	0.00	111.00	52.92	17.87	0.00	112.00
	Std. Score	81.96	16.10	54.00	146.00	81.48	16.18	0.00	146.00
	NCE	26.35	20.10	1	99	25.78	19.98	1	99
Word Attack	Raw Score	10.08	8.75	0.00	42.00	10.01	8.51	0.00	40.00
	Std. Score	103.77	14.25	0.00	146.00	104.05	13.48	74.00	141.00
	W-Score	471.54	18.49	440	532	471.52	18.04	440	525
Word Identification	Raw Score	27.90	16.90	0.00	104.00	27.96	16.95	0.00	76.00
	Std. Score	104.35	14.53	66.00	153.00	104.68	14.57	9.00	149.00
	W-Score	416.57	33.01	340	561	416.61	33.12	340	507
Age		7.04	0.32	6.31	8.39	7.02	0.31	6.11	8.13
Test Date		-	-	01/22/2007	05/17/2007	-	-	01/28/2008	05/19/2008
Second Grade									
Expressive Vocabulary	Raw Score	63.19	18.12	3.00	120.00	-	-	-	-
	Std. Score	83.20	15.71	54.00	146.00	-	-	-	-
	NCE	27.69	20.02	1	99	-	-	-	-
Word Attack	Raw Score	18.36	10.34	0.00	41.00	-	-	-	-

¹ Note: This exhibit presents **unadjusted** means and standard deviations of raw scores, standard scores, W scores, and NCEs.

² Measures are: Expressive One-Word Picture Vocabulary Test (EOWPVT)—“Expressive Vocabulary”; Woodcock Reading Mastery Tests-Revised/Normative Update (WRMT-R/NU) subtests—Word Attack, Word Identification.

	Std. Score	102.35	14.81	0.00	148.00	-	-	-	-
	W-Score	486.53	17.93	440	528				
Word Identification	Raw Score	47.27	14.00	0.00	84.00	-	-	-	-
	Std. Score	100.27	12.12	9.00	135.00	-	-	-	-
	W-Score	452.54	26.00	340	520				
Age		8.12	0.31	7.34	9.26	-	-	-	-
Test Date		-	-	02/21/2008	05/19/2008	-	-	-	-

Table 11: Cohort 1 and 2 BTL versus Control Group Differences

	1 st Cohort			2 nd Cohort		
	Expressive Vocabulary	Word Attack	Word ID	Expressive Vocabulary	Word Attack	Word ID
Kindergarten						
BTL Mean	22.95	454.11	380.91	23.54	456.99	380.88
Control Mean	22.23	452.02	378.09	22.82	455.05	382.24
BTL-Control Difference	0.72 (0.562)	2.09 (0.133)	2.82 (0.277)	-0.73 (0.646)	1.94 (0.285)	-1.36 (0.695)
1st Grade						
BTL Mean	27.03	470.17	412.16	24.94	469.94	413.80
Control Mean	26.27	470.19	415.56	28.07	471.46	416.22
BTL-Control Difference	0.76 (0.591)	-0.02 (0.991)	-3.40 (0.321)	-3.13 (0.147)	-1.52 (0.504)	-2.42 (0.619)
2nd Grade						
BTL Mean	26.94	478.69	434.20	-	-	-
Control Mean	27.28	478.40	434.55	-	-	-
BTL-Control Difference	-0.35 (0.826)	0.294 (0.905)	-0.35 (0.922)	-	-	-
Difference-in-Differences						
Kindergarten to 1 st Grade	0.05 (0.959)	-2.10 (0.127)	-6.22* (0.017)	-2.40 (0.144)	-3.46* (0.032)	-1.06 (0.773)
Kindergarten to 2 nd Grade	-1.07 (0.375)	-1.79 (0.419)	-3.17 (0.263)	-	-	-
1 st Grade to 2 nd Grade	-1.12 (0.402)	0.31 (0.898)	3.05 (0.393)	-	-	-

Notes: Scores are Normal Curve Equivalents for EOWPVT, and W-scores for WRMT-Word ID and WRMT-Word Attack. P-values are in parentheses. A two-tailed test of significance was used, and where applicable, statistically significant findings at the $p \leq 0.05$ level are indicated by asterisk (*).

Table 12: Cohort 1 and 2 BTL versus Control Group Differences in the ELL Subgroup

	1 st Cohort			2 nd Cohort		
	Expressive Vocabulary	Word Attack	Word ID	Expressive Vocabulary	Word Attack	Word ID
Kindergarten						
BTL Mean	20.53	451.07	375.18	22.12	457.36	380.45
Control Mean	19.69	448.87	371.11	23.17	452.86	374.65
BTL-Control Difference	0.83 (0.635)	2.2 (0.09)	4.07 (0.141)	-1.05 (0.567)	4.5* (0.04)	5.8 (0.138)
1st Grade						
BTL Mean	21.23	470.18	410.68	23.92	470.85	415.95
Control Mean	20.94	471.41	413.88	26.90	470.6	413.92
BTL-Control Difference	0.28 (0.888)	-1.23 (0.545)	-3.2 (0.471)	-2.98 (0.286)	0.25 (0.935)	2.03 (0.760)
2nd Grade						
BTL Mean	19.63	480.66	434.34	-	-	-
Control Mean	19.58	479.65	433.05	-	-	-
BTL-Control Difference	0.04 (0.986)	1.01 (0.575)	1.29 (0.757)	-	-	-
Difference-in-Differences						
Kindergarten to 1 st Grade	-0.55 (0.691)	-3.43 (0.073)	-7.27 (0.068)	-1.93 (0.406)	-4.25 (0.087)	-3.77 (0.512)
Kindergarten to 2 nd Grade	-0.79 (0.656)	-1.19 (0.473)	-2.78 (0.447)	-	-	-
1 st Grade to 2 nd Grade	-0.24 (0.905)	2.24 (0.325)	4.49 (0.370)	-	-	-

Notes: Scores are Normal Curve Equivalents for EOWPVT, and W-scores for WRMT-Word ID and WRMT-Word Attack. P-values are in parentheses. A two-tailed test of significance was used, and where applicable, statistically significant findings at the $p \leq 0.05$ level are indicated by asterisk (*).

Table 13: Cohort 1 and 2 BTL versus Control Group Differences in the Non-ELL Subgroup

	1 st Cohort			2 nd Cohort		
	Expressive Vocabulary	Word Attack	Word ID	Expressive Vocabulary	Word Attack	Word ID
Kindergarten						
BTL Mean	28.06	457.02	386.24	30.95	454.8	378.78
Control Mean	27.71	454.67	382.86	32.1	455.88	387.12
BTL-Control Difference	0.35 (0.858)	2.35 (0.201)	3.38 (0.314)	-1.15 (0.687)	-1.08 (0.628)	-8.34 (0.075)
1st Grade						
BTL Mean	36.32	471	418.13	30.61	467.6	409.24
Control Mean	35.16	469.86	419.54	34.18	472.26	417.98
BTL-Control Difference	1.16 (0.574)	1.14 (0.628)	-1.41 (0.732)	-3.57 (0.283)	-4.66 (0.141)	-8.74 (0.156)
2nd Grade						
BTL Mean	39.28	477.13	437.29	-	-	-
Control Mean	40.53	477.29	437.05	-	-	-
BTL-Control Difference	-1.25 (0.561)	-0.16 (0.956)	0.24 (0.956)	-	-	-
Difference-in-Differences						
Kindergarten to 1 st Grade	0.81 (0.526)	-1.21 (0.546)	-4.79 (0.115)	-2.42 (0.257)	-3.58 (0.181)	-0.4 (0.930)
Kindergarten to 2 nd Grade	-1.6 (0.257)	-2.51 (0.361)	-3.14 (0.338)	-	-	-
1 st Grade to 2 nd Grade	-2.41 (0.090)	-1.31 (0.66)	1.65 (0.673)	-	-	-

Notes: Scores are Normal Curve Equivalents for EOWPVT, and W-scores for WRMT-Word ID and WRMT-Word Attack. P-values are in parentheses. A two-tailed test of significance was used, and where applicable, statistically significant findings at the $p \leq 0.05$ level are indicated by asterisk (*).

Table 14: Two-Level HLM Model Examining Productivity, Incorrect Spelling, and Spelling Strategies in Kindergarten Writing in May (165 students, 9 classrooms, 2 school districts).

	Word Count	Number of Incorrectly Spelled Words	Proportion of Words that are Spelled Incorrectly	First-Subsequent Consonant	Vocalic Nucleus	Advanced Spontaneous Spelling
Fixed Effects						
Intercept	11.67** (2.48)	4.05***(0.69)	.387***(.045)	.603***(.08)	.757***(.07)	.182** (.06)
WestDesMoines ^a	18.40***(4.21)	7.69***(1.14)	.038 (.076)	.232~ (.13)	.207~ (.12)	.058 (.096)
Random Effects						
Student	122.78***(13.89)	20.44***(2.31)	.056***(.006)	.086***(.010)	.080***(.009)	.055***(.006)
Classroom	29.05~ (19.01)	1.56 (1.41)	.009~ (.006)	.031* (.019)	.024~ (.015)	.016~ (.01)
Variance Breakdown						
Classroom	10.8%	7.1%	10.9%	26.5%	23.1%	22.5%
Deviance (-2LL)	1266.55	969.20	11.91	86.16	72.49	9.91

^a Indicator variable, indicating West Des Moines School District. The reference category is Chicago Public Schools.

~ p < .10; * p < .05; ** p < .01; *** p < .001.

Table 15: Estimated Linear Growth Models for Productivity (Word Count), Incorrectly Spelled Words (raw frequency and proportion of total words), and Spelling Strategies (proportion of incorrectly spelled words) in Kindergarten Writing (measured in February, April, and May, n=71 students in 3 classrooms).

	Word Count	Number of Incorrectly Spelled Words	Proportion of Words that are Spelled Incorrectly	First- Subsequent Consonant	Vocalic Nucleus	Advanced Spontaneous Spelling
Fixed Effects						
Intercept	16.47*** (2.61)	7.38*** (1.12)	.450*** (.03)	.822*** (.04)	.857*** (.04)	.145*** (.02)
Classroom1 ^a	2.34 (.51)	2.12 (1.42)	.033 (.04)	-.069 (.05)	-.002 (.03)	-.025 (.03)
Classroom2 ^a	10.54** (3.62)	2.97* (1.46)	-.061 (.04)	-.088 (.05)	-.033 (.03)	-.033 (.03)
Rate of Change	4.63*** (.86)	1.37*** (0.41)	-.008 (.01)	.029~ (.02)	.052** (.02)	.045*** (.01)
Variance Components						
Within-person, σ_{ε}^2	97.7*** (12.1)	22.19*** (2.75)	.021*** (.003)	.029*** (.005)	.026*** (.005)	.019*** (.003)
Initial status, σ_{01}	83.21** (30.7)	18.47** (6.47)	.011** (.005)	.050*** (.01)	.058*** (.016)	.0004 (.004)
Rate of change, σ_1^2	--	--	--	.003 (.004)	.007~ (.005)	.0004 (.002)
Covariance, σ_{01}	21.2~ (12.4)	-0.91 (2.27)	.0004 (.002)	-.012~ (.006)	-.024** (.008)	.003 (.002)
Deviance (-2LL)	1609.67	1289.07	-119.83	-34.26	-102.57	-168.05
Pseudo-R_{ε}^2	.181	.072	.00	.065	.188	.095

^a Indicator variables – Classroom1 and Classroom2 – represent the three kindergarten classrooms in the sample. The reference category is Classroom3.

~ p < .10; * p < .05; ** p < .01; *** p < .001.

Project Narrative

Charts

Attachment 1:

Title: **Hurtig-charts** Pages: **0** Uploaded File: **M:\hurtiglab\climbers 2009 annual report\hurtig-figures.pdf**

Figure 1A. Expressive Vocabulary — First Cohort

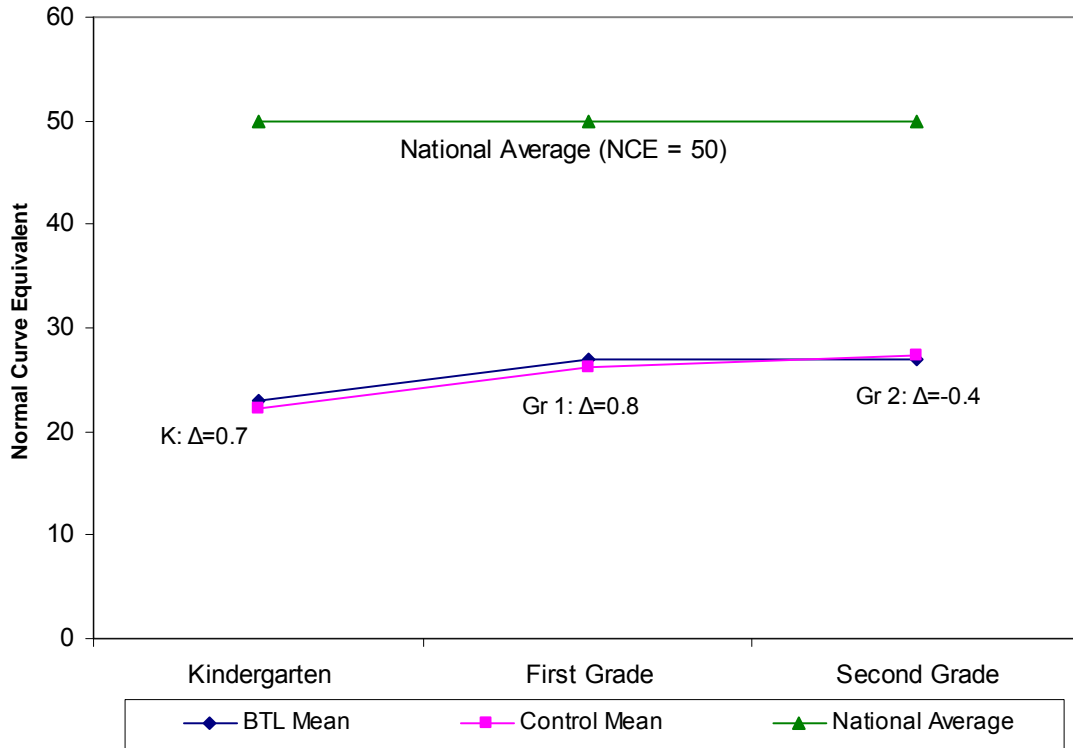


Figure Reads: There are no statistically significant differences between BTL and Control groups in Expressive Vocabulary at the end of Kindergarten, Grade 1, or Grade 2 for cohort 1. Similarly, there are no statistically significant BTL versus Control group differences between Kindergarten and Grade 1, between Grade 1 and Grade 2, or between Kindergarten and Grade 2. Both BTL and Control groups are performing below the national average of 50 NCE at all time points.

Figure 1B. Expressive Vocabulary — Second Cohort

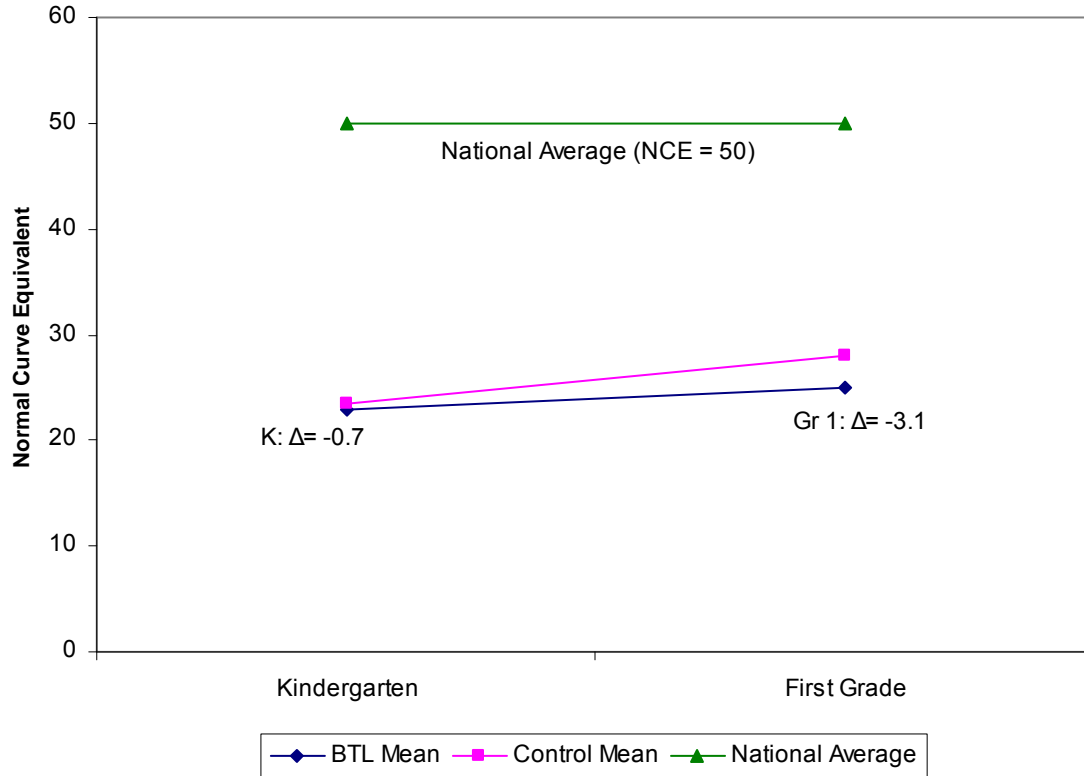


Figure Reads: There are no statistically significant differences between BTL and Control groups in Expressive Vocabulary at the end of Kindergarten or Grade 1 for cohort 2. Similarly, there are no statistically significant BTL versus Control group differences-in-differences from Kindergarten to Grade 1. Both BTL and Control groups are performing below the national average of 50 NCE at all time points.

Figure 1C. Expressive Vocabulary — First Cohort — by ELL Subgroup

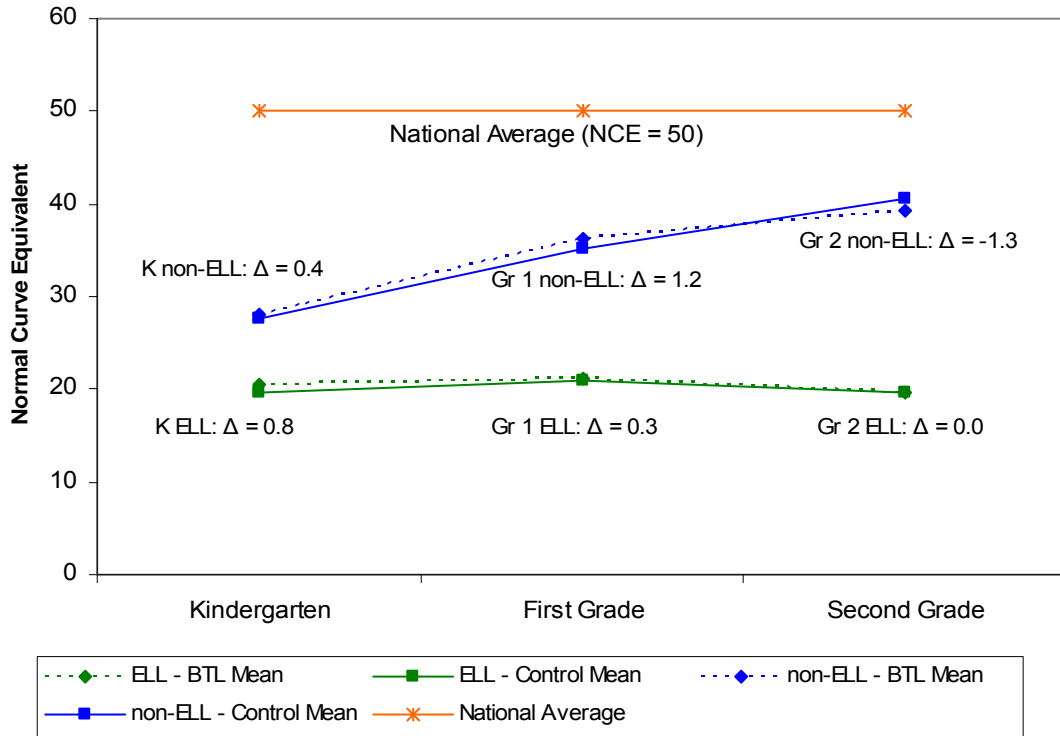


Figure Reads: There are no statistically significant differences between BTL and Control groups within ELL or non-ELL subgroups in Expressive Vocabulary at the end of Kindergarten, Grade 1, or Grade 2 for cohort 1. Similarly, there are no statistically significant BTL versus Control group differences-in-differences between Kindergarten and Grade 1, between Grade 1 and Grade 2, or between Kindergarten and Grade 2 within either subgroup. Both BTL and Control children within both ELL and non-ELL subgroups are performing below the national average of 50 NCE at all time points.

Figure 1D. Expressive Vocabulary — Second Cohort — by ELL Subgroup

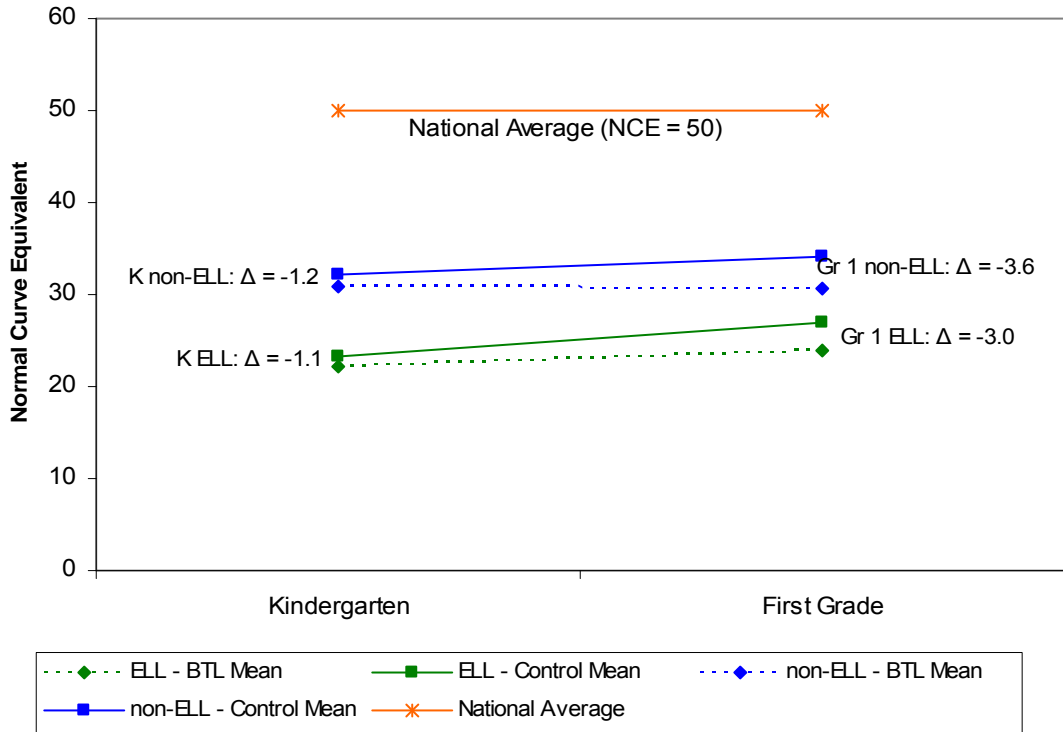


Figure Reads: There are no statistically significant differences between BTL and Control groups within ELL or non-ELL subgroups in Expressive Vocabulary at the end of Kindergarten or Grade 1 for cohort 2. Similarly, there are no statistically significant BTL versus Control group differences-in-differences between Kindergarten and Grade 1 within either subgroup. Both BTL and Control children within both ELL and non-ELL subgroups are performing below the national average of 50 NCE at all time points.

Figure 2A. Word Attack — First Cohort

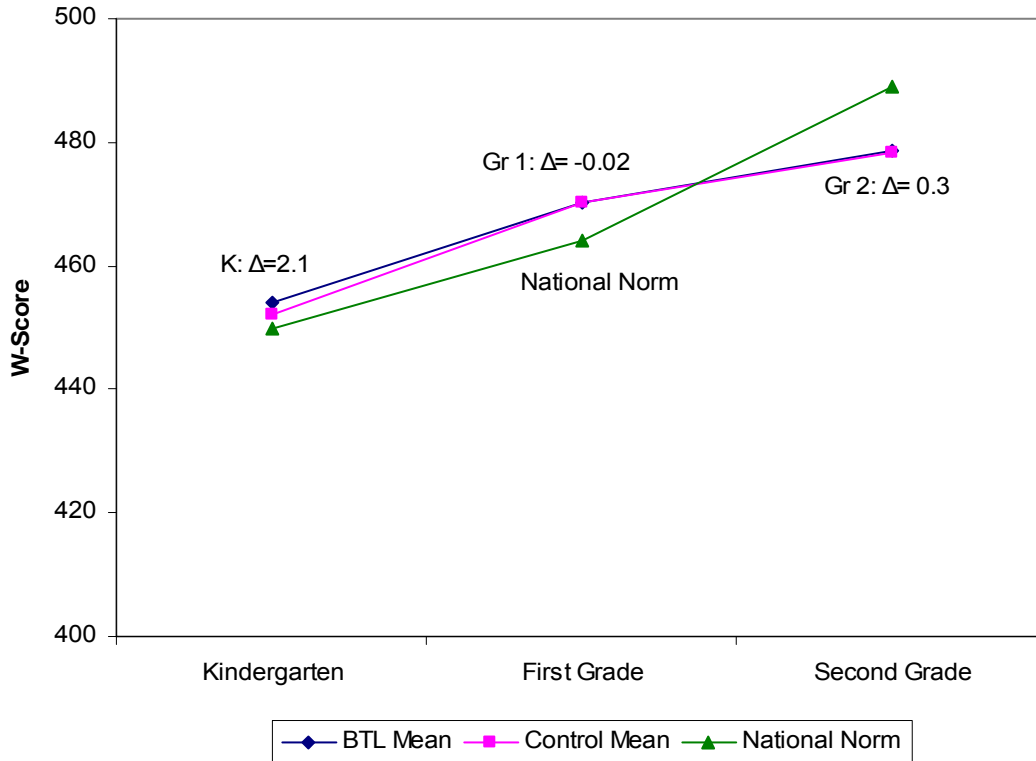


Figure Reads: There are no statistically significant differences between BTL and Control groups in Word Attack at the end of Kindergarten, Grade 1, or Grade 2 for cohort 1. Similarly, there are no statistically significant BTL versus Control group differences-in-differences between Kindergarten and Grade 1, between Grade 1 and Grade 2, or between Kindergarten and Grade 2. Both BTL and Control groups are performing above the national norm at the end of Kindergarten and Grade 1, but then fall below the norm by the end of Grade 2.

Figure 2B. Word Attack — Second Cohort

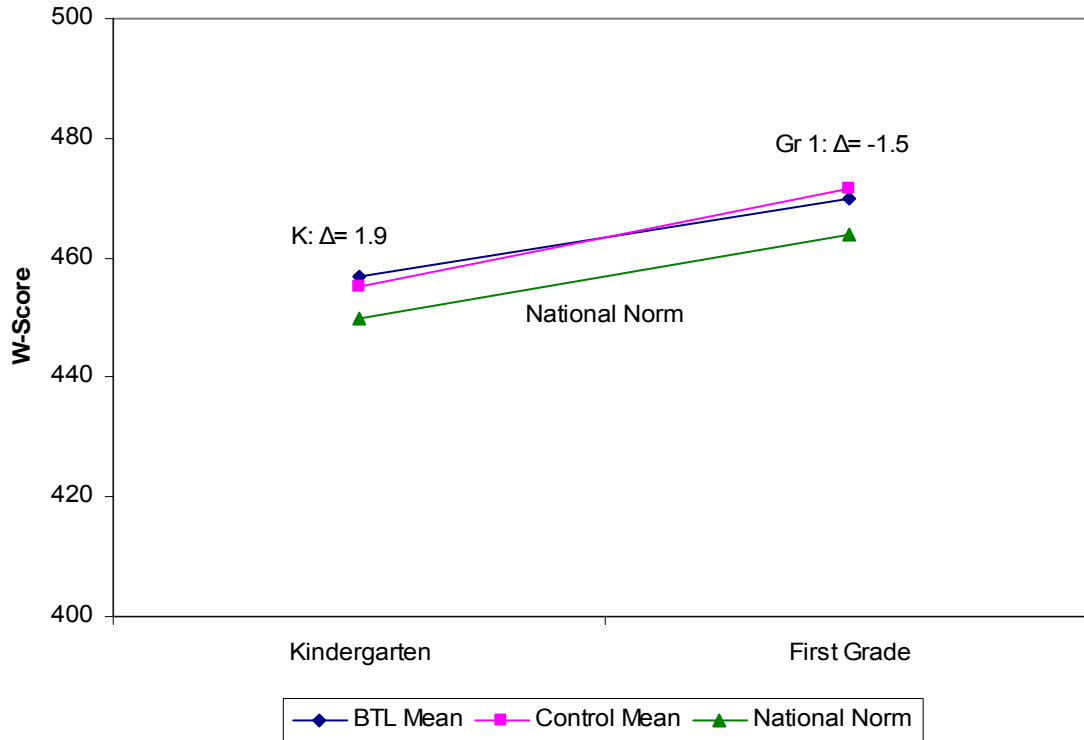


Figure Reads: There are no statistically significant differences between BTL and Control groups in Word Attack at the end of Kindergarten or Grade 1 for cohort 2. There is, however, a statistically significant BTL versus Control group difference-in-differences (favoring Control) between Kindergarten and Grade 1. Both BTL and Control groups are performing above the national norm at the end of Kindergarten and Grade 1.

Figure 2C. Word Attack — First Cohort — by ELL Subgroup

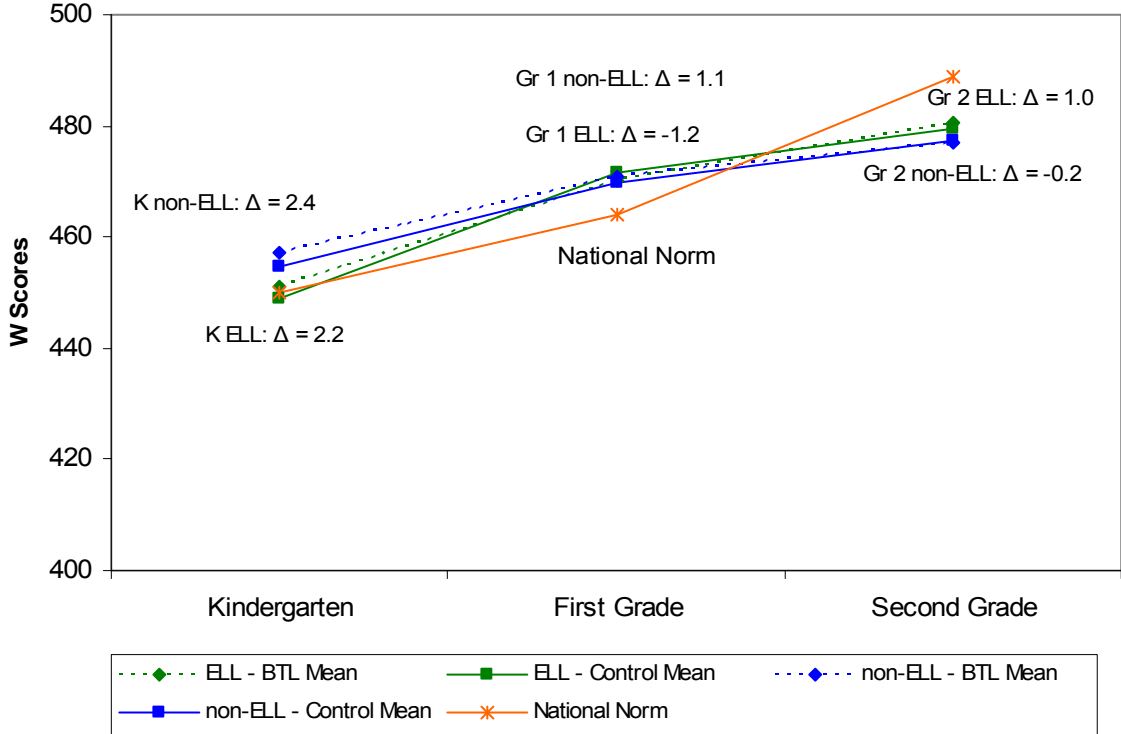


Figure Reads: There are no statistically significant differences between BTL and Control groups within ELL or non-ELL subgroups in Word Attack at the end of Kindergarten, Grade 1, or Grade 2 for cohort 1. Similarly, there are no statistically significant BTL versus Control group differences-in-differences within either subgroup between Kindergarten and Grade 1, between Grade 1 and Grade 2, or between Kindergarten and Grade 2. Both BTL and Control groups within ELL and non-ELL subgroups are performing above the national norm at the end of Kindergarten and Grade 1, but then fall below the norm by the end of Grade 2.

Figure 2D. Word Attack — Second Cohort — by ELL Subgroup

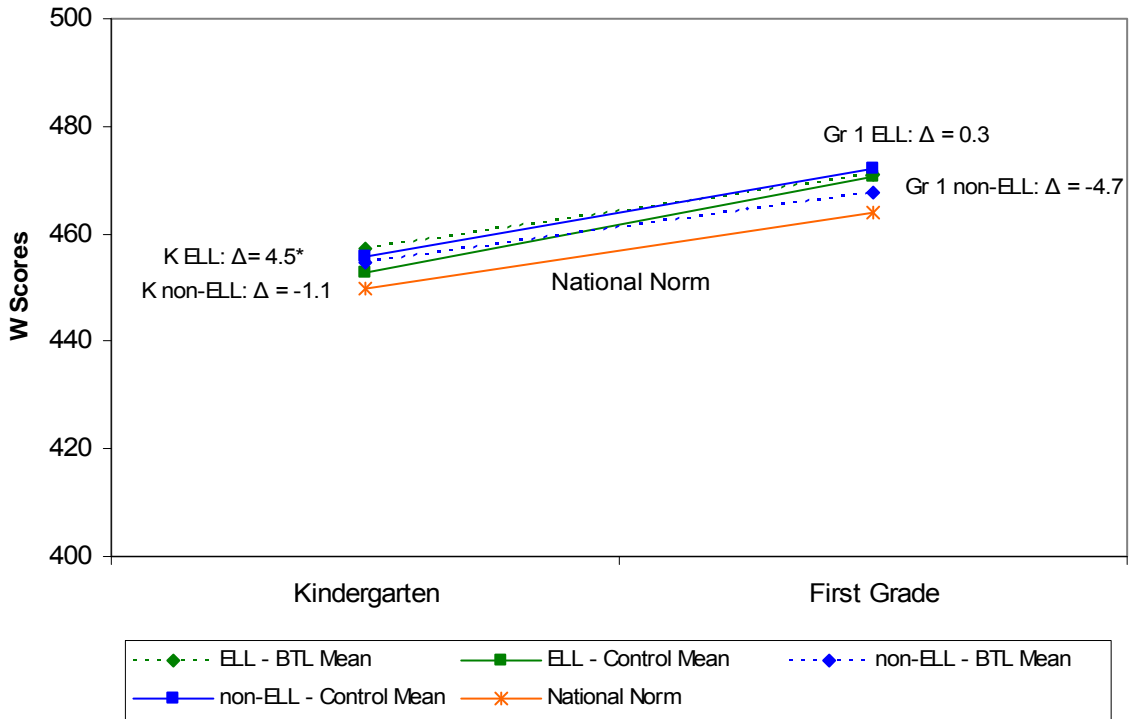


Figure Reads: There is a statistically significant difference between BTL and Control groups at the end of Kindergarten in the ELL subgroup (favoring BTL), but not in the non-ELL subgroup. There are no statistically significant differences between BTL and Control groups at the end of Grade 1 for either subgroup. There is not a statistically significant BTL versus Control group difference-in-differences between Kindergarten and Grade 1 for either subgroup. Both BTL and Control groups within ELL and non-ELL subgroups are performing above the national norm at the end of Kindergarten and Grade 1.

Figure 3A. Word ID — First Cohort

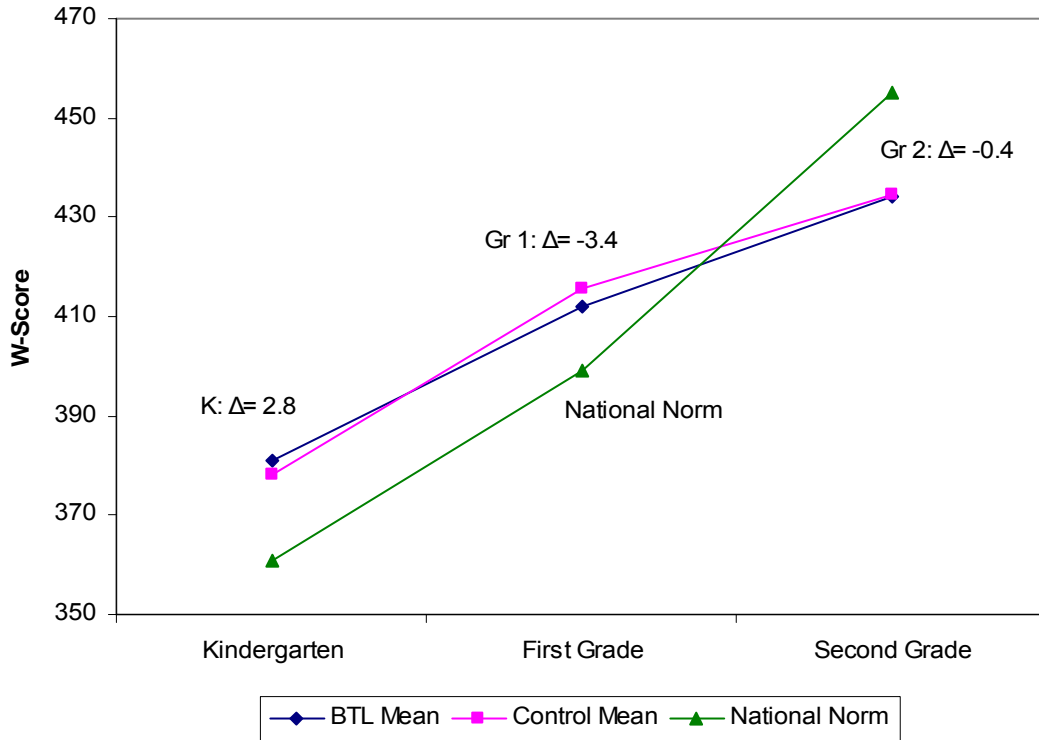


Figure Reads: There are no statistically significant differences between BTL and Control groups in Word Identification at the end of Kindergarten, Grade 1, or Grade 2 for cohort 1. Similarly, there are no statistically significant BTL versus Control group differences-in-differences between Kindergarten and Grade 2 or between Grade 1 and Grade 2. There is, however, a statistically significant difference-in-differences between Kindergarten and Grade 1 (favoring Control). Both BTL and Control groups are performing above the national norm at the end of Kindergarten and Grade 1, but then fall below the norm by the end of Grade 2.

Figure 3B. Word ID — Second Cohort

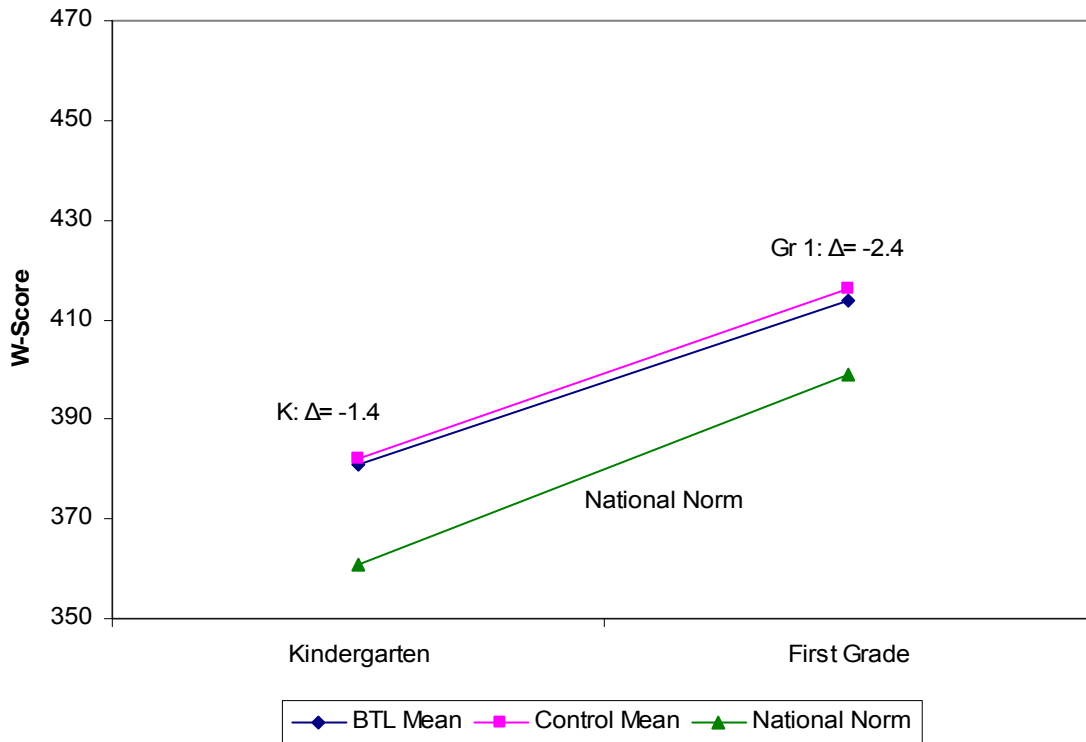


Figure Reads: There are no statistically significant differences between BTL and Control groups in Word Attack at the end of Kindergarten or Grade 1 for cohort 2. There is also no statistically significant BTL versus Control group difference-in-differences between Kindergarten and Grade 1. Both BTL and Control groups are performing above the national norm at the end of Kindergarten and Grade 1.

Figure 3C. Word ID — First Cohort — by ELL Subgroup

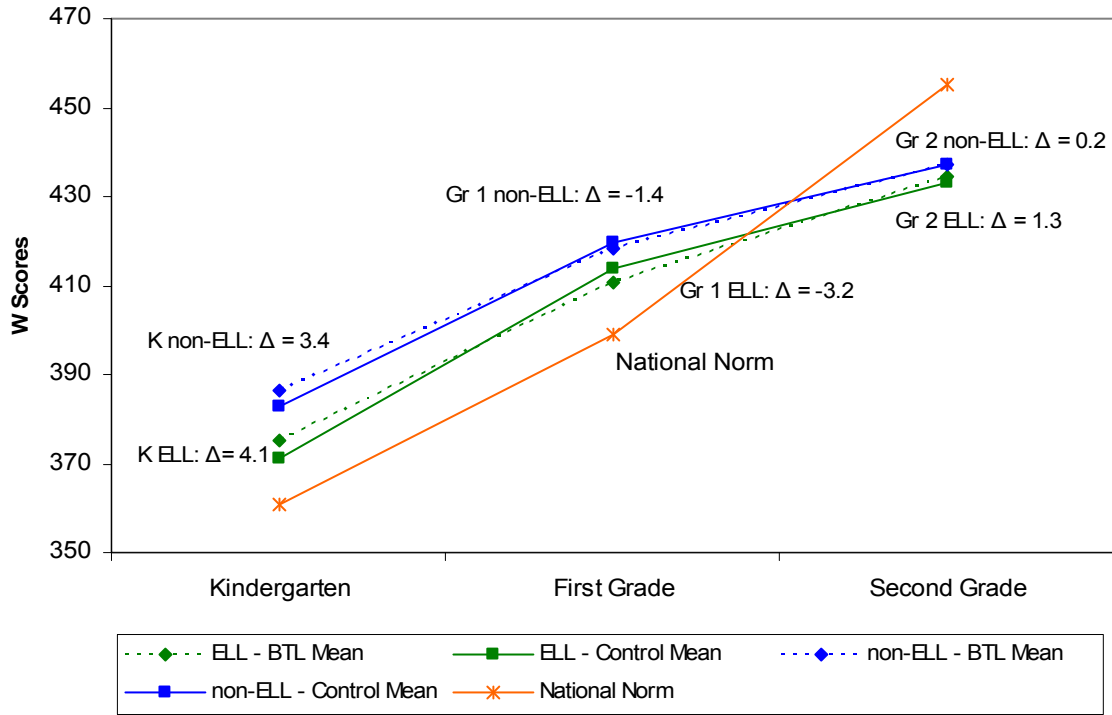


Figure Reads: There are no statistically significant differences between BTL and Control groups within ELL or non-ELL subgroups in Word Identification at the end of Kindergarten, Grade 1, or Grade 2 for cohort 1. Similarly, there are no statistically significant BTL versus Control group differences-in-differences within either subgroup between Kindergarten and Grade 1, between Grade 1 and Grade 2, or between Kindergarten and Grade 2. Both BTL and Control groups are performing above the national norm at the end of Kindergarten and Grade 1, but then fall below the norm by the end of Grade 2.

Figure 3D. Word ID — Second Cohort — by ELL Subgroup

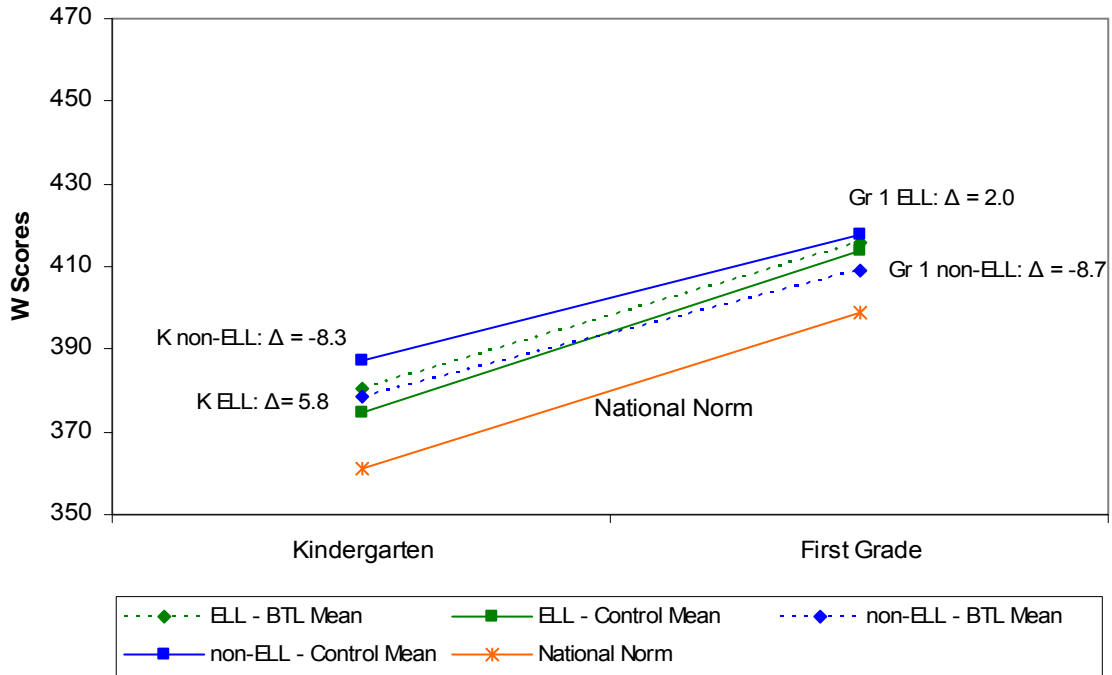


Figure Reads: There are no statistically significant differences between BTL and Control groups within ELL or non-ELL subgroups in Word Identification at the end of Kindergarten or Grade 1 for cohort 2. There is also no statistically significant BTL versus Control group difference-in-differences between Kindergarten and Grade 1 within either subgroup. Both BTL and Control groups within ELL and non-ELL subgroups are performing above the national norm at the end of Kindergarten and Grade 1.

Project Narrative

Program Specific Requirements

Attachment 1:

Title: **Hurtig-narrative** Pages: **0** Uploaded File: **M:\hurtiglab\climbers 2009 annual report\hurtig-narrative.pdf**



U.S. Department of Education
 Grant Performance Report (ED 524B)
 Project Status Chart

OMB No. 1890 - 0004
 Expiration: 10-31-2007

PR/Award #:
R305G040145

Project Narrative (See Instructions. Use as many pages as necessary.)

CLIMBERs Report: June 2009

INTRODUCTION:2

THE CURRICULUM: A Description of Breakthrough to Literacy2

JUSTIFICATION FOR AND GOALS OF THE CURRENT STUDY3

RESEARCH QUESTIONS AND STUDY DESIGN4

IMPACT ANALYSIS – WITH UPDATED ELL STATUS6

Summary6

Research Questions6

Data Collection7

Measures7

Findings9

Classroom Outcomes10

Student Measures12

LONGITUDINAL ANALYSIS, INCLUDING ELL SUBGROUP ANALYSIS14

Results16

Exploratory Analyses of ELL versus Non-ELL Patterns of Change18

WRITING SUPPLEMENT REPORT19

ClimbWrite: An analysis of elicited writing in young children.19

Research Questions20

REFERENCES22

APPENDIX A: ATTRITION RATES25

APPENDIX B: EFFECT SIZE CALCULATION26

APPENDIX C: ESTIMATION MODELS27

APPENDIX D: Three-level Hierarchical Model30

APPENDIX E: POOLED BTL-CONTROL DIFFERENCES ACROSS COHORTS35

INTRODUCTION:

This report presents findings from the Chicago Literacy Initiative: Making Better Early Readers study (CLIMBERs). This five-year study is an experimental evaluation of the impact of Breakthrough to Literacy (BTL), an early literacy curriculum taken to scale in the Chicago Public Schools (CPS), on classroom practice and student literacy outcomes. Abt Associates Inc., along with its research partners at the University of Iowa, is conducting this study, which is supported by a grant from the Institute for Education Sciences at the U.S. Department of Education. The study is focused on gathering data that will inform the status, progress, and future of early learning literacy initiatives in America's third largest school system.

THE CURRICULUM: A Description of Breakthrough to Literacy

Breakthrough to Literacy includes elements shown by reading research to be critical not only to reading instruction, but to changing teachers and improving student performance (Follow-Through, 1977; Berends et al., 2002; Snow, 2002; National Reading Panel Report, 2000). A primary tenet of *Breakthrough to Literacy* is the prediction that embedding explicit and direct oral language instruction and comprehension strategies in meaningful, print-related activities, accompanied by individualized, exploration-based software instruction, will enhance vocabulary knowledge, foster the development of inferential and meta-cognitive processes, and produce long-term improvements in children's word recognition, fluency, and reading comprehension abilities.

Breakthrough to Literacy is built on a conceptual framework for language and literacy instruction that takes seriously the sciences required for changing teacher practice and improving children's performance. It was developed as a result of an intensive research effort begun at the University of Iowa over twenty years ago, to identify experiences that are essential in building the foundations of reading: oral language and vocabulary, phonemic awareness, knowledge of the alphabetic principle, and word recognition skills. The instructional model for the program is based on research indicating that all of these elements are predictive of reading success. While the conceptual framework is systematic and direct, it is flexible and wholly supportive of a child's individual learning needs. It also supports teachers' needs for long term, intensive professional development to translate the program's research philosophy and findings into successful practice (See Snow, 2002).

A key component of the *Breakthrough to Literacy* curriculum is the use of interactive computer modules that gave children control of the auditory, picture, and print support needed to make the critical link between sound, meaning, and print. The program systematically weans each child off of picture and sound cues until each child was reading text alone. The curriculum moves from the systematic instruction in vocabulary, phonemic

awareness, phonics, and alphabet skills, to a similar model of instruction for comprehension of text. The classroom implementation of *Breakthrough to Literacy* helps teachers organize whole-group, small-group, and center activities. *Breakthrough to Literacy* also brings parents into the literacy and language process.

BTL provides two years of professional development for teachers and administrators, as well as in-classroom support for teachers (at least monthly) to assure systematic instruction and fidelity of the implementation process. The stated goal of the *Breakthrough to Literacy* professional development is to make each teacher a diagnostician of each child's needs. This takes many, substantive interactions between the product developers and teachers. In this way, *Breakthrough to Literacy* provides the teachers of the CPS the professional development, support, and tools that are likely to be of great benefit to their students. The program provides literacy coaches with backgrounds in early childhood education to train teachers on the *Breakthrough to Literacy* model and support their efforts to move children systematically yet comfortably from oral language to print. Literacy coaches were expected to visit school sites a minimum of one day per month. In the first year of implementation, literacy coaches conducted three full days of onsite professional development training. Level I training took place just prior to children's introduction to the program. It focused on integrating comprehension strategies and vocabulary development during daily instruction. Level II training was scheduled approximately one month following Level I and focused on the structure of language and on the Explore Words and Explore Alphabet components of the software curriculum. Level III training typically occurred several months later and focused on interpreting software reports and customizing instruction.

In addition to professional development and in-classroom support, *Breakthrough to Literacy* provided teachers with several tools to help them scaffold instruction for individual children. On-going computer reports provided detailed information about each child's progress in the software component of the curriculum. Numerous teacher resources were also provided to complement existing curricula and classroom themes.

JUSTIFICATION FOR AND GOALS OF THE CURRENT STUDY

The current study is designed to test whether *Breakthrough to Literacy* (BTL) is ready for scale-up; that is, whether full implementation of the curriculum, including its professional development package, can produce meaningful effects when taken to scale (i.e., implemented in a large number of sites in as close to "natural" conditions of the real world as possible). The central aim of BTL is to develop children's oral language skills. Thus, the study will demonstrate whether the BTL approach to an 'oral language'-rich curriculum can improve student outcomes. We will examine whether oral language-focused activities increased in the classroom and whether that translated into measurable impacts on students.

The *Breakthrough to Literacy* conceptual framework is one that weaves together the theoretical research-based behavioral predictors of language and literacy with the environmental predictors of reading success (lap reading, books in home, books the child

owns, oral language at home). Current research on best practice directly supports all components of the *Breakthrough to Literacy* intervention: the use of scaffolding (Mooney, 1990; Fountas and Pinnell, 1996; Pressley, 1998; Pearson, 1999; Routman, 2003); the application of basic skills (Foorman, 1998); activities that promote student independence and self-regulation; instructional balance, flexibility, and density (Pressley, 2001); integration of reading and writing (Pressley, 2001); maximizing student academic engagement (Rosenshine, 1976; Pressley, 1998; Pearson, 1999); the use of themes and authentic tasks (Fountas and Pinnell, 1996); ongoing student feedback and teacher assessments (Clay, 1993); and the development of high levels of teacher expertise and skills (Pearson, 1999).

Hart and Risley (1995) documented that children from poverty have exposure to approximately 10 million words by age four, compared to the 50 million exposures of more affluent children. Vocabulary development mirrors this finding. Many of the children from poverty being served in preschool settings have less than a fourth of the cumulative vocabulary words of affluent children. Another significant difference contributing to variations in language and literacy development is the number of hours of lap reading experienced by the children in each of these groups. Children who come to school with rich language and literacy experiences typically have 1000+ hours of lap reading (Adams, 1990). These children not only have richer vocabulary experiences, they know about the purpose and culture of reading. Many children from poverty have little or no lap reading experiences upon preschool entry.

The use of the technology has been one of the great equalizers for the children who come from poverty. It is reliable, provides independent practice, increases listening and attention skills, and provides choice and control to children.

RESEARCH QUESTIONS AND STUDY DESIGN

This impact study is designed to answer the following questions:

1. What are the impacts of BTL on the language development and the precursors of reading fluency of children at the end of preschool, after they have been exposed to BTL for one year?
2. What are the impacts of BTL on teacher instruction and classroom environment at the end of the first year of implementation?
3. What are the impacts of BTL on the language development and early reading comprehension of children at the end of Kindergarten, after they have been exposed to BTL for two years?
4. What are the impacts of BTL on teacher instruction and classroom environment at the end of the second year of implementation?
5. What are the long-term impacts of BTL on reading comprehension and literacy skills of children in grade 3?

We report data that were collected on two successive cohorts of preschoolers and their classrooms. When the first cohort of preschoolers was studied, their teachers were in their first year of implementation, and as the second cohort entered their preschool year most of the teachers were in their second year of implementation. In previous reports, we presented analyses that began to address the first, second, and fourth research questions.

In last year's report we presented the analysis of the two successive cohorts of kindergarten children and their classrooms. That report addressed three research questions about impacts of BTL on kindergarten students, based on a combined sample across the two cohorts:

1. What are the impacts of BTL on the language development and early reading comprehension of children at the end of kindergarten?
2. Do the impacts of BTL differ by students' ELL status?
3. Do kindergarten children who have been exposed to BTL for two years (preschool and kindergarten) outperform children who have been exposed to BTL for one year (kindergarten only)?

A fourth research question addressed impacts on teaching practices for teachers separately after one versus two years of implementing BTL:

4. What are the impacts of BTL on kindergarten teachers' instruction and classroom environment at the end of the first year of implementation? At the end of the second year?

The CLIMBERS project is designed as a randomized cluster design study investigating whether Breakthrough to Literacy (BTL) has an impact on both teacher and student outcomes. In this design, randomization to either a BTL condition or a comparison ("as-is") condition occurs at the school level. Therefore, all studied classrooms and students within a school are in the same condition. The outcome data, whether measured on classrooms or students, are clustered, or nested within school. All analyses reported account for this nesting by using hierarchical linear models (HLM), which estimate treatment impacts while accounting for the fact that the units being measured are not completely independent.

Forty-four schools were recruited and randomly assigned to either implement BTL in their preschool classrooms, or to a comparison group who would continue to implement their current preschool curricula. Prior to randomization, the schools were stratified with regard to their geographic location (North or South Chicago¹) and whether or not the instruction was likely to be offered partly or entirely in Spanish (or another language).² This created four strata or blocks of schools within which schools were randomly assigned to either implement BTL or remain in the "as-is" comparison condition. As mentioned above, the study followed two consecutive cohorts of students in order to have enough statistical power to detect small impacts on students and moderate impacts on teachers.

Last year we also presented preliminary analyses of data from assessments of study children conducted at the end of first grade. This year's report includes analyses of data that includes assessments conducted at the end of the second grade as well as preliminary analyses of kindergarten writing samples.

¹ The North and South designations are based on the Chicago Public Schools administrative areas designations. Administrative Areas 1-9 were designated as North, while Areas 10-18 were designated as South.

² Schools were coded 'yes' based on data from the State Pre-Kindergarten (SPK) program on language of instruction in SPK classrooms. Information was provided by SPK administrators. In addition, using information from the Chicago Public Schools database schools were coded as 'yes' if the proportion of ELL students ("LEP") in a given school was so high (80% or above, according to school report card data) that it was appropriate to assume that at least a portion of the instruction was delivered in Spanish or another language.

IMPACT ANALYSIS – WITH UPDATED ELL STATUS

This report updates the report of June 2008 that presented findings on the impacts of BTL on kindergarten students and teachers at the end of the second year of implementation (2006-07). Findings are updated based on the reclassification of 230 students' ELL status (from non-ELL to ELL), which was recently corrected in our data files. The findings in this report are based on a sample of 43 schools (22 BTL and 21 control), 134 kindergarten classrooms where BTL had been in place for one or two years,³ and two successive cohorts of kindergarten students. The full student sample consists of 3,107 Kindergarten students with spring scores (1,788 BTL and 1,319 C). We begin with a brief summary of the findings. We then present the research questions that guided these analyses and more detailed explanation of methods and results.

Summary

- Scattered findings on teachers/classrooms after 1 year, disappear at end of 2nd year
- No child effects overall or for ELL students

Research Questions

These analyses address three research questions about impacts of BTL on kindergarten students, based on a combined sample across the two cohorts:

1. What are the impacts of BTL on the language development and early reading comprehension of children at the end of kindergarten?
2. Do the impacts of BTL differ by students' ELL status?
3. Do kindergarten children who have been exposed to BTL for two years (preschool and kindergarten) outperform children who have been exposed to BTL for one year (kindergarten only)?

A fourth research question looks at impacts on teaching practices for teachers separately after one versus two years of implementing BTL:

4. What are the impacts of BTL on kindergarten teachers' instruction and classroom environment at the end of the first year of implementation? At the end of the second year?

At this point in the study, pretest, first, and second year outcome data have been collected on two successive cohorts of kindergartners and their classrooms. Some of the students in

³ There are a total of 134 classes total in the student achievement dataset. Students in 108 classes are tested in year 1 and students in 108 classes are tested in year 2. Eighty-two classes are included in both years, while 26 are only in year 1 and 26 others are only in year 2. In the teacher dataset, there are 104 first-year teachers from 2005-2006 and 27 first-year teachers from 2006-2007, for a total of 131 first-year teachers. In addition, there are 80 second-year teachers; 77 of which are included in the first-year dataset. Of the three teachers who were in the second-year dataset but not in the first, all were control teachers. One declined to participate in the first year, but agreed to do so in the second; one was a Pre-K teacher for the first two years of the study and a K teacher for the third year. The last one was incorrectly identified as a Pre-K teacher in year one of the study, but correctly identified as a Kindergarten teacher in year two. The analytic decision was made to include all three of these teachers in the second-year teacher dataset.

BTL kindergarten classrooms were also in BTL preschool classrooms; therefore, we can address the third research question listed above (although note caveats below). When the first cohort of kindergartners was studied, their teachers were in their first year of implementation, and as the second cohort entered their kindergarten year the teachers were in their second year of implementation. The analyses in this memo include data for both cohorts of students, and both years of implementation for teachers.

Data Collection

Below we describe the number of subjects in the sample for these two years of data collection and the measures used.

Sample

Two cohorts of classrooms and students comprise the study sample. Table 1 below displays the numbers of schools, classrooms, and students included in the study sample. During the 2005-06 school year, baseline observation and assessment data were collected in the fall from 108 kindergarten classrooms and 1,156 students. Follow-up data were collected in the spring on 1,574 students.⁴ In the 2006-07 school year, baseline observation data were collected in the (26) kindergarten classrooms that were new to the study, having replaced classrooms that dropped from the study after Year 2. Baseline assessment data were also collected in the fall in 108 kindergarten classrooms on 1,261 students. Follow-up data were collected on 1,533 students in the spring. Across both cohorts, impacts are estimated on data from 134 classrooms and 3,107 students.⁵ For the purposes of analysis, teachers who joined the study in the second year of implementation were included in the sample of teachers who had implemented the curriculum for one year (in both treatment and control groups). Thus, the sample for 2-year teachers is much smaller than that for 1-year teachers (because although they were in the same classrooms in which 1-year teachers had been, the replacement teachers had only been in the study for one year). Attrition rates for both cohorts of children are presented in Appendix A.

Measures

Classroom measures

The measures used for classroom observation were the same as those used in the pre-K classrooms: the QUEST, CLOC, RAP, and Arnett were administered at baseline (fall of the teacher's first year implementing BTL) and the full OMLIT (Snapshot, CLIP, RAP, QUILL, and CLOC) plus the Arnett were administered at follow-up in the spring of the teachers' first and second years in the study (first year only for those teachers who were new to the study in the 2006-07 school year). (Please see Table 2a below.) Descriptions of all of these measures can be found in the reports for the previous years of the study.

Child outcome measures

Each cohort of children was given two tests at baseline (fall) and a battery of four tests at follow-up (spring). All of the tests are individually administered, and all are described below:

⁴ Note that not all students were tested at both time points. In the fall, a sample of students (with parental permission) from each classroom was tested and their scores were used to create a classroom- and a school-level pretest mean for use in analyses. *All* students (with parental permission) in each classroom were tested in the spring.

⁵ Two BTL classrooms (one from each cohort) were eliminated from spring student outcome analyses because the classrooms dropped out of the study at some point during the year. Their fall data, however, were included in calculations of school-level pretest means.

Peabody Picture Vocabulary Test – Third Edition (PPVT-III; Dunn & Dunn, 1997)

This assessment measures receptive vocabulary. It was administered at baseline, in English. It has been normed, and is standardized to have a mean of 100 and a standard deviation of 15.

Preschool Comprehensive Test of Phonological and Print Processing (Pre-CTOPPP, Lonigan, Wagner, Torgesen, & Rashotte, 2002), **Print Knowledge** subtest: This subtest measures early knowledge about written language conventions and form as well as alphabet letters. The test requires children to identify examples of aspects of print, identify letters and written words, point to specific letters, name specific letters, say the sounds associated with specific letters, and identify letters associated with specific sounds. The test was used to provide some parity with earlier tests (this test would be administered three times to a child tested in pre-K and kindergarten—once at the start of each year and once in spring of the pre-K year. The Pre-CTOPPP has now been normed and standardized and is known as the Test of Preschool Early Literacy (TOPEL; Lonigan, Wagner, Torgesen, & Rashotte, 2006). An algorithm was used to convert scores from the Pre-CTOPPP to TOPEL scores to be standardized.

Expressive One-Word Picture Vocabulary Test – Third Edition (EOWPVT; Brownell, 2000)

This test is a standardized, norm-referenced measure of an individual’s English speaking vocabulary. Standard scores are set to have a mean of 100 and standard deviation of 15.

Woodcock Reading Mastery Test – Revised/Normative Update (WRMT-R/NU; Woodcock, 1998)

This test battery measures several important aspects of reading ability. We used three subtests from this battery: **Letter Identification** (at kindergarten only), **Word Identification**, and **Word Attack**. While the Letter Identification subtest is part of the Readiness cluster, the Word Identification and Word Attack subtests combined form the Basic Skills cluster. Each subtest is described briefly below. This test has also been normed

Letter Identification

This test measures the child’s ability to identify uppercase and lowercase letters. The letter forms presented include roman, italic, bold type, serif and sans serif type styles, cursive characters and special type styles. The child is shown the letter and asked to provide the name of the letter.

Word Identification

This test measures the child’s ability to identify isolated words that appear in large type on the stimulus pages. To get credit, the child must produce a natural reading of the word within 5 seconds.

Word Attack

This test requires the child to read either nonsense words or words with extremely low frequency. Nearly all phonemes in the English language are represented in at least one of their major spelling patterns in the items. The test measures the child’s ability to apply phonic and structural analysis skills in order to pronounce words with which s/he may be unfamiliar.

Findings

Descriptive Statistics for Study Schools

The final analytic sample for the kindergarten analyses consists of 43 schools.⁶ As can be seen in Table 2c, the schools serve largely low-income students of color; on average, 89.3% of the students attending these schools are categorized as coming from low-income families, and 85.1% of the students are minorities. The schools in the sample are large; the average school enrollment is 791, with a range of 139 to 1969 students. Mobility rates are also high in this sample of schools, averaging 23.7% and ranging from a low of 5.7% to a high of 56.1%.⁷

Baseline Equivalence of School, Classroom, Teacher and Student Characteristics

The process of random assignment is intended to ensure that the BTL group and control group are statistically equivalent. At the same time, with smaller samples, it is possible that random assignment resulted in two groups that were different on one or more characteristics that could be related to study outcomes. The purposes of analyzing the baseline data collected before the implementation of BTL are to both describe the study sample and identify where the random assignment procedure did not yield equivalent groups. Variables on which the schools, classrooms, teachers, or students differ were then considered as likely candidates for inclusion as covariates in our impact models.⁸

School Characteristics

As seen in the first panel of Table 3, BTL and control schools are the same in terms of the percentage of students from low income families, the total enrollment, and mobility rate. However, BTL schools serve more minority students, on average, than control schools. The .42 of a standard deviation difference is statistically significant at the .05 level. This indicates that it could be an important covariate in the impact models.⁹

Differences in Classroom Characteristics and Baseline Measures

Reported baseline differences here are estimated using a 2-level HLM model (classrooms nested within schools with school-level random error terms). This model employs the sample stratifiers (north, Spanish, north and Spanish interaction) and an observation year indicator as covariates.

The second panel of Table 3 summarizes teacher characteristics, for which there are not statistically significant differences, on average, across BTL and control classrooms. The third panel of Table 3 summarizes classroom characteristics for both the BTL and control classrooms. Prior to randomization, schools were stratified by whether or not instruction was likely to be delivered in Spanish or another non-English language. The baseline equivalence models indicate that, on average, BTL and control schools did not differ any of the 6 classroom and teacher characteristics tested, including percentages of non-English and English speaking students and classroom size.

In terms of classroom interactions at baseline (fourth and fifth panels of Table 3), on the QUEST and the Arnett scores, there were no statistically significant baseline differences between the two groups on any of the QUEST or Arnett measures at $p \leq .05$ level. However,

⁶ This differs from the pre-K sample; one school in the pre-K sample is not in the kindergarten sample and two schools in the K sample are not in the pre-K sample.

⁷ While criteria for eligibility in the study included a mobility rate of less than 40%, data were obtained from the CPS website, were for two years before the kindergarten data collection (and had changed by the time of the data collection). In addition, these school data from the CPS database are based on the school-age population and do not include data on pre-K students, who were in the State Pre-Kindergarten Program.

⁸ Baseline equivalence models were rerun using the updated kindergarten sample of schools.

⁹ For the formula used to calculate effect sizes, see Appendix B.

there is one QUEST subscale – supporting social and emotional development – for which the effect size could be considered substantially meaningful even though the difference is not statistically significant. We do not have a hypothesis for what kind of difference this would make in the outcomes. (This analysis was based on 111 classrooms observed in Fall 2005 and 25 classrooms that were introduced to the study in Fall 2006.)

Differences in Student Measures

As seen in the sixth panel of Table 3, there were no differences in student characteristics, on average, between children in BTL and control schools. Further, students in BTL and control schools on average did not differ significantly on their scores on either the PPVT or the Print Knowledge subscale of the Pre-CTOPP.

The following analyses summarize impacts on kindergarten teachers' instructional behaviors, their classrooms, and their students' early literacy outcomes.

Method

Impacts reported here are estimated through a 2-level HLM model (classrooms nested within schools) which uses sample stratifiers as well as an observation year indicator and percentage of minority students at the school level, which was significantly different at treatment and control schools at baseline. For the Arnett measures, baseline values on these variables are used as covariates. For the model for a typical outcome, see Appendix C.

Classroom Outcomes

Tables 4 and 5 summarize the impact of BTL on OMLIT and Arnett scores after one and two year(s) of implementation in kindergarten classrooms. Analyses of OMLIT measures after one year of implementation are based on observations of 131 classrooms, 104 of which were observed in Spring 2006. The remaining 27 classrooms were those of teachers who replaced teachers who left the study; these teachers were introduced to the study in the 2006-2007 school year and observed in Spring 2007. Arnett analyses are based on the same sample of 131 classrooms. These analyses employ school-level baseline measures as covariates.¹⁰ Analyses of OMLIT and ARNETT measures at the end of two years of BTL implementation are based on observations of 80 classrooms, all of which were observed in Spring 2007.

Teachers after one year of implementing BTL.

As seen in Table 4a, there is a substantial and statistically significant impact of BTL on kindergarten teachers' oral language instruction and on print motivation at the end of one year of implementation. This suggests that teachers are providing students with more opportunities to participate in classroom and instructional activities aimed at improving oral language skills, especially through book-focused activities, a key emphasis of the BTL curriculum. Examples of activities to develop children's oral language include discussions ("sharing," book-related discussions, discussions aimed at building children's vocabulary and concept knowledge), questioning as part of a shared book reading ("dialogic reading"), and

¹⁰ Baseline Arnett measures were available for 120 classrooms. Using these, school-level baseline covariates are created and used in the analyses. 11 classrooms were not observed because teachers replaced other teachers after the fall data collection had taken place.

conversations involving an adult and one or more children focusing on topics other than management.

Impact estimates on two other OMLIT measures are also statistically significant: proportion of time spent in literacy-related activities, and proportion of time spent in computer activities. Teachers in BTL classrooms were encouraged by coaches to weave literacy throughout the day, and both BTL training and coaching provided BTL teachers with explicit explanation and practice in designing activities to reinforce vocabulary and concepts throughout the day. Thus, it is not surprising that observers documented more literacy activities in BTL classrooms. The finding that BTL students spent 4% of their time, on average, using the computer, whereas their counterparts in control schools spent 1% of their time working with computers, is not surprising, given that individualized software instruction is another key component of BTL. Computer-based activities in BTL classrooms are all interactive and include electronic book reading (designed to simulate lap reading); as well as activities focused on letter and word knowledge, phonological sensitivity, vocabulary and concept knowledge, and auditory comprehension skills.

Finally, effect sizes of impacts on three OMLIT measures (print knowledge, ELL students, and literacy resources) are all larger than 0.2, which could be considered substantially meaningful, even though these impacts are not statistically significant.

Teachers after two years of implementing BTL.

As seen in Table 4b, nearly all of the significant impacts shown at the end of one year of BTL vanished at the end of the second year. The only remaining statistically significant impact is the proportion of time spent in computer activities (ES = 4.588, $p \leq 0.0001$), which is dramatic but not surprising given that many control group classrooms did not have or use computers in the classroom.

Another impact is large and nearly statistically significant—the approach to working with ELL students. The effect is more than three-quarters of a standard deviation in *favor* of the control group. This is a troubling trend, but we do not have a hypothesis regarding why implementation of BTL might have affected the classrooms in this way or what else was going on in the district that might have had a positive impact on the way teachers in the control group work with ELLs.

BTL classrooms may also have spent more time in literacy activities (ES = 0.534), but the impact is not statistically significant ($p = .062$).

One reason for the failure to reach statistical significance is that the year-two teacher group is so much smaller—much smaller than the study design called for (thus, the second year analysis is under-powered to detect moderately-sized effects). The one-year group included 131 classrooms, while the two-year group included only 80, so the standard errors are much larger in the two-year analysis, and several sizeable impacts do not attain statistical significance.

Impacts on Arnett scores.

Impact estimates on Arnett measures are presented in Tables 5a and 5b. Although none of these estimates are statistically significant, effect sizes of two measures (*positive* and

permissive) in the analysis of classrooms of one-year teachers and of three measures (*positive*, *detachment*, and the Arnett standardized composite) in the analysis of two-year teachers are larger than 0.25. We do not have a hypothesis about what the effect of these differences might be on the implementation of the curriculum or on classroom outcomes nor of why implementation of the curriculum might affect scores on the Arnett subscales.

Student Measures

Overall Impacts

Method

Table 6 presents the impact of BTL on student test scores; first separately for each cohort and then across two cohorts. The estimates reported here are estimated through a three-level HLM model that represents the nested structure of the data (students nested within classrooms, which are nested within schools). This model employs a-priori selected covariates (sample stratification indicators, cohort indicator, school-level pretest, and percentage of minority students at the school level). For the models used for each outcome, see Appendix C. A number of other covariates (gender, age, ELL status, and percentage of ELL students at the classroom level) were tested for inclusion using the backwards elimination technique.¹¹

Results and Discussion

As seen in the first panel of Table 6, overall there were no statistically significant differences between the BTL group and control group on any of the four tests of early literacy skills at the end of kindergarten across both cohorts. Analyses by cohort, presented in the second and third panels, indicate the same pattern as the overall sample, with no significant differences across BTL and control students for either cohort. These results provide support for the pooling of the two cohorts in analyses. It is also worth noting that not only is there no statistically significant impact of BTL, but there is also no impact of any substantial magnitude that would indicate a trend towards an impact of the curriculum.

We are considering further analyses to explain why the curriculum might not have produced an impact. Our hypotheses fall into three main types: insufficient implementation in the treatment classrooms, high quality instruction in the counterfactual, and insufficient time to make a substantial impact on outcomes.

Insufficient implementation.

The first group includes the fact that teachers did not receive as much support from coaches as the developers suggested. In other instances of implementation of this curriculum, coaches visited teachers every two weeks or no less than monthly, whereas in this implementation, some teachers received five coaching visits during the year, others received fewer. A second implementation problem was that the interactive software component of the curriculum was unevenly implemented; some teachers did it very well, others did not adjust the settings of the computers (rendering them ineffective—students would continue endlessly at the same level), and that technical difficulties with computers prevented teachers from using them to the fullest extent possible. A third problem with the implementation is that the core books for BTL are intended to be used as just one book to be

¹¹ Covariates were retained in the model if they had a p-value of less than or equal to .20 (Budtz-Jorgensen et al., 2006; Maldonado & Greenland, 2006). Results of this process resulted in models that retained all covariates for Expressive One Word Vocabulary and Letter ID. The final model for Word Attack included all covariates except gender, while the final model for Word ID included all covariates except student level ELL status.

used in a unit, and the developers expected that teachers would supplement these books with other trade books. To the extent that this was not done consistently, an extremely limited vocabulary would have been presented to the children. This would explain why, even with more time spent on book-related activities and discussions, if those activities are based on reduced-language books, then the impact on children's vocabulary development will be smaller.

High quality instruction in control classrooms.

Our observations do not lend support to this hypothesis. We observed teachers in the treatment group classrooms spending more time doing activities that should have led to better early literacy outcomes for children.

Insufficient time spent on high-quality instruction to make an impact.

Finally, it is possible that although the difference in key instructional behaviors was statistically significant, it was still not substantial enough to produce a statistically significant impact in children's outcomes. It is possible that if the teachers in the treatment classrooms spend 6% more time on an activity, for example, than teachers in the control classrooms, and if the time observed is the only time in the day that those activities occur (i.e., if what we observed represented 100% of the literacy instruction in the day), then 6% would be 9 minutes. It is possible that 9 additional minutes each day may not be sufficient to produce a statistically significant impact.

Differences in impact by language of child

Table 7 presents the results of subgroup analyses based on ELL status. Once the sample is divided into ELL and non-ELL students, the number of schools drops for the ELL group from 43 to 29 because there are 14 schools in the sample that have no ELL students. No statistically significant differences are found between BTL and control students in the ELL or non-ELL subgroup.

Exposure

In Table 8, we present test score comparisons of BTL students who were exposed to the program one, two, and three years. It is very important to keep in mind that these analyses deviate from the experimental setting as these groups were not determined randomly. In addition, since we do not have enrollment histories on either group of students (BTL or control), our findings are limited by the lack of information about students' school experiences prior to being in our sample. These results, therefore, should be interpreted with caution.

For each outcome measure (Expressive Vocabulary, Word Attack, Word Identification, Letter Identification), we conducted three initial comparisons:

- 1 year BTL exposure versus 1 year control (n = 2,340)
- 2 years BTL exposure versus 2 years control (n = 710), and
- 3 years BTL exposure versus 3 years control (n = 57).

In addition, because the 3 year exposure sample is extremely small, we also combined the 2 and 3 year exposure groups and compared their outcomes.

As shown in Table 8, there were no statistically significant differences between BTL and control students with 1 or 2 years of exposure. For those students with 3 years of exposure, there was a statistically significant difference between BTL and control students in

Expressive Vocabulary and Letter ID (7.934 and 6.792 points, respectively) in favor of BTL students. However, once we combined the 2- and 3-year exposure groups, these differences were no longer statistically significant.

LONGITUDINAL ANALYSIS, INCLUDING ELL SUBGROUP ANALYSIS

Summary

In this analysis, we tested impacts of BTL by examining differences in performance on three language outcomes (expressive vocabulary, decoding, and word reading¹²), at the end of Kindergarten, Grade 1, and Grade 2. There were no statistically significant differences between Treatment and Control at the end of Kindergarten, first, or second grade for either cohort. We also tested impacts on the growth in children's abilities from Kindergarten to Grades 1 and 2. Compared with children in BTL classrooms, children in Control classrooms, on average, showed significantly more growth from Kindergarten to Grade 1 in both decoding and word reading, although the difference in growth was small. The amount of growth between Grade 1 and Grade 2 could only be tested for the first cohort of children in the sample, and there was no significant difference between the growth of children in BTL and Control group classrooms. Once data become available, we will test whether there is any impact of the treatment on children's growth from Grade 1 to Grade 2 for cohort 2 and for the two cohorts combined.

We examined impacts as a function of English Language Learner (ELL) status by conducting separate, parallel analyses for ELL and non-ELL subgroups on the same outcomes. For the ELL children, there was a statistically significant impact favoring BTL over Control on one outcome (decoding) at the end of Kindergarten, although this difference did not persist to the end of Grade 1 (or Grade 2). For the non-ELL subgroup, there was no impact on decoding at the end of Kindergarten, Grade 1, or Grade 2. There were no impacts on expressive vocabulary or word reading at the end of Kindergarten, first, or second grade for either the ELL or non-ELL subgroup.

Note that because we tested so many hypotheses in this analysis, there is a strong possibility that the few impacts detected likely occurred by chance rather than as a result of the intervention.

Measuring Changes in Impacts on Students Over Time: By Cohort and by ELL Subgroup

The overarching research question for these analyses is as follows: Is there any long-term effect of BTL (exposure) in Kindergarten on student language and literacy outcomes at the end of first or second grade? To understand this, we examine student growth on language and literacy outcomes from Kindergarten through first and second grade. The data for this

¹² Measures used were: English One-Word Picture Vocabulary Test (EOWPVT, expressive vocabulary), Woodcock Reading Mastery Test: Word Attack (WRMT: Word Attack; decoding), and WRMT: Word Identification (Word ID; word reading).

analysis are standardized test scores through the end of second grade for the first cohort of Kindergarten students in our sample and through the end of first grade for the second cohort of Kindergarten students. In this analysis, therefore, we can examine change for cohort 1 from the spring of Kindergarten to the spring of second grade and from the spring of Kindergarten to the spring of first grade for cohort 2.

A second research question asks about differential impacts for ELL versus non-ELL children.

As shown in Table 9, the total sample consists of 7,382 students (4,557 ELL students and 2,825 non-ELL students across all three grades) in 43 schools (22 assigned to the Treatment condition, and 21 assigned to the Control condition). A detailed breakdown by cohort, grade, ELL status, and treatment status is presented below.

The analysis is based on a difference-in-differences approach, in which we test to see whether there are statistically significant differences between BTL and Control students *at* each time point (spring of Kindergarten, spring of Grade 1, and spring of Grade 2) and then test whether the differences at each time point are different *across* time points (and whether these differences-in-differences are statistically significant). We refer to the difference at each time point as Δ (Δ = mean score of Treatment group – mean score of Control group), while the difference across time points is the “difference-in-differences” or relative growth, calculated as $\Delta_2 - \Delta_1$ (subtracting the value for time 1 from the value for time 2). All analyses control for Kindergarten pre-test scores at the school level. Separate models are run for three outcomes: Expressive One-Word Picture Vocabulary Test (EOWPVT), WRMT: Word Identification (Word ID), WRMT: Word Attack. Scores are Normal Curve Equivalents for EOWPVT, and W-scores for Word ID and Word Attack.¹³

We use a hierarchical model to conduct these analyses because our data are nested. In this case, the data have three levels of nesting: time is nested within individual students who are, in turn, nested within schools. See Appendix D for the three-level hierarchical linear growth model that is used in analyses.

Because we have different numbers of test points for the two cohorts of students, the results of these analyses are reported separately by cohort. For completeness, we present results from analyses that estimate pooled BTL-Control group differences across the two cohorts for the sample overall and for ELL and non-ELL subgroups in Appendix E. In general, these pooled cohort differences display patterns very similar to those of the by-cohort differences presented below. Once we have data at the end of second grade for cohort 2, we will test for differences across the cohorts from Kindergarten to Grade 2. If none are found, we will combine the cohorts for all analyses. If there are differences, we will continue to present the cohort results separately.

¹³ Test scores have been converted into metrics that allow for comparison on a continuous, equal-interval scale spanning multiple grades. For the subtests of the WMRT-R, W scores, which are the result of a mathematical transformation of raw scores into Rasch-based ability scores, are used. The range of W scores for the WMRT-R tests extends from a low of 340 to a high of 606. Because W scores are not available for the EOWPVT, Normal Curve Equivalent (NCE) scores, which range from 1 to 99 (standard deviation = 21.09), are used.

Results

Descriptive statistics for the sample are presented by measure in Table 10. Findings for each outcome are presented by cohort and then by ELL subgroup by cohort. Findings are presented graphically in Figures 1A – 4D below; estimates used to create these graphs are presented in Tables 11-13

Trends in the estimated effect of BTL exposure on each outcome, by cohort and for ELL and non-ELL subgroups by cohort, are summarized below.

Caveat on Interpretation

When interpreting these results, it is important to keep the “multiple comparisons” or “multiple hypothesis testing” issue in mind. Multiple hypothesis testing is a problem because as the number of tests conducted increase, the probability of making a Type I error (i.e. finding a difference when in fact there is none) increases. More specifically, when 20 hypothesis tests are performed in a setting where there are no true differences between two conditions at the usual $p < 0.05$ significance level, we expect to have one false significant difference (i.e., one difference by chance). Note that we conducted 81 tests when examining the BTL-control differences (excluding the exploratory analyses), and one would expect to observe 4 significant differences (two favoring the BTL group and two favoring the control group) by chance alone (when there were no “true differences”). Hence, we suggest that the three statistically significant findings presented below should be interpreted with caution as they may well just be due to chance.

Expressive Vocabulary

At the three time points indicated by the Δ in the graphs (Kindergarten, Grade 1, and Grade 2), for cohort 1, the difference between the average expressive vocabulary (EOWPVT) test scores of BTL and Control students is 0.7, 0.8, and -0.3 points in Kindergarten, first, and second grade respectively (Figure 1A). None of these differences is statistically significant at the $p \leq 0.05$ level. When we tested for “differences in differences,” we found that there were no statistically significant differences in the amount of growth for BTL versus Control between Kindergarten and Grade 1, Grade 1 and Grade 2, or between Kindergarten and Grade 2. The same results hold for cohort 2 (Figure 1B), with no statistically significant differences between BTL and Control students at either Kindergarten or first grade and no statistically significant difference between the BTL versus Control differences between Kindergarten and Grade 1. For both cohorts, BTL and Control groups are performing below the national average of 50 NCE at all time points.

For the comparison between the ELL and non-ELL subgroups by cohort (see Figures 1C and 1D), there were no statistically significant differences were found between BTL and Control groups for either ELL or non-ELL students at the end of K, Grade 1 or Grade 2 for Expressive Vocabulary. There were also no statistically significant “differences in differences” between grades within either subgroup for either cohort. For both cohorts, BTL and Control groups within ELL and non-ELL subgroups are performing below the national average of 50 NCE at all time points.

Decoding

As indicated in Figure 2A, the difference in the average word attack test scores (WRMT-R: Word Attack) of BTL and Control students is 2.1, -0.02, and 0.3 points in Kindergarten, first, and second grade respectively. None of these differences is statistically significant at the $p \leq 0.05$ level. When we tested for “differences in differences,” we found that there were no

statistically significant differences between the BTL versus Control differences from Kindergarten to Grade 1, from Grade 1 to Grade 2, or from Kindergarten to Grade 2. Both BTL and Control groups are performing above the national norm at the end of Kindergarten and Grade 1, but then fall below the norm by the end of Grade 2. For cohort 2 (Figure 2B), BTL students outperformed Control students in Kindergarten by 1.9 W-score points, while Control students outscored BTL students in first grade by 1.5 W-score points. However, the difference between BTL and Control students is not statistically significant at either Kindergarten or first grade. On the other hand, we do find a statistically significant “difference in differences.” That is, the difference in the BTL versus Control differences between Kindergarten and Grade 1 is 3.4 W-score points¹⁴ (effect size = 0.21) and it is statistically significant at the $p = 0.03$ level, most likely because there was a difference in Kindergarten in favor of the BTL group (a positive number) that then became a difference in first grade in favor of the Control group (a negative number), in other words, the Treatment group not only failed to maintain their lead from Kindergarten but actually fell behind the Control group at Grade 1, so their relative growth rate was negative. Results from the assessments to be conducted in Spring 2009 will reveal whether the pattern of cohort 1 is repeated in cohort 2. Both BTL and Control groups in cohort 2 were performing above the national norm at the end of Kindergarten and Grade 1.

Figures 2C and 2D present results for ELL and non-ELL subgroups for cohorts 1 and 2. For cohort 1, we find the same results as above, i.e., no statistically significant differences between BTL and Control groups within ELL or non-ELL subgroups at the end of K, Grade 1 or Grade 2 and no statistically significant “difference in differences” between grades. Both BTL and Control groups within each subgroup are performing above the national norm at the end of Kindergarten and Grade 1, but then fall below the norm by the end of Grade 2. For cohort 2, we find a slightly different pattern from that described above, i.e., there is a statistically significant difference in favor of the Treatment group at the end of Kindergarten within the ELL subgroup (4.5 W score points; effect size = 0.31; $p \leq 0.05$), although this difference does not persist until the end of Grade 1. For non-ELLs, there are no statistically significant differences between BTL and Control groups at the end of Kindergarten or first grade. There is not a statistically significant “difference in differences” between Kindergarten and Grade 1 for either ELL or non-ELL subgroups. Both BTL and Control groups within ELL and non-ELL subgroups are performing above the national norm at the end of Kindergarten and Grade 1.

Word Reading

As indicated in Figure 3A, the differences in the average word identification test scores (WRMT-R: Word Identification) of BTL and Control students are 2.8, -3.4, and -0.4 points in Kindergarten, first, and second grade respectively. None of these differences is statistically significant at the $p \leq 0.05$ level. When we tested for “differences in differences,” we found that there **was** a statistically significant difference in differences between BTL versus Control scores from Kindergarten to Grade 1 (6.2 W-score points, effect size = 0.20, $p = 0.02$). Once again, this is likely because although BTL students outperformed Control students in Kindergarten, Control students outperformed BTL students in first grade. By the end of Grade 2, BTL students had almost caught up to Control students. Both BTL and Control groups are performing above the national norm at the end of Kindergarten and Grade 1, but then fall below the norm by the end of Grade 2. For cohort 2 (Figure 3B), there were no statistically significant differences between BTL and Control students at either Kindergarten

¹⁴ The calculation, $\Delta_2 - \Delta_1$, $(-1.5 - 1.9 = -3.4)$.

or first grade and no statistically significant difference between the BTL versus Control differences between Kindergarten and Grade 1. Both BTL and Control groups are performing above the national norm at the end of Kindergarten and Grade 1.

Similar results were obtained for tests of impacts of BTL for the ELL and non-ELL subgroups by cohort (see Figures 3C and 3D). In both cohorts, no statistically significant differences were found between BTL and Control groups within ELL or non-ELL subgroups at the end of K, Grade 1 or Grade 2 for word reading. There were also no statistically significant “differences in differences” between grades within either subgroup for either cohort. For cohort 1, both BTL and Control groups within ELL and non-ELL subgroups are performing above the national norm at the end of Kindergarten and Grade 1, but then fall below the national norm by the end of Grade 2. For cohort 2, both BTL and Control groups within ELL and non-ELL subgroups are performing above the national norm at the end of Kindergarten and Grade 1.

Exploratory Analyses of ELL versus Non-ELL Patterns of Change

An examination of Figures 2C-2D and 3C-3D indicates that, for two of the outcomes – Word Attack and Word ID – patterns of change for each outcome are similar for the ELL and non-ELL subgroups and across BTL and Control groups within each cohort. With or without BTL, ELLs and non-ELLs in our sample are performing similarly on these two outcomes. This is an interesting finding in and of itself. In addition, at the last data collection point, the difference between study children’s scores and the national norm is not statistically significant on both outcomes, irrespective of treatment or ELL status. (The last data collection point is Grade 2 for cohort 1 and Grade 1 for cohort 2.)

When we tested Word Attack outcomes for cohort 1, we found that although there were statistically significant differences between ELL and non-ELL children in the BTL group and statistically significant differences between ELL and non-ELL children in the Control group at the end of Kindergarten, these differences no longer existed at the end of Grade 1 or Grade 2. This means that although ELL children in both BTL and Control groups in cohort 1 are starting out significantly lower than the non-ELL children in their same treatment group in Kindergarten, they catch up with the non-ELL children by the end of Grade 1 and perform similarly to the non-ELL children through the end of Grade 2. For cohort 2, there are no statistically significant differences between ELL and non-ELL children in the BTL group or between ELL and non-ELL children in the Control group at the end of Kindergarten or Grade 1.

For Word ID, we found a similar pattern for cohort 1, with statistically significant differences between ELL and non-ELL children within both the BTL and Control groups at the end of Kindergarten, but no statistically significant differences at the end of Grade 1 or Grade 2. For cohort 2, we found a statistically significant difference at the end of Kindergarten between ELL and non-ELL children in the Control group, but not in the BTL group. As with Word Attack, however, this difference is no longer statistically significant at the end of Grade 1 within either group (BTL or Control).

Interestingly, a very different pattern is evident when we examine the expressive vocabulary outcome (Figures 1C and 1D). Especially in cohort 1, it appears that ELLs and non-ELLs, independent of their treatment status, are performing differently. While both ELL and non-ELL subgroups are performing below the national norm (NCE = 50), the non-ELL subgroup

appears to be approaching the norm from Kindergarten to Grade 2, while the ELL subgroup's performance remains flat, even losing a little ground from Kindergarten to Grade 2.¹⁵ As a result, the gap between the two groups widens over time. For cohort 2, the pattern is similar but there is less of a gap between the ELL and non-ELL children.

For cohort 1, we found statistically significant differences between ELL and non-ELL children in the BTL group and statistically significant differences between ELL and non-ELL children in the Control group at the end of Kindergarten, Grade 1 and Grade 2. This means that ELL children in both BTL and Control groups are starting out significantly lower than the non-ELL children in their same treatment group in Kindergarten and remain at a significantly lower level through Grade 2. For cohort 2, the same pattern held at the end of Kindergarten and Grade 1. Spring 2009 data from cohort 2 will show whether cohort 2 data continue to follow the same pattern as cohort 1 data.

In general, these findings indicate a shortfall in ELL children's vocabulary knowledge compared with non-ELL children's vocabulary knowledge (whether in BTL or Control), whereas ELL and non-ELL children differed only in Kindergarten, if at all, in their performance on tests of decoding and word reading, with ELL children catching up to non-ELL children in Grade 1 (cohorts 1 and 2) and Grade 2 (cohort 1). In addition, on tests of decoding and word reading, both ELL and non-ELL children's scores were close to the grade level norm, while in vocabulary, both ELL and non-ELL children performed statistically significantly below the national norm, with only non-ELL children making some progress toward it by the end of Grade 2.

WRITING SUPPLEMENT REPORT

ClimbWrite: An analysis of elicited writing in young children.

The Breakthrough to Literacy (BTL) curriculum that is being assessed in the large randomized cluster study in the Chicago public schools includes writing as an essential classroom practice. When implemented with fidelity, the BTL curriculum integrates listening, reading, writing and speaking, around a focus book in whole group, small group and individual instruction. Repeated opportunities for exposure to language and cognitive activities within the context of familiar daily practices set the stage for building both the deep and surface structures of language.

The Chicago Study (CLIMBERS) offers us the opportunity to examine a large sample of writing at the Pre-K and Kindergarten levels to more precisely examine the relationship of early reading and writing. At the time we submitted the original grant we did not have a tool

¹⁵ The non-ELL subgroup's initial mean scores on EOWPVT are slightly more than one standard deviation below the mean. By the end of Grade 2, both BTL and Control non-ELLs' scores are less than half of a standard deviation below the mean. For the ELL subgroup, initial mean scores are 1.4 standard deviations below the national norm and remain at approximately the same level through Grade 2.

to systematically examine the writing of children in Pre-K and Kindergarten, and so did not include an analysis of children’s writing and its relationship to other emerging literacy skills. We are now in a position to be able to also systematically examine the impact of the BTL curriculum on the emergent writing skills of children. The supplement has allowed us to more fully assess the impact of BTL as it is implemented on a large scale in Chicago by allowing us to look at both reading and writing.

Utilizing the Picture prompts developed for this project we collected a pilot sample of writing samples during the spring of 2008. Teachers asked kindergarten students to complete a writing task using picture prompts and standard instructions. We examined the writing samples that kindergarteners in the Chicago and West Des Moines public schools produced in response to this task. Writing samples from Chicago were collected from classrooms identified by Abt as classrooms (BTL and non-BTL classrooms) in which teacher were observed engaging the children in significant writing activities (based on OMLIT: Quill observation scores). The West Des Moines samples came from classrooms that were using the BTL curriculum. Each sample was scanned into a specialized database developed at the University of Iowa. The samples were then coded utilizing a computer assisted coding scheme.

Research Questions

In this preliminary analysis, we examined the children’s overall productivity and spelling strategies to address two research questions:

1. How do students vary in the length of written text, prevalence of incorrectly spelled words, and spelling strategies at the end of kindergarten?
2. How does students’ writing change during the second half of the kindergarten year, in terms of the length of written text, prevalence of incorrectly spelled words, and spelling strategies?

RQ1. Writing and Spelling at the End of Kindergarten

We examined descriptive statistics and bivariate correlations for total word count, incorrect spelling, and spelling strategies in kindergarten writing collected at the end of the kindergarten year. The sample included 165 students in 9 classrooms in Chicago and West Des Moines public schools. We estimated a two-level model to test for differences in kindergarten writing outcomes between students and classrooms. We included a school district indicator variable in the Level-2 model to test for school district differences in writing outcomes. For each writing outcome, we estimated the following model:

Student-level (Level-1): $Y_{ij} = \beta_{0j} + \varepsilon_{ij}$, where $\varepsilon_{ij} \sim N(0, \sigma^2)$

Classroom-level (Level-2): $\beta_{0j} = \gamma_{00} + \gamma_{01}DesMoines_j + \mu_{0j}$, where $\mu_{0j} \sim N(0, \tau_{00})$

Findings indicated that students from the West Des Moines sample differed substantially from students from the Chicago sample (See Table 14). The average writing sample length among Chicago students was 11.7 words, while the average length among West Des Moines students was 30.1 words ($p < .0001$). Approximately 11% of the variation in the length of students’ writing was between classrooms ($p < .10$), indicating that children demonstrated more productivity in their writing in some Chicago classrooms than in other Chicago classrooms, and children’s writing was more productive in some West Des Moines classrooms than in others. By writing longer text, students in West Des Moines used more invented spelling (average of 11.7 incorrectly spelled words in West Des Moines) than in

Chicago (average of 4.1 incorrectly spelled words; $p < .0001$), which provided them more opportunity to employ spelling strategies reflecting correct consonant and vowel sounds. Only 5% of the students in Chicago produced written text of 21 or more words and a high incidence of invented spelling; however, 53% of students in the West Des Moines sample produced this type of written text. The site and classroom variations in students' writing can be attributed to variations in the level of implementation of writing opportunities in the site and classroom curriculum.

RQ2. Growth in Writing and Spelling Strategies during Kindergarten

Students completed the writing task at three time points (February, April, and May) in the West Des Moines sample (71 students in 3 classrooms). We estimated the following linear growth model to investigate change in students' writing productivity, incorrect spelling, and spelling strategies over the second half of kindergarten.

Within-student (Level-1): $Y_{ij} = \pi_{0j} + \pi_{1j}(TIME)_{ij} + \varepsilon_{ij}$ where $\varepsilon_{ij} \sim N(0, \sigma^2)$

Between-student (Level-2): $\pi_{0j} = \gamma_{00} + \gamma_{01}(Classroom1)_{0j} + \gamma_{02}(Classroom2)_{0j} + \mu_{0j}$

$\pi_{1j} = \gamma_{10} + \mu_{1j}$

where $\begin{pmatrix} \mu_{0j} \\ \mu_{1j} \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{pmatrix} \right]$

Findings indicated that students' writing productivity and incidence of invented spelling increased over the second half of the kindergarten year (See Table 15). In February, the average length of written text was 16 to 18 words in two of the W. Des Moines classrooms and was 27 words in a third classroom. Students' writing increased in length by an average of 4.7 words at each subsequent time point that writing was elicited ($p < .001$). In writing longer texts, students had greater opportunity for employing their invented spelling strategies; the number of incorrectly spelled words in students' writing increased by an average of 1.4 words at each time point ($p < .001$). In particular, students employed advanced spelling strategies (i.e., correct consonant and vowel sounds in misspelled multisyllabic words) with an increasing proportion of incorrectly spelled words over the second half of kindergarten. On average, in February, students employed advanced spelling strategies in 15% of their incorrectly spelled words, and this percentage increased by an average of 5 percentage points at each subsequent time point ($p < .001$). By the end of kindergarten, students were using advanced spontaneous spelling strategies in 25% of their incorrectly spelled words, suggesting that students were using more multi-syllabic words and more advanced spelling strategies over the second half of kindergarten.

A complete set of writing samples were obtained at three time points during the 2008-2009 academic year (Fall, Winter, Spring) from the kindergarten classrooms in a sub-sample of the Chicago public schools in order to apply the analytic techniques we developed to look at growth.

REFERENCES

- Achilles, C.M. & Finn, J.D. (2000). Should class size be a cornerstone for educational policy? *The CEIC review*, 91 (2), 15, 23.
- Adams, M.J. (1990). *Beginning to read: thinking and learning about print*. MIT Press: Cambridge, MA.
- Alexander, K. L., & Entwisle, D. R. 1988. Achievement in the first two years of school: Patterns and Processes. *Monographs of the Society for Research in Child Development* 53(2, Serial No. 218)..
- Bloom, B.S., Englehart, M.B., Furst, E.J., Hill, W.H., and Krathwohi, O.R. (1956). *Taxonomy of educational objectives: The classification of educational goals*. Handbook 1: The cognitive domain. New York: Longman.
- Bloom, H. S. (2003). *Sample Design for an Evaluation of the Reading First Program*. MDRC Working Papers on Research Methodology, for the U.S. Department of Education.
- Bredenkamp, S. & Copple, C., (eds. 1997). *Learning to read and write: developmentally appropriate practices for young children* (revised edition). National Association for the Education of Young Children: Washington D.C.
- Budtz-Jorgensen E, Keiding N, Grandjean EM, Weihe P. Confounder selection in environmental epidemiology. Assessment of health effects of prenatal mercury exposure. *Annals of Epidemiology* 2006.
- Clay, M.M. (1993). *An observation survey of early literacy achievement*. Portsmouth, NH: Heinemann.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.
- Dunn, L. & S. Kontos (1997). *What Have We Learned about Developmentally Appropriate Practice?* Vol. 52 / Num. 5
- Foorman, B.R., Francis, D.J., Fletcher, J.M., Schatschneider, C., Mehta, P. (1998). The role of instruction in learning to read: Prevention reading failure in at-risk children. *Journal of Education Psychology*, 90, 37-55.
- Fountas, I.C. & Pinnell, G.S. (1996). *Guided reading: Good first teaching for all children*. Portsmouth NH: Heinemann.
- Gardner, H. (1991). *The Unschooled Mind: How Children Think & How Schools Should Teach*. New York: Basic Books.
- Hart and Risley (1995), *Meaningful Differences in the Everyday Experience of Young American Children*. Baltimore: Paul H. Brookes Publishing Co, Inc.

Helburn, S. W. (1995). *Cost, Quality and Child outcomes in Child Care Centers, Technical Report*. Denver, Department of Economics, Center for Research in Economic and Social Policy, University of Colorado, Denver.

Jorm, A.F. & Share, D.L. (1983). Phonological recoding and reading acquisition. *Applied Psycholinguistics*, 4, 103-147.

Juel, C. (1994). *Learning to read and write in one elementary school*. Springer Verlag: New York.

Konald, T.R.; Juel, C.; McKinnon, M. (1999). Building an integrated model of early reading acquisition. Center for the Improvement of Early Reading Achievement CIERA (Online). Available: <http://www.ciera.org/library/reports/inquiry-1/1-003/1-003.pdf>

Maldonado G, Greenland S. Simulation study of confounder-selection strategies. *Am J Epidemiol* 2006; 138:923-936.

Mooney, M. (1990). *Reading to, with, and by children*. Richard D. Owen Publishers, Inc.: Katonah, NY.

National Assessment of Educational Progress. (1999). Major findings from the NAEP reading assessment (Online), Available: http://nces.ed.gov/nationsreportcard/reading/read_new_findings.asp

National Association for the Education of Young Children. (1996). *Phonics and whole language learning: A balanced approach to beginning reading* (Online). Available: <http://www.ericps.crc.uiuc.edu/npin/respar/texts/home/phonics.html>

National Reading Panel. (2000) *Teaching Children to Read: An Evidence Based Assessment of the Scientific Research Literature on Reading and Its Implications for Reading Instruction*.

Pearson, P.D. & Raphael, T.E. (1999). Toward an ecologically balanced literacy curriculum. In Gambrell, L.B., Morrow, L.M., Neuman, S.B., & Pressley, M. (Eds.) *Best practices in literacy instruction* 22-33. The Guilford Press: New York.

Perfetti, C. A., & Sandak, R. (2000). *Literacy Education*. In N. J. Smelser, & P. B. Baltes (Eds.), *International Encyclopedia of the Social and Behavioral Sciences*. Elsevier.

Piaget, J. (1970). Piaget's theory. In P. Mussen (Ed.), *Handbook of child psychology* (3rd ed.) (Vol. 1, pp. 703-732). New York: Wiley.

Pressley, M. (1998). *Reading instruction that works: The case for balanced teaching*. The Guilford Press: New York.

Pressley, M.; Allington, R.L.; Warton-McDonald, R.; Block, C.C.; Morrow, L.M. (2001). *Learning to read: Lessons from exemplary first grade classrooms*. The Guilford Press: New York.

RAND Education and RAND Science and Technology Policy Institute for the U.S. Department of Education's Office of Educational Research and Improvement. (2002) *Reading for Understanding: Toward an R&D Program in Reading Comprehension*. RAND Distribution Services.

Raudenbush, S.W. & Bryk, A.S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. 2nd ed. Thousand Oaks, CA: Sage Publications.

Rosenshine, B. (1976) Classroom instruction. In N. L. Gage (Ed.), *The psychology of teaching methods*. Seventy-fifth yearbook of the National Society for the Study of Education. Chicago: University of Chicago Press.

Routman, (2003) *Reading Essentials: The Specifics You Need to Teach Reading Well*. Heinemann.

Shadish, W.R., Cook, T.D., & Cambell, D.T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin Company.

Shaywitz, S.E. and Shaywitz. (1994) *Measuring and analyzing change*. In: G. Reid Lyon (ED.), Frames of reference for the assessment of learning disabilities: New views on measurement issues. Baltimore:MD

Snow, C.E., Burns, M.S., & Griffin, P. (Eds.). (1998) *Preventing reading difficulties in young children*. Washington D.C.: National Academy Press.

Vygotsky, L.S. (1978). *Mind in society* (M.Cole, V. John-Steiner, S. Scribner, and E. Souberman, eds.). Cambridge, Mass.: Harvard University Press.

Wagner, R.K. & Torgeson, J.K. (1987) The nature of phonological processing and its causal role in the acquisition of reading skills. *Psychological Bulletin*, 101, 192-212.

Whitebook, M., Howes, C., & Phillips, D. (1990). *Who Cares? Child Care Teachers and the Quality of Care in America. Final Report of the National Child Care Staffing Study*. Oakland, CA: Child Care Employee Project.

APPENDIX A: ATTRITION RATES

Attrition rates for each cohort are presented below. Attrition rates across treatment and control groups are similar for both cohorts.

Cohort 1: Attrition Estimates: Spring 2005 PreK to Spring 2006 K or Spring 2006 PreK*			
	Treatment	Control	Total
Stayed in the K sample	157 (55%)	187 (53%)	344 (54%)
Stayed in the PreK sample	45 (16%)	45 (13%)	90 (14%)
Left the sample	82 (29%)	124(34%)	206(32%)
Total	284(100%)	356 (100%)	640 (100%)

*Children who were 3 years old in pre-K in 2004-2005 were in pre-K again in 2005-2006.

Cohort 2: Attrition Estimates: Spring 2006 PreK to Spring 2007 K			
	Treatment	Control	Total
Stayed in the sample	293 (49%)	226 (44%)	519 (47%)
Left the sample	305 (51%)	282 (56%)	587 (53%)
Total	598 (100%)	508 (100%)	1106 (100%)

APPENDIX B: EFFECT SIZE CALCULATION

Based on Cohen (1988), we defined d (the effect size) as the difference between the means, $M_t - M_c$, divided by standard deviation, σ_c , of the control group.

$$d = M_t - M_c / \sigma_c$$

In our analyses, M_c is the estimated mean for the control group, adjusting for covariates in the models. M_T is derived by adding the impact estimate to M_c , so $M_t - M_c$ is the same as the impact estimate.

APPENDIX C: ESTIMATION MODELS

Classroom level outcomes are estimated using a 2-level HLM model (classrooms nested within schools) that employs sample stratifiers as well as an observation year indicator and percentage of minority students at the school level, which was significantly different between treatment and control schools at baseline. For the Arnett measures, baseline values on these variables are used as covariates. The model for a typical outcome is as follows:

Level 1 (classroom):

$$Y_{jk} = \pi_{0k} + \pi_{1k} Year_{jk} + \varepsilon_{jk}$$

where:

Y_{jk} is the outcome measure for classroom j in school k ,

$Year_{jk}$ is the observation year indicator of the outcome measure, Y_{jk} . In particular, it equals one if the measure is from Spring 07 and zero otherwise.

π_{0k} is the mean of the outcome measure in classroom j , in school k , and

ε_{jk} is the residual associated with student i , in classroom j , in school k . All error terms are assumed to be normally distributed and independent of one another.

Level 2 (school):

$$\begin{aligned} \pi_{0k} = & \gamma_{00} + \gamma_{01} BTL_k + \gamma_{02} north_k + \gamma_{03} Spanish_k + \gamma_{04} N_k * S_k \\ & + \gamma_{05} Minority_k + \gamma_{06} Y_k^{pre} u_k \end{aligned}$$

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

where:

BTL_k is an indicator variable equaling 1 for BTL schools and 0 for control schools,

$Minority_k$ is the school level percentage of minority students,

Y_k^{pre} is the baseline value of the outcome measure in school k . Note that this covariate is only used for Arnett outcomes,

γ_{00} is the grand mean of the outcome measure for the average control school,

γ_{01} is the estimated impact of BTL, or the difference between the average control school and the average BTL school,

γ_{02} , γ_{03} , and γ_{04} are the effects associated with the stratification indicators used during the random assignment process, and their interaction, to control for the study design, γ_{05} is the effect associated with school-level percentage of minority students, γ_{10} is the estimated overall difference in the outcome measure in Spring 06 and Spring 07, γ_{06} is the effect associated with the baseline measure, and u_k is the residual associated with school k .

Student level outcomes are estimated using a three-level HLM model that represents the nested structure of the data (students nested within classrooms, which are nested within schools). This model employs a-priori selected covariates (sample stratification indicators, cohort indicator, school-level pretest, and percentage of minority students at the school level). A number of other covariates (gender, age, ELL status, and percentage of ELL students at the classroom level) were tested for inclusion using the backwards elimination technique.¹⁶ Results of this process resulted in the inclusion of the following covariates for each of the four outcomes:

Outcome	Covariates Retained
Expressive One Word Vocabulary	Gender, age, student level ELL status, classroom level percentage of ELL students
Work Attack	Age, student level ELL status, classroom level percentage of ELL students
Word ID	Gender, age, classroom level percentage of ELL students
Letter ID	Gender, age, student level ELL status, classroom level percentage of ELL students

The general HLM model for student outcomes can be represented as follows:

Level 1 (student):

$$Y_{ijk} = \pi_{0jk} + \pi_{1jk} Cohort_{ijk} + \sum_{n=2}^N \pi_{njk} Cov_{ijk}^n + \mathcal{E}_{ijk}$$

where:

Y_{ijk} is the outcome measure for student i , in classroom j , in school k ,
 $Cohort_{ijk}$ is cohort indicator for student i , in classroom j , in school k . In particular, it equals one if the student is from the second cohort of kindergartners and zero otherwise,
 Cov_{ijk}^n is the n^{th} student-level covariate that is tested for inclusion using backwards elimination. These covariates are gender, age, and ELL status,

¹⁶ Covariates were retained in the model if they had a p-value of less than or equal to .20 (Budtz-Jorgensen et al., 2006; Maldonado & Greenland, 2006).

π_{0jk} is the mean of the outcome measure in classroom j , in school k , and

ε_{ijk} is the residual associated with student i , in classroom j , in school k . All error terms are assumed to be normally distributed and independent of one another.

Level 2 (classroom):

$$\pi_{0jk} = \beta_{00k} + \beta_{01k} \text{PercentELL}_{jk} + r_{jk}$$

$$\pi_{njk} = \beta_{n0k} \text{ for } n = 2, 3, \dots, N$$

where:

β_{00k} is the mean of the outcome measure in school k ,

PercentELL_{jk} is the percentage of ELL students in classroom j in school k . This covariate is also tested for inclusion using backwards elimination,

r_{jk} is the residual associated with classroom j , in school k .

Level 3 (school):

$$\beta_{00k} = \gamma_{000} + \gamma_{001} \text{BTL}_k + \gamma_{002} \text{north}_k + \gamma_{003} \text{Spanish}_k + \gamma_{004} N_k * S_k + \gamma_{005} \text{pretest}_k + u_k$$

$$\beta_{n0k} = \gamma_{n00} \text{ for } n = 2, \dots, N$$

where:

γ_{000} is the grand mean of the outcome measure for the average control school,

BTL_k is an indicator variable equaling 1 for BTL schools and 0 for control schools,

γ_{001} is the estimated impact of BTL, or the difference between the average control school and the average BTL school,

γ_{002} , γ_{003} , and γ_{004} are the effects associated with the stratification indicators used during the random assignment process, and their interaction, to control for the study design,

γ_{005} is the effect associated with school-level pretests, and

u_k is the residual associated with school k .

APPENDIX D: Three-level Hierarchical Model

For purposes of illustration, we present here a three-level hierarchical linear growth model. The first level of our model represents each student's development in the form of an individual growth trajectory, whose parameters then become the outcome variables in the between-student level of our model. Those parameters can vary among students and schools as a function of child- or teacher/school-level variables. The within-student or repeated observations model (Level-1) is denoted as follows:

$$(1) Y_{ij} = \pi_{0ij} + \pi_{1ij} \text{Grade1}_{ij} + \pi_{2ij} \text{Grade2}_{ij} + \alpha_{ij}$$

where,

Y_{ij} = an observed outcome measure (e.g., EOWPVT score) for student i in school j at time t ;

Grade1_{ij} = the first grade indicator for the record of student i in school j at time t , such that when t = spring of first grade, $\text{Grade1}_{ij}=1$ and $\text{Grade1}_{ij}=0$ for other times.

Grade2_{ij} = the second grade indicator for the record of student i in school j at time t , such that when t = spring of second grade, $\text{Grade2}_{ij}=1$ and $\text{Grade2}_{ij}=0$ for other times.

π_{0ij} = the score (intercept) for student i in school j at time t = spring of Kindergarten;

π_{1ij} = the growth rate parameter (rate of change from K to first grade) for student i in school j ;

π_{2ij} = the growth rate parameter (rate of change from K to second grade) for student i in school j ;

α_{ij} = a random error term for student i in school j at time t . The distribution of within-student errors is assumed to be normal, with mean=0 and variance= ϕ .

In the between-student model (Level-2), variation in the growth parameters π_{ij} can be modeled as a function of student background characteristics (e.g. gender and age) and cohort. In this second level of the model, the π_{ij} are random outcome variables. A between-student model can be formulated for the intercept, π_{0ij} , and growth rate parameters, π_{1ij} and π_{2ij} , as follows:

$$(2) \pi_{0ij} = \beta_{00j} + \beta_{01j} \text{Cohort2}_{ij} + \sum_{q=1}^Q \beta_{0(q+1)} (X_{qij} - \bar{X}_q) + \sum_{q=1}^Q \beta_{0(q+Q+1)} (X_{qij} - \bar{X}_q) \text{Cohort2}_{ij} + \delta_{0ij}$$

$$(3) \pi_{1ij} = \beta_{10j} + \beta_{11j} \text{Cohort2}_{ij} + \sum_{q=1}^Q \beta_{1(q+1)} (X_{qij} - \bar{X}_q) + \sum_{q=1}^Q \beta_{1(q+Q+1)} (X_{qij} - \bar{X}_q) \text{Cohort2}_{ij} + \delta_{1ij}$$

$$(4) \pi_{2ij} = \beta_{20j} + \sum_{q=1}^Q \beta_{2q} (X_{qij} - \bar{X}_q) + \delta_{2ij}$$

where,

π_{0ij} = the score for student i in school j , defined in Kindergarten (i.e., the level-1 intercept);

π_{1ij} = the growth rate parameter (from K to first grade) for student i in school j ;

π_{2ij} = the growth rate parameter (from K to second grade) for student i in school j ;
 $Cohort2_{ij}$ = the indicator for the second cohort students. $Cohort2_{ij} = 1$ if student i is a second cohort student and $Cohort2_{ij} = 0$ otherwise;
 $X_{qij} - \bar{X}_q$ = the q^{th} measured background characteristic for student i in school j .
 Each of these Q characteristics are centered at the overall mean value;
 β_{00j} = the average covariate adjusted score across all *first cohort* students in school j in Kindergarten;
 β_{01j} = the difference in the average covariate adjusted Kindergarten scores of *first and second cohort* students in school j ;
 $\beta_{0(q+1)}$ = a vector of $q = 1 \dots Q$ regression coefficients, which capture the effects of the X_i predictor variables on the student-specific outcome score in Kindergarten averaged across all *first cohort* students at all schools;
 $\beta_{0(q+Q+1)}$ = a vector of $q = 1 \dots Q$ regression coefficients, which capture the difference in the effects of the X_i predictor variables on the *first and second cohort* student-specific outcome score in Kindergarten at all schools;
 β_{10j} = the average covariate adjusted Kindergarten to first grade score growth across all *first cohort* students in school j ;
 β_{11j} = the difference in the average covariate adjusted Kindergarten to first grade score growth of *first and second cohort* of students in school j ;
 $\beta_{1(q+1)}$ = a vector of $q = 1 \dots Q$ regression coefficients, which capture the effects of the X_i predictor variables on the student-specific Kindergarten to first grade score growth averaged across all *first cohort* students at all schools;
 $\beta_{1(q+Q+1)}$ = a vector of $q = 1 \dots Q$ regression coefficients, which capture the difference in the effects of the X_i predictor variables on the *first and second cohort* student-specific Kindergarten to first grade score growth at all schools;
 β_{20j} = the average covariate adjusted Kindergarten to second grade score growth across all *first cohort* students in school j . Note that equation [4] does not include the $Cohort2_{ij}$ indicator as the sample does not yet include any second grade observations of second cohort students.
 β_{2q} = a vector of $q = 1 \dots Q$ regression coefficients, which capture the effects of the X_i predictor variables on the student-specific Kindergarten to second grade score growth averaged across all *first cohort* students at all schools;
 δ_{0ij} = random error term indicating the deviance between the score for student i in school j in Kindergarten and the average score for students in school j in Kindergarten, after controlling for the vector of student-level characteristics;
 δ_{1ij} = random error term indicating the deviance between the Kindergarten to first grade growth rate for student i in school j and the average growth rate for students in school j , after controlling for the vector of student-level characteristics;
 δ_{2ij} = random error term indicating the deviance between the Kindergarten to second grade growth rate for student i in school j and the average growth rate for students in school j , after controlling for the vector of student-level characteristics;

The between-student error terms are assumed to have a joint normal distribution, with mean 0, variances σ_0^2 , σ_1^2 , and σ_2^2 and covariance $\sigma_{10} = \sigma_{01}$, $\sigma_{20} = \sigma_{02}$, and $\sigma_{12} = \sigma_{21}$. These assumptions are expressed as follows¹⁷:

$$\begin{bmatrix} \delta_{0ij} \\ \delta_{1ij} \\ \delta_{2ij} \end{bmatrix} \sim \mathbf{N} \left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} & \sigma_{02} \\ \sigma_{10} & \sigma_1^2 & \sigma_{12} \\ \sigma_{20} & \sigma_{21} & \sigma_2^2 \end{bmatrix} \right).$$

The impact of the BTL treatment in Kindergarten on student growth is tested in a school-level model. In the school-level model (Level-3), variation in school mean scores of children in Kindergarten (β_{00j} and β_{01j}) and variation in school mean growth rates (β_{10j} , β_{11j} , and β_{20j}) can be modeled as a function of BTL treatment status and other school characteristics.¹⁸ The school-level model is specified as follows:

$$(5) \beta_{00j} = \gamma_{000} + \gamma_{001} BTL_j + \sum_{k=1}^K \gamma_{00(k+1)} (Z_{kj} - \bar{Z}_k) + \nu_{00j}$$

$$(6) \beta_{01j} = \gamma_{010} + \gamma_{011} BTL_j + \sum_{k=1}^K \gamma_{01(k+1)} (Z_{kj} - \bar{Z}_k) + \nu_{01j}$$

$$(7) \beta_{10j} = \gamma_{100} + \gamma_{101} BTL_j + \sum_{k=1}^K \gamma_{10(k+1)} (Z_{kj} - \bar{Z}_k) + \nu_{10j}$$

$$(8) \beta_{11j} = \gamma_{110} + \gamma_{111} BTL_j + \sum_{k=1}^K \gamma_{11(k+1)} (Z_{kj} - \bar{Z}_k) + \nu_{11j}$$

$$(9) \beta_{20j} = \gamma_{200} + \gamma_{201} BTL_j + \sum_{k=1}^K \gamma_{20(k+1)} (Z_{kj} - \bar{Z}_k) + \nu_{20j}$$

where,

β_{00j} = the average covariate-adjusted score across all first cohort students in school j in Kindergarten;

β_{01j} = the difference in the average covariate adjusted Kindergarten scores of *first and second cohort* of students in school j ;

β_{10j} = the average covariate adjusted Kindergarten to first grade score growth across all *first cohort* students in school j ;

β_{11j} = the difference in the average covariate adjusted Kindergarten to first grade score growth of *first and second cohort* of students in school j ;

¹⁷ For Word Attack, all three student level random error terms are used. For Expressive Vocabulary, two of these terms (δ_{1ij} and δ_{2ij}) and for Word Identification, one term (δ_{2ij}) are not included in the analysis because after controlling for student and school characteristics and including school-level random errors, there was not much variation left at the student level that would be captured by those particular student-level random errors.

¹⁸ In these analyses, we include the stratification variables, “North,” “Spanish,” and the interaction of North and Spanish, as well as a school-level pretest score from the fall of Kindergarten as covariates at the school level. For EOWPVT, we use the mean school-level standardized PPVT score, while for Word Attack and Word Identification, we use the mean school-level raw score on the Print Knowledge subtest of the Pre-CTOPPP.

β_{20j} = the average covariate adjusted Kindergarten to second grade score growth across all *first cohort* students in school j .
 BTL_j = treatment status, where $BTL = 1$ for a school assigned to the BTL treatment and $BTL = 0$ for a school assigned to the Control group;
 $Z_{kj} - \bar{Z}_k$ = school covariates for school j , each centered around the grand mean.
 Y_{000} = the grand or overall covariate-adjusted mean value of *first cohort* students' Kindergarten scores in Control schools;
 Y_{001} = the average treatment effect on school mean Kindergarten scores of *first cohort* students;
 $Y_{00(k+1)}$ = a vector of K regression coefficients indicating the effect of each school characteristic on school mean Kindergarten scores of *first cohort* students;
 Y_{010} = the overall covariate-adjusted difference in the school mean Kindergarten scores of *first and second cohort* students in Control schools;
 Y_{011} = the average treatment effect on the covariate-adjusted difference in the school mean Kindergarten scores of *first and second cohort* students;
 $Y_{01(k+1)}$ = a vector of K regression coefficients indicating the effect of each school characteristic on the difference in the school mean Kindergarten scores of *first and second cohort* students;
 Y_{100} = the overall covariate-adjusted school mean Kindergarten to first grade score growth of *first cohort* students in Control schools;
 Y_{101} = the average treatment effect on the covariate-adjusted school mean Kindergarten to first grade score growth of *first cohort* students;
 $Y_{10(k+1)}$ = a vector of K regression coefficients indicating the effect of each school characteristic on school mean Kindergarten to first grade score growth of *first cohort* students;
 Y_{110} = the overall covariate-adjusted difference in the school mean Kindergarten to first grade score growth of *first and second cohort* students in Control schools;
 Y_{111} = the average treatment effect on the covariate-adjusted difference in the school mean Kindergarten to first grade score growth of *first and second cohort* students;
 $Y_{11(k+1)}$ = a vector of K regression coefficients indicating the effect of each school characteristic on the school mean Kindergarten to first grade score growth of *first and second cohort* students;
 Y_{200} = the overall covariate-adjusted school mean Kindergarten to second grade score growth of *first cohort* students in Control schools;
 Y_{201} = the average treatment effect on the covariate adjusted school mean Kindergarten to second grade score growth of *first cohort* students;
 $Y_{20(k+1)}$ = a vector of K regression coefficients indicating the effect of each school characteristic on school mean Kindergarten to first grade score growth of *first cohort* students;
 U_{00j} = the error term capturing the j^{th} school's deviation from the grand mean value of *first cohort* students' Kindergarten scores, controlling for school-level covariates and treatment effects;
 U_{01j} = the error term capturing the j^{th} school's deviation from the grand mean difference in the school mean Kindergarten scores of *first and second cohort*, controlling for school-level covariates and treatment effects;

- $u_{10,j}$ = the error term capturing the j^{th} school's deviation from the grand mean
 Kindergarten to first grade score growth of *first cohort* students, controlling for school-level covariates and treatment effects;
- $u_{11,j}$ = the error term capturing the j^{th} school's deviation from the grand mean
 difference in the Kindergarten to first grade score growth of *first and second cohort* students, controlling for school-level covariates and treatment effects;
- $u_{20,j}$ = the error term capturing the j^{th} school's deviation from the grand mean
 Kindergarten to second grade score growth of *first cohort* students, controlling for school-level covariates and treatment effects;

The school-level error terms are assumed to have a joint normal distribution, with mean 0 and the following variance and covariance terms:

$$\begin{bmatrix} u_{00,j} \\ u_{01,j} \\ u_{10,j} \\ u_{11,j} \\ u_{20,j} \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{00}^2 & \tau_{0001} & \tau_{0010} & \tau_{0011} & \tau_{0020} \\ \tau_{0100} & \tau_{01}^2 & \tau_{0110} & \tau_{0111} & \tau_{0120} \\ \tau_{1000} & \tau_{1001} & \tau_{10}^2 & \tau_{1011} & \tau_{1020} \\ \tau_{1100} & \tau_{1101} & \tau_{1110} & \tau_{11}^2 & \tau_{1120} \\ \tau_{2000} & \tau_{2001} & \tau_{2010} & \tau_{2011} & \tau_{20}^2 \end{bmatrix} \right)$$

Note that linear combinations of the coefficients in equations [5] through [9] yield various overall means and treatment effects of interest. More specifically:

- Y_{000} = mean of the first cohort Kindergarten scores in the Control group;
 $Y_{000} + Y_{001}$ = mean of the first cohort Kindergarten scores in the treatment group;
 Y_{001} = treatment effect for the first cohort students in Kindergarten;
 $Y_{000} + Y_{010}$ = mean of the second cohort Kindergarten scores in the Control group;
 $Y_{000} + Y_{001} + Y_{010} + Y_{011}$ = mean of the second cohort Kindergarten scores in the Treatment group;
 $Y_{001} + Y_{011}$ = treatment effect for the second cohort students in Kindergarten;
- $Y_{000} + Y_{100}$ = mean of the first cohort first grade scores in the Control group;
 $Y_{000} + Y_{100} + Y_{001} + Y_{101}$ = mean of the first cohort first grade scores in the Treatment group;
 $Y_{001} + Y_{101}$ = treatment effect for the first cohort students in the first grade;
 $Y_{000} + Y_{100} + Y_{010} + Y_{110}$ = mean of the second cohort first grade scores in the Control group;
 $Y_{000} + Y_{100} + Y_{010} + Y_{110} + Y_{001} + Y_{011} + Y_{101} + Y_{111}$ = mean of the second cohort first grade scores in the Treatment group;
 $Y_{001} + Y_{011} + Y_{101} + Y_{111}$ = treatment effect for the second cohort students in the first grade;
- $Y_{000} + Y_{200}$ = mean of the first cohort second grade scores in the Control group;
 $Y_{000} + Y_{200} + Y_{001} + Y_{201}$ = mean of the first cohort second grade scores in the Treatment group;
 $Y_{001} + Y_{201}$ = treatment effect for the first cohort students in the second grade.

APPENDIX E: POOLED BTL-CONTROL DIFFERENCES ACROSS COHORTS

In this appendix, we present results from analyses that estimate pooled BTL-Control group differences across the two cohorts. The model used in these analyses is a three-level HLM model (observations nested within students nested within schools), but unlike the one presented earlier, it does not include the cohort indicator or its interactions with the other covariates. These results are presented in Table E.1.

Notice that these pooled cohort differences display very similar patterns to the by-cohort differences presented earlier. The only statistically significant BTL-Control difference is observed in the average **word attack** test scores at Kindergarten within the ELL subgroup (3.3 W score points, effect size = 0.22; $p = 0.04$)—a small positive impact (favoring BTL) for ELLs. In addition, we found two small statistically significant “difference-in-differences”, both favoring the Control group: from Kindergarten to Grade 1 in the whole sample (-2.89 W score points, effect size = -0.09; $p = 0.01$) and within the ELL subgroup (-3.93 W score points, effect size = -0.13; $p = 0.01$).

Table E.1: BTL versus Control Group Differences Pooled Across Two Cohorts

	Expressive Vocabulary			Word Attack			Word ID		
	Overall	ELL	Non-ELL	Overall	ELL	Non-ELL	Overall	ELL	Non-ELL
Kindergarten									
BTL Mean	22.88	21.15	29.19	455.59	454.7	456.41	381.75	378.71	385.07
Control Mean	22.78	21.29	29.43	453.58	451.4	455.51	380.76	373.79	386.7
BTL-Control Diff.	0.09 (0.927)	-0.14 (0.925)	-0.24 (0.890)	2.01 (0.098)	3.3* (0.014)	0.9 (0.568)	0.99 (0.675)	4.92 (0.068)	-1.63 (0.578)
1st Grade									
BTL Mean	25.9	22.49	33.67	469.85	470.3	469.55	412.32	412.41	414.18
Control Mean	27.15	24.04	34.75	470.74	470.94	470.97	415.37	413.34	417.83
BTL-Control Diff.	-1.26 (0.302)	-1.55 (0.341)	-1.08 (0.576)	-0.89 (0.565)	-0.63 (0.720)	-1.43 (0.486)	-3.05 (0.322)	-0.93 (0.810)	-3.66 (0.328)
2nd Grade									
BTL Mean	27.76	23.58	37.37	479.28	481.31	477.26	434.11	436.06	432.69
Control Mean	29.31	24.67	39.85	479.25	479.78	478.71	435.14	433.67	435.57
BTL-Control Diff.	-1.54 (0.293)	-1.09 (0.618)	-2.48 (0.228)	0.03 (0.990)	1.52 (0.425)	-1.45 (0.621)	-1.04 (0.764)	2.39 (0.569)	-2.87 (0.472)
Difference-in-									

Differences									
K to 1 st Grade	-1.35 (0.115)	-1.42 (0.175)	-0.84 (0.475)	-2.89* (0.011)	-3.93* (0.007)	-2.32 (0.155)	-4.04 (0.065)	-5.85 (0.065)	-2.03 (0.452)
K to 2 nd Grade	-0.29 (0.831)	0.47 (0.808)	-1.41 (0.363)	0.92 (0.696)	2.15 (0.287)	-0.03 (0.992)	2.02 (0.541)	3.33 (0.462)	0.78 (0.834)
1 st Gr. to 2 nd Gr.	-1.64 (0.168)	-0.95 (0.595)	-2.24 (0.100)	-1.98 (0.358)	-1.78 (0.274)	-2.35 (0.378)	-2.03 (0.449)	-2.53 (0.475)	-1.24 (0.682)
N –Total	6887	4307	2580	6886	4309	2577	6888	4309	2579
N –BTL	4003	2453	1550	4000	2453	1547	4002	2453	1549
N –Control	2884	1854	1030	2886	1856	1030	2886	1856	1030

Note: Total Ns indicate students with post-test scores for each outcome at every time point.