UPSTART Program Evaluation

Year 6 Program Results

Submitted to the Utah Office of Education
January, 2016

Evaluation and Training Institute
100 Corporate Pointe, Suite 387
Culver City, CA 90230
www.eticonsulting.org

All correspondence should be directed to:
Jon Hobbs, Ph.D.
jhobbs@eticonsulting.org

Table of Contents

Executive Summary

Utah Preparing Students Today for a Rewarding Tomorrow (UPSTART) is a home-based preschool program developed and provided by Waterford to prepare preschool children for school and future academic success. The Evaluation and Training Institute (ETI), the external evaluator of UPSTART since 2009, has prepared this report for the Utah State Office of Education (USOE) to document UPSTART's impact in its 6th year of implementation (Cohort 6/2014-2015 program year).

The evaluation of UPSTART's sixth cohort moved from using a nonequivalent control group seen in previous years to a pre-test/post-test design with a statistically matched control group to assess the program's impact on developing children's early literacy skills in preschool. Our research findings cover two areas, how the program was implemented and what types of impacts it had on children's literacy.

Program Implementation

Enrollment has increased across the state of Utah and UPSTART has reached families in both rural and urban areas. Half of the children enrolled in Year 6 lived in families with incomes less than 200% of the federal poverty level and the majority of the children were White (83%) and English speaking (92%). UPSTART enrollment increased from 1,577 children in Year 5 to 5,091 children in Year 6, an increase of over 300 percent.

Findings about UPSTART usage are summarized below:

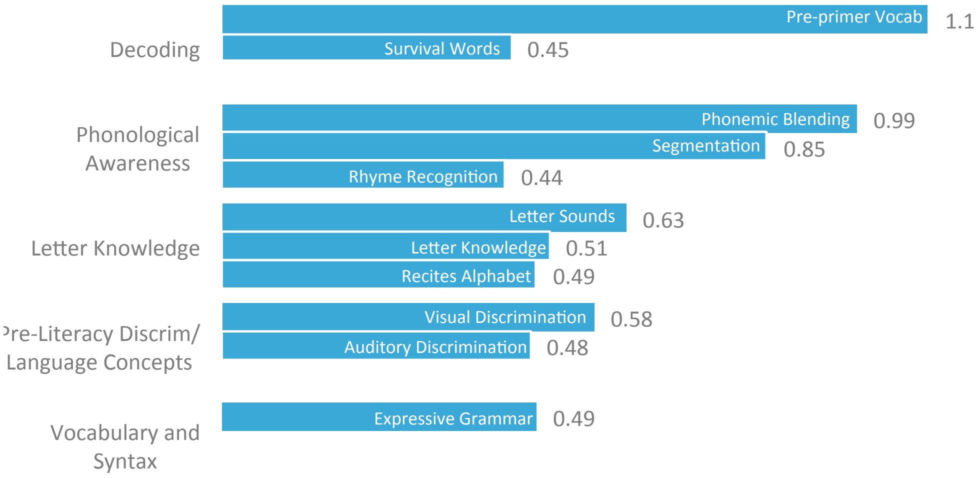
- The average level of UPSTART curriculum usage in Year 6 was 67 hours.
- The UPSTART graduation rate with Cohort 6 was 92%, slightly lower than the graduate rate of 94% in Cohort 5.
- UPSTART graduates had an average program use of 70 hours.
- UPSTART curriculum usage was significantly positively correlated with literacy skills measured by the Bader and Brigance post-tests.

Impacts on Literacy

Results from effect size and growth score analyses indicated that participation in UPSTART had a strong impact on children's emerging literacy skills. Children enrolled in UPSTART produced large effects (ES = .81) compared to control children on the Brigance composite, an instrument that measures decoding skills, letter knowledge, vocabulary and syntax, and pre-literacy discrimination. Similarly, UPSTART participants experienced large effects (ES = .95) on the Bader, an instrument assessing children's phonological awareness. Detailed findings by literacy construct are summarized below:

UPSTART had a medium to large impact on a majority of early literacy domains

Effect Size Estimates



UPSTART had a strong impact on children's **word decoding** skills:

- Children participating in UPSTART had significantly higher post-test scores on decoding pre-primer words (large ES = 1.1) and reading survival sight words (medium ES = .45)
- UPSTART children had stronger growth scores on reading pre-primer vocabulary and survival sight words subtests compared to children who were not enrolled in the program.

Children's **phonological awareness** abilities were significantly improved as a result of UPSTART:

- UPSTART students had significantly higher phonemic blending skills (large ES = .99), phoneme segmenting skills (large ES = .85), and facility with rhyme recognition (medium ES = .44)
- Compared to control children, students participating in UPSTART had significantly higher increases from the pre-test to the post-test on all three phonological awareness subscales.

Students who participated in UPSTART experienced a moderate improvement on their **letter knowledge** skills:

- UPSTART children had medium effects in their learning how to recite (ES = .49), identify (ES = .51), and sound out (ES = .63) letters of the alphabet.
- Compared to control students, UPSTART participants showed significantly stronger growth rates in learning how to pronounce letter sounds and identifying lowercase letters.

UPSTART participants showed a moderate impact on **pre-literacy discrimination and language concepts**:

- UPSTART had a medium effect on children's ability to discriminate between different shapes, letters, and words (ES = .58) as well as their ability to distinguish whether or not two words sounds the same (ES = .48).
- Children in UPSTART had stronger growth scores on their auditory discrimination of words when contrasted with children not enrolled in UPSTART.

Impact of the UPSTART program on children's **vocabulary and syntax skills** was mixed:

- The UPSTART program had a medium effect (ES = .49) on expressive grammar
- UPSTART did not have significant effects on receptive or expressive vocabulary.
- Children enrolled in UPSTART did not have significantly different growth rates on vocabulary and syntax subscales when compared to control children.

Recommendations

The UPSTART program shows continued success at helping preschool age children develop literacy skills and prepare for school. These outcomes would have specific benefits to at-risk children, whose families struggle with poverty and other issues, and often lack the resources to help their children develop the literacy skills needed to succeed in school.

Enrollment increased in the 2014-2015 program year, which resulted in more families benefitting from the computer-based instruction. However, slightly less C6 students were classified as graduates when compared to previous cohorts (92% vs 94% for C5, for example). In addition to a slight drop in graduation rates, average program usage dropped approximately 4 hours when compared to the previous year (C5 average use was 71 hours vs. 67 hours of average use for C6). While slight, these reductions need to be monitored to be sure it is not a trend due to the demands of increased enrollment.

Given the success at improving literacy test scores, we recommend that the state continue providing the UPSTART program to children. The strong program effects support wide-scale implementation across at-risk preschool populations. In addition, we recommend that the program vendor work with the evaluator and USOE staff to monitor program implementation carefully and to be sure that increased enrollment does not erode graduation or usage rates, two key areas for ensuring strong student literacy achievement and future program success.

Preface

The Utah State Office of Education (USOE) hired the Evaluation and Training Institute (ETI), a non-profit research and consulting firm, to conduct a multi-year evaluation of the UPSTART program to determine the effectiveness of the home-based preschool program in academically preparing children for school success. This report includes evaluation results for UPSTART's sixth year of implementation during the 2014-2015 program year, hereafter referred to as Cohort 6 (C6).

The 2014-2015 program year saw the program's use increased to reach more families than in any previous cohort. This expansion was due in part to empirical evidence from previous positive program evaluation findings (Evaluation and Training Institute 2011, 2012a, 2012b, 2013, 2014). As the program scaled-up, the evaluation had to be adapted to accommodate larger numbers of program students, and higher stakes related to greater resource allocation for the program. While the scale and stakes increased, our research objectives remained constant: we continued to evaluate the program's *impact on developing children's early literacy skills in preschool* to help the state and stakeholders determine the benefits from participating in the program.

We enhanced the established evaluation design to meet a higher level of accountability for the Cohort 6 students, and ensure that the program resources were having a positive impact on school readiness. The Cohort 6 evaluation included a balanced one-to-one match of treatment and control students. While requiring a larger sample size, the matching process enhanced our ability to detect treatment effects and, in general, improve the accuracy of the evaluation results.

In addition to documenting program effects on early literacy skills, other objectives included: (a) documenting the extent to which participants used the computerized curriculum; (b) establishing the relationship between curriculum usage and literacy outcomes; and (c) documenting the program's completion or "graduation" rate as measured by the proportion of the enrollment that met the criteria established for usage of the program's curriculum.

A note to readers: This report is intended for multiple audiences, from technical reviewers to high-level policy makers. For those seeking to know the big-picture findings without technical details, the executive summary contains information about the program's impacts on preschool children. The main body of the report contains detailed results and technical information about the findings.

Introduction

UPSTART Program Description

Utah Preparing Students Today for a Rewarding Tomorrow (UPSTART) is a pilot project established by the Utah state legislature that uses a home-based education technology approach to develop the school readiness skills of preschool children. In its sixth year of operation during the 2014-15 school year, the project's implementation contractor – the Waterford Institute – enrolled 5,091 preschool children and provided them with an adaptive program of computer-based early literacy instruction to prepare them academically for kindergarten. The 5,091 children enrolled in the sixth year cohort, hereafter referred to as Cohort 6 (C6), participated in UPSTART from September 2014 through June 2015. Cohort 6 is the largest since the program's rollout.

The UPSTART software uses adaptive lessons, digital books, songs, and activities to deliver early literacy content. The reading skills taught by the Waterford Early Learning Program at Level 1 of the curriculum[1] include:

- Phonological Awareness: phonemic segmenting and blending
- Phonics: letter name knowledge, sound knowledge, and word reading
- Comprehension and Vocabulary: vocabulary knowledge
- Language Concepts: oral reading fluency

Children are encouraged to use the UPSTART program for 15 minutes a day, 5 days a week and families are provided with parental resources and technical support from Waterford customer service representatives.

Evaluation Research Questions

Our evaluation is framed by research questions. We hypothesized that if UPSTART has no effect on improving early literacy skills, then the preschool children who participated in UPSTART – the treatment group – would be expected to perform at the same level as a comparison control group (children who were not exposed to UPSTART) on post-test measures of early literacy development at the beginning of Kindergarten. If UPSTART does have an effect on improving early literacy, then the treatment group should perform significantly better than the control group on the post-test at the beginning of Kindergarten. For purposes of triangulation, we also wanted to take a slightly different look at the data by examining growth rates from pre-test to post-test. If UPSTART shows stronger literacy growth rates, then the treatment group would be expected to show greater gain scores (post-test score minus pre-test score) relative to the comparison group on the various literacy subtests and total test scores.

With respect to concerns for school readiness, our research questions for the C6 evaluation study were as follows:

1. Do UPSTART students have better early literacy skills at kindergarten compared to control group students?

[1] Level One is the beginning point of the curriculum where the preschool child begins as a nonreader and is introduced to skills designed to teach the child to read.

2. Do UPSTART students show stronger literacy growth rates from preschool to kindergarten compared to control group students?

In the preschool analysis, the outcomes of interest were measures of early literacy skills relevant to emerging readers such as phonological awareness, letter recognition, and letter sound knowledge and vocabulary development. Results for research questions 1 and 2 are presented in the **UPSTART Program Impacts on Literacy** section of the report.

The Utah State Office of Education (USOE) and the Utah State Legislature were also interested in outcomes related to the implementation of UPSTART. Research questions along this line included:

3. What was the extent of UPSTART curriculum usage in terms of the amount of exposure per participant, as measured in minutes or hours of instruction per week?

4. What percent of the participants completed the full implementation program (i.e., "graduated" as defined by the Waterford Institute)?

5. How does the level of UPSTART curriculum usage relate to reading readiness outcomes?

Data for research questions 3 and 4 were obtained from records maintained by the Waterford Institute and are answered in this report by descriptive statistics. The answer to Research Question 5 was derived from the relationship between exposure to the computer-assisted program of instruction (measured by program records documenting minutes of computer usage for each enrolled student) and the measured literacy outcomes of interest. Results for research questions 3 through 5 are presented in the **UPSTART Program Implementation** section of the report.

Research Methods

The following section presents information about the research methods used to conduct the evaluation, including: the research design, creation of treatment (UPSTART students) and control (non-UPSTART students) samples, outcome measures, and ETI's data collection and analyses procedures.

Research Design

To evaluate the impact of the UPSTART program, we collected literacy data for a "treatment group" of UPSTART participants and a comparison "control group" of students who did not participate in the program. We collected pre-test and post-test data on children in each group over a 12-month interval during the year prior to enrollment in Kindergarten. Due to the legislative mandate that all children interested in enrolling in the program be allowed to participate, children could not be randomly assigned to groups, which resulted in a "quasi-experimental research design" as diagrammed below:

		Year 1		**Year 2**	
Non-Random Assignment	Treatment	Pre-Test	**UPSTART**	Post-Test	Kindergarten
	Control	Pre-Test		Post-Test	

The use of both a pre-test and a comparison group facilitates our ability to examine potential threats to validity, which could jeopardize a clear interpretation of the results (Shadish, Cook, & Campbell, 2002). Because students could not be randomly assigned to treatment or control groups, the groups begin as nonequivalent by definition, and consequently selection bias can be assumed to operate to some degree in some manner. The pre-test allows us to examine the potential for selection bias by determining the nature of the bias as well as its size and direction (i.e., which group is favored over the other by a particular inequality).

C6 Evaluation Samples

The C6 evaluation moved from a using a nonequivalent group approach seen in previous years (Evaluation and Training Institute 2011, 2012a, 2012b, 2013, 2014), to using a statistically matched control group balanced across meaningful variables that contribute to achievement outcomes. Simply put, using a matching process to develop our treatment and control groups is a stronger method for ruling out the influence of preexisting differences between groups on program outcomes.

To help readers make the transition from the previous evaluation design, this report contains information for two samples formed for the study: nonequivalent and matched treatment and control groups.

1. Nonequivalent treatment and control groups were used in previous evaluations, are efficient, and allow for statistical comparisons between groups when controlling for differences in important variables that might also predict students' achievement beyond the UPSTART program (such as pre-test scores, gender, ethnicity, and others). The students in the control group are not matched to treatment students, and have different starting points, such as pre-test scores, so they are called "nonequivalent groups" to designate the lack of matching.

2. A matched treatment-control group is made by statistically matching control students to certain characteristics of treatment students to make two equal or "balanced" groups across a set of important predictor variables. With the appropriate resources, the matching process creates groups that are equivalent before any treatment effects are taken into account. To do this, however, students who are not matched one-to-one must be removed from the final research sample. The process depends on having a sufficiently large enough subject pool to draw from for both treatment and, especially, control students.

ETI's methods for generating each sample are described in more detail below.

Nonequivalent C6 Evaluation Sample

The C6 study recruited a total of 529 preschool children: 200 treatment group children who had enrolled in UPSTART for Year 6 of the program (the 2014-15 school year) and 329 nonparticipating control group children. The children were not randomly assigned to the treatment or control groups.

Treatment children. The 200 treatment group children came from an initial random sample of C6 UPSTART enrollees whose families were contacted about participating in the C6 evaluation[2]. The 200 UPSTART children subsequently participated in pre-testing prior to entering the program over the summer of 2014 and post-tests were conducted the following year upon the conclusion of the program and before children entered kindergarten.

Control children. Data from control children consisted of panel data collected from non-UPSTART participants. The control children were recruited using a variety of strategies, including targeting preschools, daycare centers, childcare organizations, Head Start centers, parent groups, and snowball sampling[3] from families who were UPSTART users.

Table 1 presents key demographic characteristics for the nonequivalent treatment and control sample. As shown in **Table 1**, control families were somewhat more advantaged compared to treatment families from the standpoint of parental education and household income level. For example, 35% of control families indicated that the primary caregiver graduated from a four-year college versus 8% of treatment families.

[2] C6 treatment families were screened based on location, parental education, child language, and known disabilities.

[3] Snowball sampling is when existing participants recruit future participants among their personal network of acquaintances.

Table 1
Nonequivalent Treatment-Control Comparisons on Key Demographics

Demographic Categories		Treatment (N=200)	Control (N=329)
Gender	Female	46%	53%
	Male	53%	47%
Ethnicity	Caucasian	88%	83%
	Hispanic	12%	11%
Child Language	English	96%	94%
Parent Education Level	High School Diploma	15%	11%
	Some College	71%	46%
	Bachelor's degree	8%	35%
	Graduate degree	2%	6%
Parent Marital Status	Married	90%	85%
Household Income	Under $10,000	5%	4%
	$10k-$24,999	8%	12%
	$25k-$49,999	29%	25%
	$50k-$74,999	35%	29%
	$75k-$99,999	22%	18%
	$100k or more	4%	11%

Studies of child development have found that parents with higher levels of education spend more time with their children in ways likely to enhance their development, hold higher expectations for their children, and use varied and complex language and speech patterns (Davis-Kean, 2005; Guryan et al, 2008; Neitzel & Stright, 2004). Thus it is important to ensure that the treatment and control groups are as comparable as possible with regard to parental education when evaluating post-test literacy outcomes.

Appendix A displays pre-existing differences between the nonequivalent treatment and control groups on measured literacy instruments (Brigance and Bader, see **Outcome Measures** section below). Significant differences between the two groups that favored the control group were found on both literacy instruments. While the use of a pre-test and covariates with the nonequivalent sample allows us to examine and statistically control for pre-existing literacy skills and demographic differences between the treatment and control groups, using these control methods can reduce our ability to detect treatment effects and to estimate their size. We determined that using a matched treatment and control group strategy would further reduce the chance that pre-existing differences influenced our ability to statistically test for treatment effects.

Matched Treatment-Control Group Sample

To combat the limitations (cited above) of using the full nonequivalent C6 sample, we used a statistical process called "Coarsened Exact Matching" (CEM) to match control students to treatment students. During the CEM procedure, each treatment child is statistically matched with a control child who is most similar to them and if no matches can be made, children are removed from the sample. Additional tests are preformed to assess the balance between the treatment and control group to ensure that the groups are as similar as possible. The resulting matched treatment-control sample consists of treatment children who have a statistical control "twin". Using CEM, we are able to construct a comparison group of control children that resemble the treatment sample as closely as possible on specific observable characteristics, such as gender, race/ethnicity, language, parental education, and performance on pre-test measures.

The CEM procedure consisted of a three-step process:

1. The C6 nonequivalent evaluation sample contained data from 200 treatment students from C6 and 329 comparison students who did not participate in the UPSTART program.
2. Students from the pool of potential controls were then matched to treatment students using CEM, which found an exact match—or twin—for treatment students from the group of control students in terms of:
    - Sex (Female/Male)
    - Ethnicity (White, Hispanic, African American, or Asian),
    - Language
    - Parent Education
    - Household income
    - Brigance Composite pre-test scores
    - Bader Composite pre-test scores
3. Statistical tests assessed the balance between treatment and control group to ensure groups are as similar as possible.

The matching process resulted in a data file with comparable students in each group so that we could improve our precision in estimating treatment effects. **Table 2** displays the demographic breakdown of the matched treatment and control groups. Note how the two groups in the matched sample are much more similar in terms of parental education than in the nonequivalent sample.

Table 2
Matched Treatment-Control Comparisons on Key Demographics

Demographic Categories		Treatment (N=138)	Control (N=138)
Child Gender	Female	49%	49%
	Male	51%	51%
Child Ethnicity	Caucasian	98%	98%
	Hispanic	1%	1%
Child Language	English	100%	100%
Parent Education Level	High School Diploma	12%	10%
	Some College	75%	75%
	Bachelor's degree	9%	9%
	Graduate degree	3%	5%
Parent Marital Status	Married	95%	89%
Household Income	Under $10,000	2%	2%
	$10k-$24,999	5%	10%
	$25k-$49,999	29%	29%
	$50k-$74,999	35%	34%
	$75k-$99,999	24%	17%
	$100k or more	5%	8%

Comparison of Analysis Samples with C6 Population

Table 3 compares the matched and non-equivalent samples with the C6 population on key demographic characteristics. The matched sample is more homogenous than the C6 or non-equivalent sample, with 94% of children being Caucasian and 100% classified as English speakers.

Table 3
Sample Comparisons on Key Demographics

Demographic Categories		C6 Population (N = 5,091)	Nonequivalent Sample (N=200)	Matched Sample (N=138)
Gender	Female	48%	46%	48%
	Male	52%	54%	52%
Ethnicity	Caucasian	83%	81%	94%
	Hispanic	12%	11%	2%
Child Language	English	92%	96%	100%
Parent Education Level	Some College	36%	78%	83%
	Bachelor's Degree	42%	1%	1%
Parent Marital Status	Married	94%	92%	95%
Poverty Status	Under 185%	45%	52%	49%

The C6 population had parents with higher education levels and slightly lower levels of poverty. Whereas 42% of the parents in the overall C6 population have a college degree, the modal level of parent education in the matched and nonequivalent sample was some college (83% and 78%, respectively). Additionally, 45% of families in the C6 sample were under the 185% federal poverty rate compared to 52% of families in the nonequivalent sample and 49% of families in the matched sample.

The nonequivalent sample is closer to representing the characteristics of the C6 population. However, the matched sample ensures that the treatment group's characteristics best mirror the control group to estimate program impact with the greatest accuracy. UPSTART outcome findings are reported in the main body of the report from the matched treatment-control sample and a comparison of the results from the matched and nonequivalent samples can be found in **Appendix B**.

Outcome Measures

The reading skills taught by the Waterford Early Learning Program at Level 1 of the curriculum[4] include:

- Phonological Awareness: phonemic segmenting and blending
- Phonics: letter name knowledge, letter sound knowledge, and word reading
- Comprehension and Vocabulary: vocabulary knowledge and oral comprehension
- Language Concepts: concepts of written language from letters and pictures to basic grammar

The outcomes of interest for the UPSTART evaluation are measures of early literacy skills that are **aligned to the UPSTART curriculum and considered to be important predictors of later reading ability**, such as phonological awareness, letter knowledge, and vocabulary. In order to measure these outcomes in our treatment and control groups, we used appropriate subscales from two standardized measures of early literacy, the Brigance Inventory of Educational Development and the Bader Reading and Language Inventory.

The Brigance. The Brigance Inventory of Educational Development (Brigance, 2014) was selected as an early literacy measure of phonics and vocabulary knowledge and as a measure of pre-Kindergarten academic and cognitive skills. Ten scales were administered from the language development and academic/cognitive domains of the Brigance. Brigance subscales measured the literacy constructs of *vocabulary and syntax*, *pre-literacy discrimination*, *letter knowledge*, and *decoding* and are described in detail in **Table 4** on the following page. A composite Brigance score to create a comprehensive score of early literacy achievement was created by adding the scores from the ten subtests. Possible scores on the Brigance composite range from a low of 0 points to a high of 240 points.

The Bader. The Bader Reading and Language Inventory (Bader, 2008) was selected as a measure of *phonological awareness*. Phonological awareness involves the child's ability to detect the sound structure of spoken words at three levels: rhyming, syllables, and phonemes. The Bader is comprised of three phonological awareness subtests (rhyme recognition, phonemic blending, phoneme segmentation), along with a composite summary phonological awareness score that was calculated by adding the scores from the three subtests.

[4] Level 1 of the UPSTART curriculum is the beginning point of the curriculum where the preschool child begins as a nonreader and is introduced to skills designed to teach the child to read.

Table 4 summarizes the alignment between the UPSTART curriculum and the literacy constructs measured by the Brigance and Bader, and also contains information about specific skills assessed by the Brigance and Bader subscales, along with possible scale ranges.

Table 4
Alignment of Outcome Measures with UPSTART Curriculum

UPSTART Curriculum	Literacy Construct	Instrument Subscale	Measured Skill	Possible Range
Language Concepts	Pre-literacy Discrimination	Auditory Discrimination	Identifies if two words sound the same	0-10
		Visual Discrimination	Identifies similarities and differences between forms, letters, and words	0-20
Comprehension/ Vocabulary	Vocabulary and Syntax	Expressive Vocabulary	Names pictures	0-27
		Receptive Vocabulary	Points to pictures named by an assessor	0-27
		Expressive Grammar	Talks about an illustration	0-12
Phonics I	Letter Knowledge	Recites Alphabet	Recites alphabet	0-26
		Lowercase Letter Knowledge	Names or recognizes lowercase letters	0-52
		Sounds of Lowercase Letters	Produces sounds of lowercase letters	0-26
Phonological Awareness	Phonological Awareness	Rhyme Recognition	Identifies word pairs that rhyme or do not rhyme	0-10
		Phonemic Blending	Blends separate word sounds into single word	0-8
		Phoneme Segmentation	Segments word into separate word sounds	0-8
Phonics II	Decoding	Survival Sight Words	Reads survival sight words that appear in public places	0-16
		Pre-Primer Vocabulary	Reads basic vocabulary words found in pre-primer reading programs	0-24

Data Collection

Data were collected for 200 treatment group children who had enrolled in UPSTART for Year 6 of the program and 329 control group children who had not enrolled in the UPSTART program. The children's parents were given an intake questionnaire during the pre-test session that collected demographic information from children, parents, and the household. The children were post-tested on the Brigance and Bader a year later before entering kindergarten.

A student data file was developed based on data collected from the intake questionnaire and from the pre-test and post-test administrations of the Brigance and Bader. The final analysis file was based on the subset of children with valid matched pre-test and post-test data, and who had not previously used the UPSTART computerized learning program as documented through the pre-screening interview.

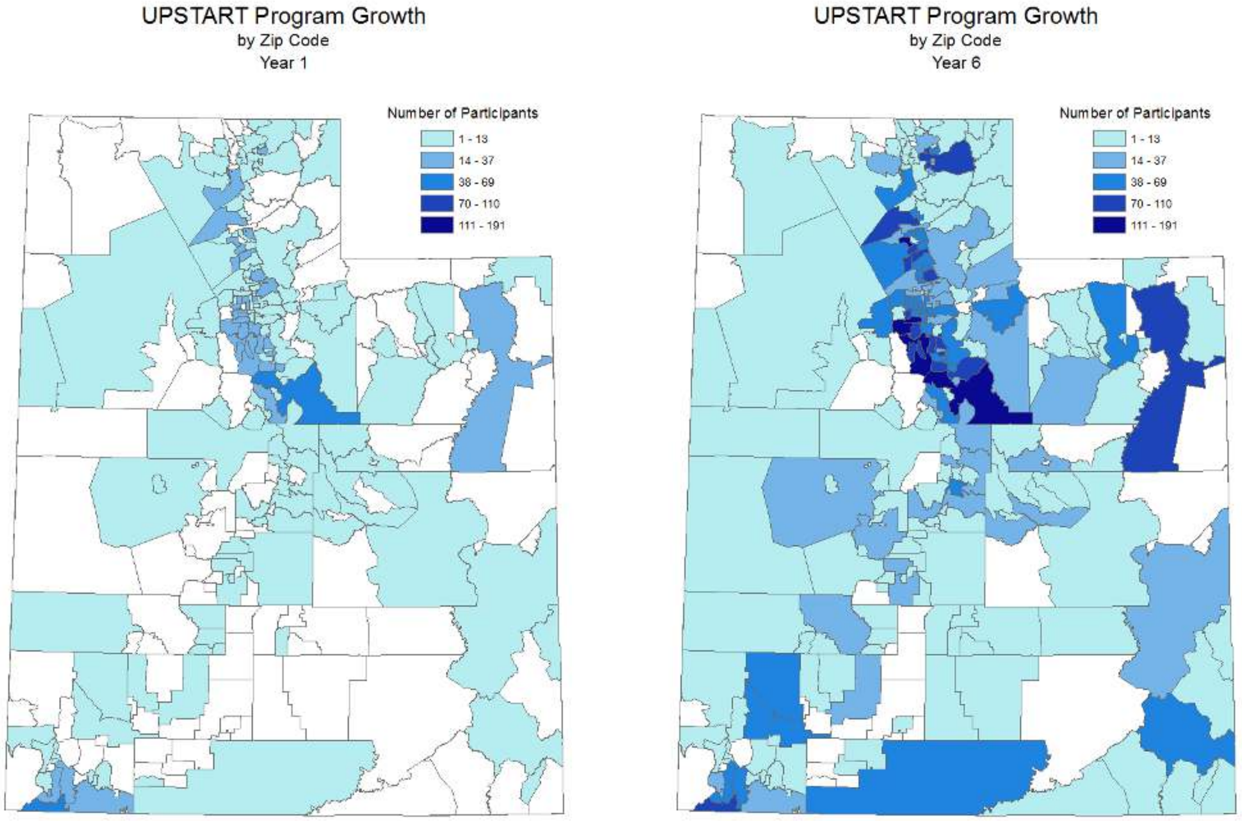
UPSTART Program Implementation

Findings reviewed in the UPSTART implementation section include sixth year enrollment, equipment provided to enrolled families by UPSTART, usage of the UPSTART curriculum in terms of instructional time logged, the proportion of UPSTART students considered to have “graduated” from the program, and the relationship between levels of UPSTART curriculum usage and literacy outcomes.

UPSTART Enrollment

The 2014-15 program year marked a breakout year for UPSTART enrollment. The number of preschool students enrolled in the program rose from 1,577 children in Year 5 to 5,091 students in Year 6, an increase of over 300 percent. The maps depicted in **Figure 1** showcase UPSTART program participation by student zip code from the inception of the program (Year 1, N=1,248) to the most recent program year (Year 6, N=5,091). As seen below in **Figure 1**, the UPSTART program has furthered its reach over the past six years and augmented enrollment in both urban and rural areas of the state.

Figure 1. Map of UPSTART program participation in Year 1 and Year 6



The Waterford Institute provided documentation for the sixth-year UPSTART enrollment of 5,091 children, including demographic information, provisioned educational technology, UPSTART program usage, and whether or not children completed program requirements. Some basic demographic characteristics of the C6 population are presented below in **Table 5**, along with characteristics of UPSTART children comprising the nonequivalent treatment sample and the matched treatment sample.

Table 5
Demographic Characteristics of C6 Population

Demographic Categories		**All C6 UPSTART** (N=5,091)	**Nonequivalent Sample** (N=200)	**Matched Treatment** (N=138)
Child's Gender	Male	48%	46%	48%
	Female	52%	54%	52%
Child's Ethnicity	White	83%	81%	94%
	Hispanic	12%	11%	2%
	Asian/Pacific Islander	3%	5%	3%
	African American	1%	1%	0%
	Native American	<1%	1%	1%
	Other	2%	2%	1%
Child's Language	English	92%	96%	100%
	Spanish	7%	4%	0%
	Other	1%	1%	0%
Parent Educational Attainment	Some High School	3%	4%	1%
	High School Graduate	10%	19%	15%
	Some College	36%	78%	83%
	College Graduate	42%	1%	1%
	Advanced Degree	9%	0%	0%
Parent Marital Status	Married	94%	92%	95%
	Otherwise	6%	8%	5%
Household Poverty Level	Under 100%	16%	19%	12%
	Under 185%	45%	52%	49%
	Under 200%	50%	57%	53%

Note: Percentages may not add to 100% due to rounding.

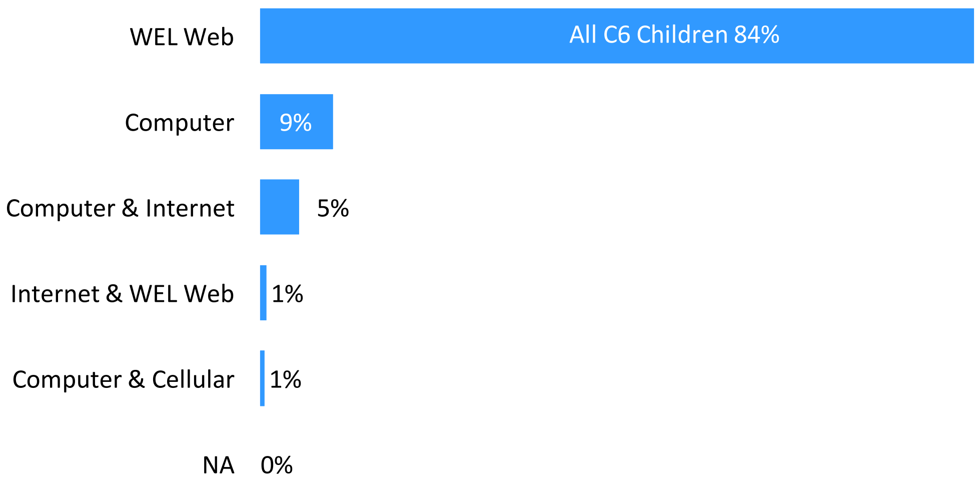
Slightly more C6 girls (52%) were enrolled than boys (28%) and in terms of ethnicity, the vast majority (83%) of the C6 enrollment was White, with 12% of the children being of Hispanic origin. Half of the C6 UPSTART participants lived in families with incomes less than 200% of the federal poverty level.[5]

[5] The federal poverty definition consists of a series of thresholds based on family size. In 2014, a 100% poverty threshold for a family of four was $23,850, while a 200% threshold for a family of four was $47,700.

Provided UPSTART Equipment

The type of education technology provided to UPSTART children in Year 6 of the program is shown in **Figure 2** for all 5,091 children enrolled and for the C6 analysis sample (N=200). The vast majority of UPSTART children (84%) used the Waterford website to retrieve the UPSTART program. This allowed families to access the UPSTART curriculum from their home computers. Similarly, students in the C6 analysis sample most often (80%) also accessed the UPSTART curriculum through the Waterford website.

Figure 2. Equipment provided to C6 Participants by Waterford



*Note: Percentages may not add to 100% due to rounding.

Second most frequently, UPSTART provided free personal computers to 9% of the C6 children while they participated in the program. Another 5% of the C6 program participants were provided with free internet subscriptions and personal computers. The remaining 7% of the C6 enrollment received various combinations of computer technology to enable them to access the UPSTART curriculum (see **Figure 2** for details).

UPSTART Usage

We reviewed program usage (time spent using the software program) for three groups: all UPSTART participants, UPSTART program graduates, and the evaluation analysis sample. The hours of instruction observed for all children documented as enrolled in the sixth year of UPSTART are summarized in **Table 6**, and are compared to program “graduates”. The average level of usage for all students enrolled in the sixth year of UPSTART (N=5,091) was approximately 67 hours of instruction; this is slightly less than the average level of usage as documented in the fifth year of the program (71 hours for C5; see Evaluation and Training Institute, 2015). The C6 academic year covered 44 weeks of instruction, beginning the week of September 1, 2014 and ending June 29, 2015.

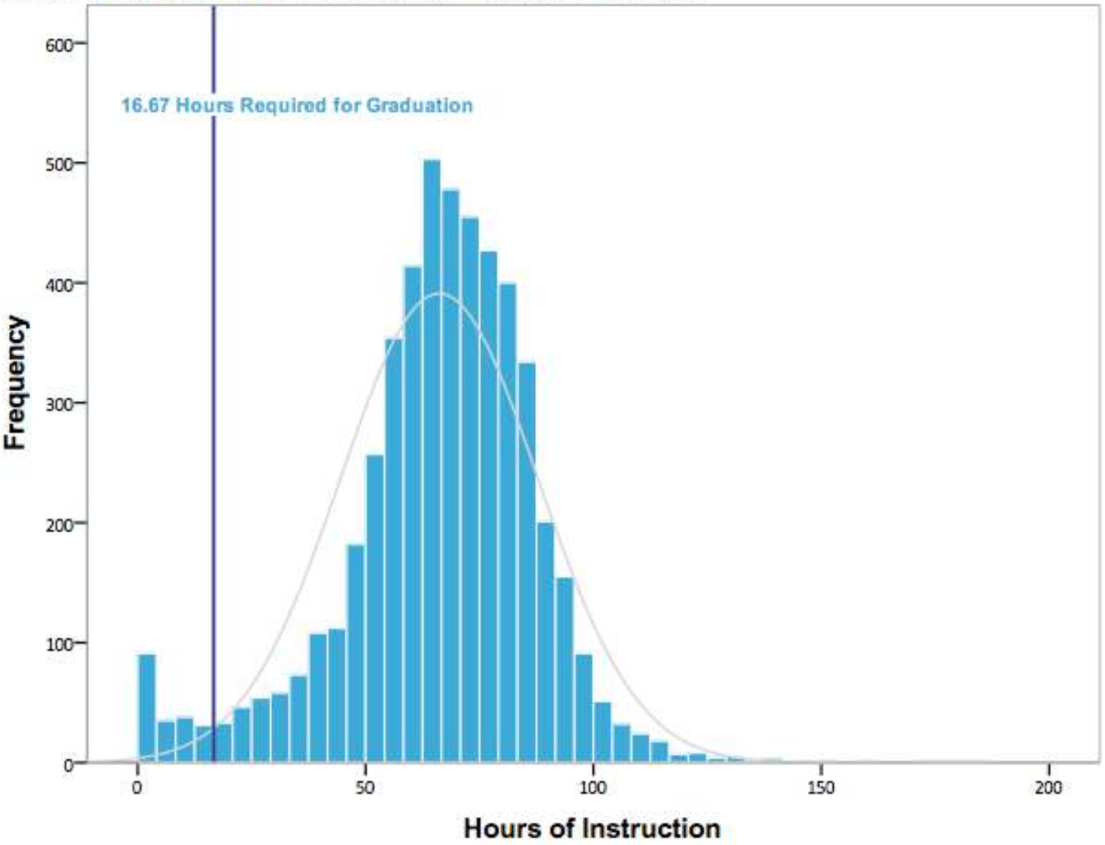
Table 6
C6 Hours of UPSTART Instruction

Group	N	Mean	SD	Range
All UPSTART	5,091	66.75	21.64	00.00 - 183.56
UPSTART Graduates	4,674	70.33	16.71	16.74 - 183.56
UPSTART Analysis Sample	200	69.42	18.04	5.63 – 114.94

Forty-five of the enrolled families who were provided instructional equipment (e.g., computers, an Internet subscription, and a computer drive) did not log any instructional time in the UPSTART curriculum during Year 6 of the program. These families dropped out of the program within eight weeks of enrollment. For enrolled families whose children did use the curriculum, the average duration in the program was approximately 41 weeks. This usage pattern is similar to that observed in the fifth year of the program.

The children in the C6 evaluation analysis sample used the UPSTART curriculum for approximately 69 hours of instruction on the average (see **Table 6**).
The histogram in **Figure 3** shows the distribution of hours of instruction for the total C6 population (N=5,091). As noted previously, forty-five of the enrolled children logged zero hours of instruction during their time in UPSTART. At the other end of the spectrum, six children logged over 150 hours of instruction.

Figure 3. Hours of Instruction for C6 Families

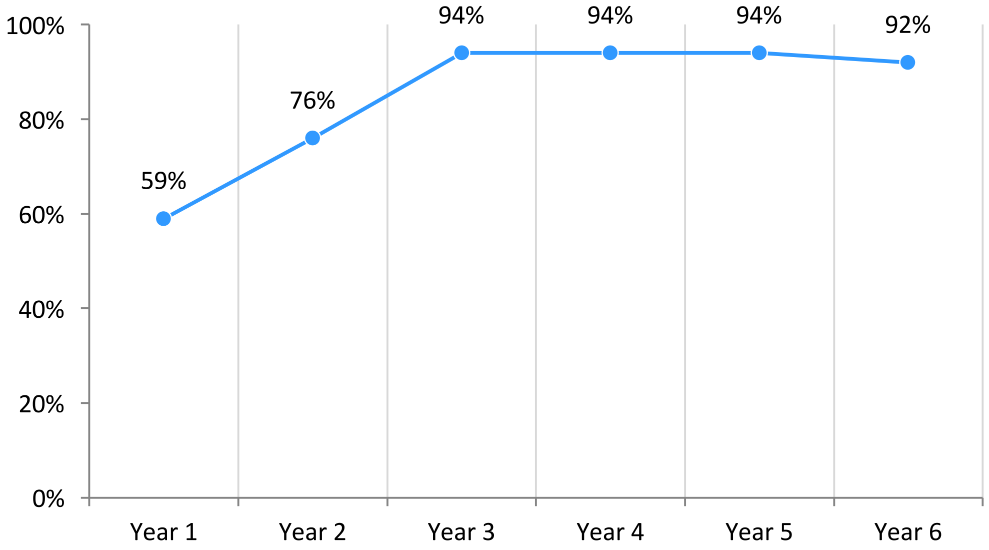


The bottom quartile of the C6 population completed 56.19 hours of instruction or less, the midpoint of the C6 distribution was 68.08 hours, and the top quartile completed in excess of 79.79 hours of instruction.

UPSTART Graduates

Of the 5,091 children documented as enrolled in UPSTART in the sixth year of the program, the Waterford Institute classified 4,674 as children who had met the program's usage criteria and were thus considered to be graduates of the program. The usage criteria involved (a) logging more than 1,000 minutes (16.67 hours of instruction) with the UPSTART curriculum and (b) averaging at least one hour of instruction per week while participating in the program. By this definition, Cohort 6 achieved a graduation rate of 92% (i.e., 4,674/5,091 = 0.92). As seen in **Figure 4**, this is a slightly lower rate than the previous three years, which may reflect the dramatic growth in participants in Year 6 as the UPSTART population increased from 1,577 children in Year 5 to over 5,000 children in Year 6.

Figure 4. UPSTART Graduation Rates over Time



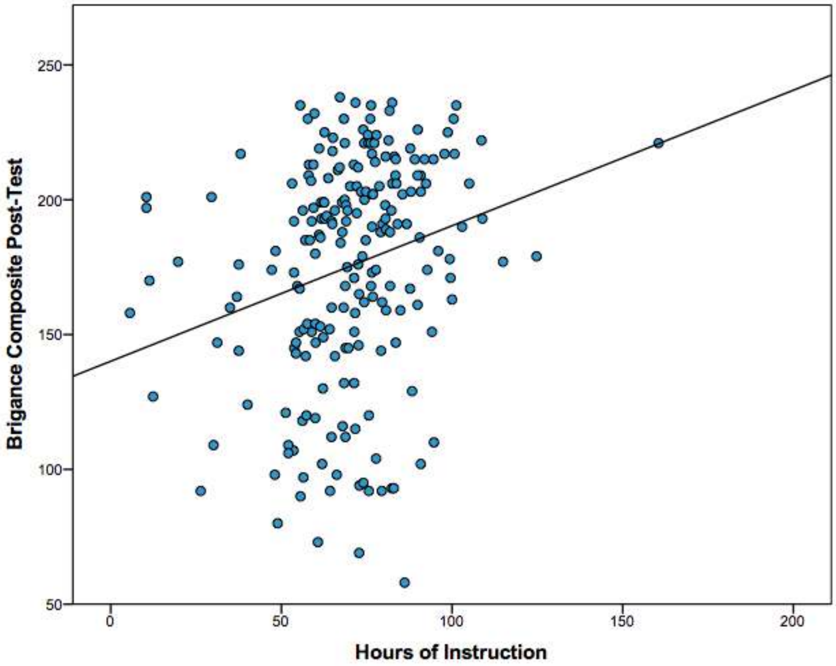
UPSTART graduate status was significantly correlated with hours of instruction (*r* = .65) and with the number of weeks in the program (*r* = .64).

UPSTART Usage and Literacy Outcomes

Similar to previous years, the sixth year evaluation of UPSTART found curriculum usage to be significantly and positively related to literacy outcomes as measured by composite scores on the Brigance and Bader instruments.

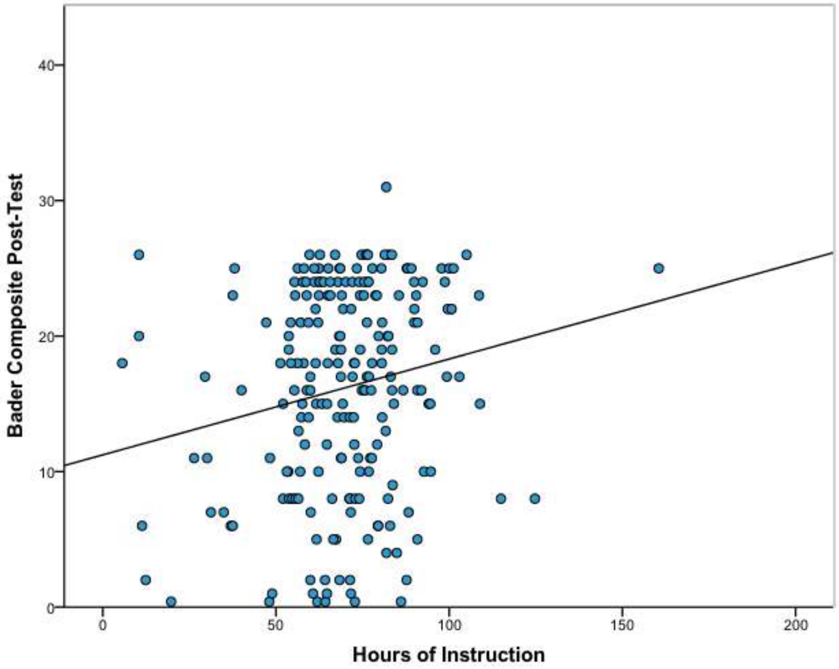
The plot in **Figure 5** on the following page shows a linear relationship between UPSTART usage (measured in hours of instruction) and Brigance post-test scores. That is, Brigance post-test scores tend to increase with increasing hours of UPSTART usage.

Figure 5. Plot of Hours of Instruction and Brigance post-test scores



Similarly, the plot presented in **Figure 6** displays the relationship between hours of UPSTART instruction and the Bader composite post-test score indicates a weak positive linear association between instruction time and scores on the Bader post-test. This suggests that the acquisition of early phonological skills as measured by the Bader tend to improve with increasing levels of exposure to UPSTART curriculum.

Figure 6. Plot of Hours of Instruction and Bader post-test scores



UPSTART Program Impacts on Literacy

This section includes results based on statistical comparisons of literacy achievement (test scores) for matched treatment and control groups. The impact of the UPSTART program is shown through two lenses: effect sizes and growth scores. Both methods provide salient feedback about the impact of UPSTART. The first method helps stakeholders understand how large an impact UPSTART had on participants, while the second method shows how UPSTART students grew (compared to control students) based on two points of time. To explore the implications of using matched vs. nonequivalent group designs, we also provide findings for the two sampling approaches in **Appendix B**.

Findings in this section were analyzed to answer the following two research questions:

> **Research Question 1:** *Do UPSTART students have better literacy skills at Kindergarten than control students?*
>
> **Research Question 2:** *Do UPSTART students show stronger literacy growth rates from preschool to Kindergarten than control students?*

The results of the matched sample are presented for each research question above, and the statistically significant (*p* < .05) findings are depicted visually[6].

Do UPSTART students have better literacy skills at entry to Kindergarten than control students?

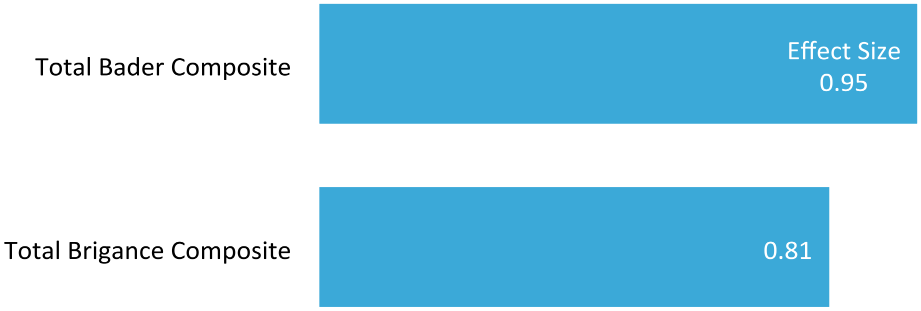
Effect sizes[7] were calculated to show the magnitude of UPSTART's impact at post-test as measured by each of the 13 literacy subtests (10 Brigance subtests and 3 Bader subtests), and the Total Brigance and Bader Composites (composites include aggregated results of the subtests). An effect size (ES) is a measure that describes the magnitude of the difference between two groups, essentially standardizing a scale so the results are easy to interpret and have meaning. Cohen (1998) categorizes effect sizes as small (0.2), medium (0.5), and large (0.8). Combined post-test results showed that UPSTART participation had a large impact on students' early literacy skill development. In the matched post-test sample[8] (N=271), UPSTART produced large effects (.95 and .81) as measured by the total Bader and Brigance composite scores (see **Figure 7**).

[6] To create a concise report that highlights the most important findings for stakeholders, we did not present findings that were non-significant in figures. Comprehensive results can be found in **Appendix B**.

[7] Effect size (Cohen's *d*) was calculated for each test as the treatment group mean minus the control group mean divided by the pooled standard deviation.

[8] Treatment Group (N = 138); Control Group (N = 133)

Figure 7. Brigance and Bader Posttest Analysis of Composite Scores



UPSTART children scored significantly higher on eleven of the thirteen Brigance and Bader subtests on the post-test, showing strong empirical evidence that UPSTART was successful helping children develop key early literacy skills. The ES estimates for individual subtests ranged from .44 (Rhyme Recognition) to 1.1 (Pre-primer Vocabulary) and would be considered medium to large effects. Expressive and Receptive Vocabulary subtests were the only subtests in which the treatment and control groups were non-significant at post-test.

The effect size estimates for each statistically significant literacy subtest (11 out of 13), as measured by the Brigance and Bader instruments, are presented below in **Figure 8**. The results are organized according to the subtests' respective literacy constructs (see **Table 4** on page 15 for a list of all 13 subtests and corresponding constructs).

Figure 8. Effect Size Estimates by Literacy Construct

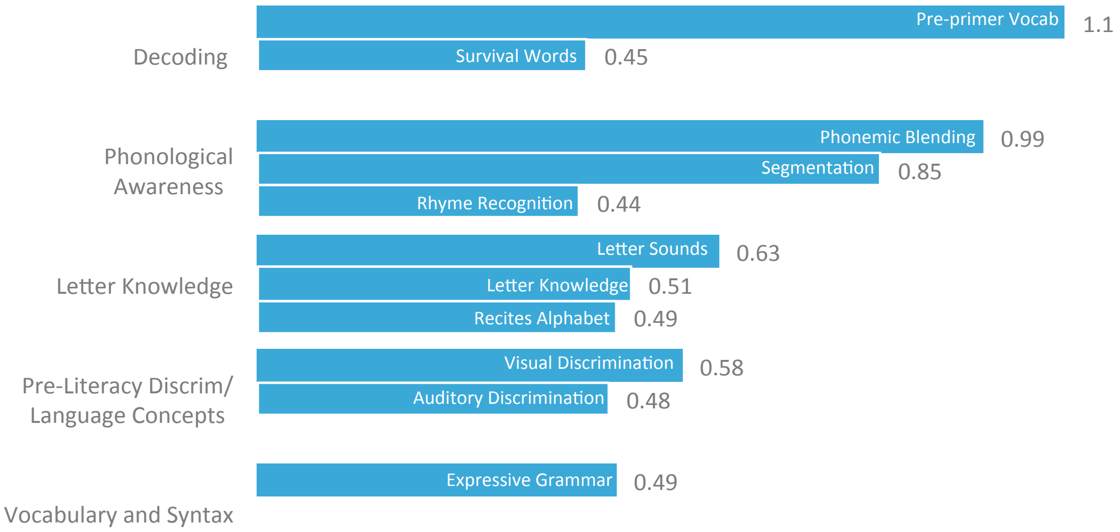
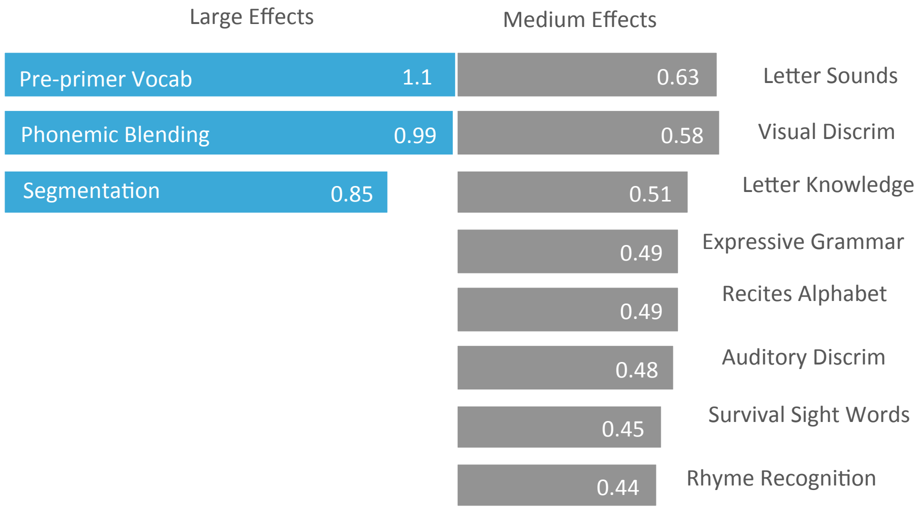


Figure 9 presents the ES of each literacy subtest based on the size of their effects (small, medium or large). UPSTART had the largest impact on pre-primer vocabulary (1.1), phonemic blending (.99), and phonemic segmentation (.85).

Figure 9. Effect size estimates by magnitude of effect



Regression Results. In addition to computing effect sizes, we ran regression analyses to determine if pre-existing differences between the treatment and control groups on demographics and pre-test measures affected the results. The regression analyses did not essentially change the initial estimate of the mean overall impact on the Bader at post-test, however the linear regression analyses improved the estimate of UPSTART's overall impact on the Brigance post-test from 33.73 to 36.24 points (see **Table 7**).

Table 7
Comparison of Deltas by Post-test Composite Measure and Analysis Method

	T-Test	Regression
Bader	6.61	6.85
Brigance	33.73	36.24

None of the significant demographic differences between the treatment and control group were significantly correlated with Brigance post-test scores in the matched sample. These variables included prior usage of a computer at home (favoring the treatment group), possession of an IPAD/tablet computer in the home and its usage in daycare (favoring the control group), and prior participation in daycare (favoring the control group).

Do UPSTART students show stronger literacy growth rates from preschool to Kindergarten than control students?

We studied literacy growth rates while in the program as an additional way to evaluate program impacts beyond outcome score comparisons. Paired samples t-tests were performed to examine growth rates as measured by the Brigance and the Bader total test batteries and subtests for the treatment and control group children. Growth rates for the treatment and control children were compared based on the observed difference scores between the post-test and the pre-test.

- The treatment group showed significantly ($p < .05$) stronger mean literacy growth rates compared to the control group on the Total Bader and Brigance Composites, with the treatment group scoring an average of 7 points higher on the Bader and 37 points higher on the Brigance.

- The treatment group showed statistically stronger ($p < .05$) literacy growth rates compared to the control group on five out of ten Brigance subtests (Letter Knowledge, Letter Sounds, Auditory Discrimination, Survival Sight Words, and Basic Vocabulary) and all three Bader subtests (Rhyme Recognition, Phonemic Blending, and Segmentation).

- There was no difference in growth rates between the treatment and control group on the following four subtests: Expressive and Receptive Vocabulary (measures vocabulary and syntax), Expressive Grammar (measures vocabulary and syntax), Visual Discrimination (measures pre-literacy discrimination), and Recites Alphabet (measures letter knowledge).

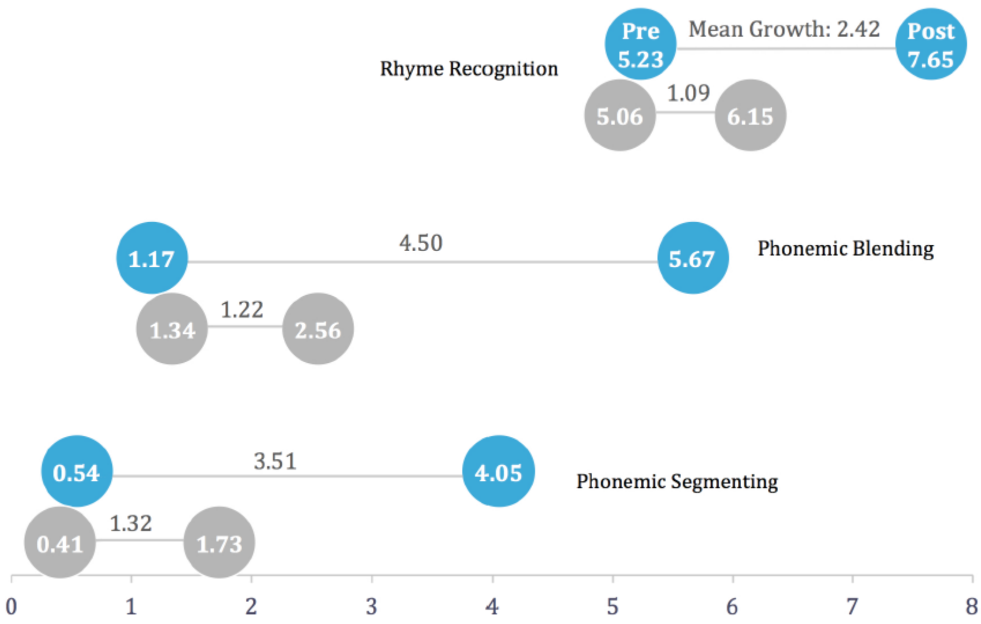
- Of the five literacy constructs in which the Brigance and Bader subtests measure, Vocabulary and Syntax was the only construct in which growth rates between the treatment and control students were not statistically significant (p<.05).

Growth rates from pre-test to post-test are shown in the figures below. Each figure categorizes the Brigance and Bader subtests that were statistically significant (p<.05) based on their respective literacy constructs, which include: **phonological awareness**, **decoding**, **pre-literacy discrimination**, and **letter knowledge**[9]. UPSTART participants' scores are depicted in blue, while their control group counterparts are in grey.

UPSTART children experienced significant, higher mean growth from pre-test to post-test compared to control children on all three subtests (rhyme recognition, phonemic blending and segmenting) that measure **Phonological Awareness**.

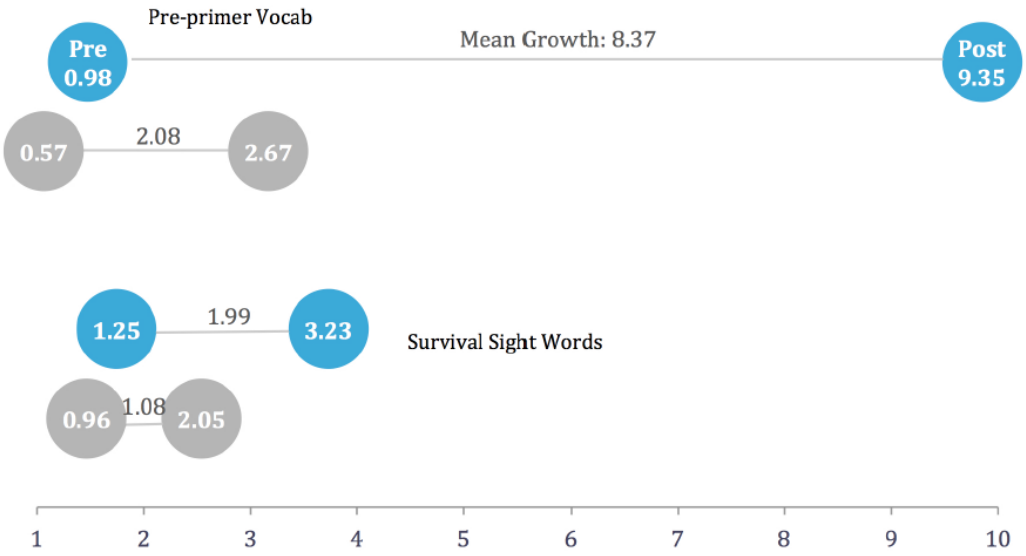
[9] This section presents outcomes that were statistically significant, and therefore a figure for vocabulary and syntax is not depicted here.

Figure 10. Phonological Awareness: Treatment and Control Group pre-and-posttest mean scores



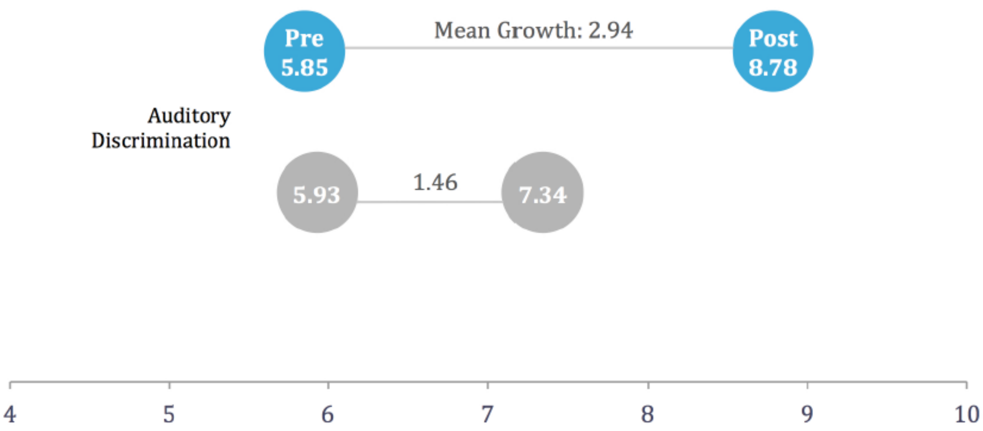
UPSTART students experienced significant, higher mean growth compared to the control group on both subtests used to measure children's **Decoding** ability, including pre-primer vocabulary and survival sight words.

Figure 11. Decoding: Treatment and Control Group pre-and-posttest mean scores



When compared to the control group, UPSTART children experienced significantly higher growth in auditory discrimination, which measures children's ability to identify if two words sound the same. Auditory discrimination is one of two subtests used to determine children's skill in **Pre-Literacy Discrimination.**

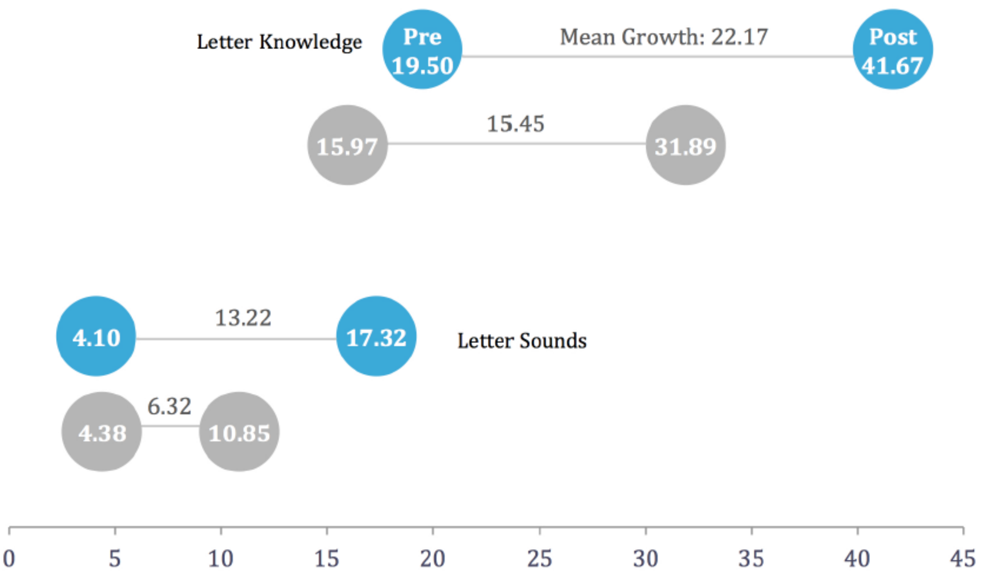
Figure 12. Pre-literacy Discrimination: Treatment and Control Group pre-and-posttest mean scores



*Note: Growth rates in *visual discrimination* were not significant between the treatment and control groups.

UPSTART children experienced significantly higher growth, compared to non-UPSTART children, in two out of three literacy subtests measuring **Letter Knowledge**. UPSTART children showed stronger growth in naming or recognizing lowercase letters (letter knowledge) and producing sounds of lower case letters (letter sounds). A significant difference in the growth rates of treatment and control students was not observed for the visual discrimination subtest, in which children identified the similarities and differences between forms, letters and words.

Figure 13. Letter Knowledge: Treatment and Control Group pre-and-posttest mean scores



*Note: Growth rates in *recites alphabet* were not significant between the treatment and control groups.

Discussion and Recommendations

The final section of the Cohort 6 (C6) evaluation report includes findings and trends for UPSTART implementation and impacts on early literacy skills. Based on the results and additional discussion about the evaluation design, we include summary recommendations for the program and future research efforts to help the state monitor its impacts.

Program Implementation

Based on the data provided by UPSTART program officers, the program was implemented with great success. UPSTART enrollment increased from 1,577 children in Year 5 to 5,091 children in Year 6, an increase of over 300 percent. Enrollment has increased across the state of Utah and UPSTART has reached families in both rural and urban areas. Half of the children enrolled in Year 6 lived in families with incomes less than 200% of the federal poverty level and the majority of the children were White (83%) and English speaking (92%).

Most of the C6 children accessed the UPSTART curriculum through the Waterford website (84%). Approximately 9% of the sixth year participants received a computer loan and 5% were provided with a computer and internet. While graduation rates were approximately 2% lower than in previous years, program usage was consistent with successful program implementation. A slight drop in graduation rates could be due to the increased enrollment across the state.

Program Impacts on Literacy Development

While program implementation findings are important for monitoring how resources were used to enroll and graduate students, findings about literacy testing outcomes is the most important indicator of program success. UPSTART participation had a strong impact on children's emerging literacy skills based on the results from effect size and growth score analyses. The program produced large statistical effects (Brigance ES = .81; Bader ES = .95) on learning compared to non-program children. The effects were seen across different measures of literacy: decoding skills, letter knowledge, pre-literacy discrimination, and phonological awareness.

We used two types of statistical comparisons to give the state multifaceted findings related to literacy achievement during the pre-kindergarten year: effect sizes and growth scores. The effect size estimates measured the differences between the treatment and control students at post-test, while the growth score analyses measured the change from pre-test to post-test for both the treatment and control groups.

We reported findings for focused literacy tests, and a majority of the results from the Brigance and Bader scales were shown to have medium to large effects (effect sizes ranged from .44 to 1.1). Overall, the results of both analyses illustrate that UPSTART program participation had a strong impact on facilitating UPSTART students' literacy skill development in a variety of key areas. The largest impacts were found for pre-primer vocab (measures decoding skills), phonemic blending and segmentation (measures phonological awareness).

UPSTART students also experienced greater growth from pre-test to post-test compared to control students in four out of five literacy constructs (phonological awareness, decoding, pre-literacy discrimination, and letter knowledge), with the exception of the Vocabulary and Syntax construct, which is comprised of the Expressive and Receptive Vocabulary and Expressive Grammar subtests. Group differences in the Expressive and Receptive Vocabulary subtests were not statistically significant in the post-test analyses, indicating that this is one of the few literacy skill areas in which UPSTART did not have a positive impact.

In general, both the post-test effect size analyses and growth score analyses were consistent, showing that UPSTART students performed better than the control group. However, the post-test analyses using effect sizes depicted three significant subtests in favor of the treatment group in which the growth score analyses did not: Expressive Grammar, Visual Discrimination, and Recites Alphabet. One explanation for the difference in the analyses results could be due to pre-test differences. For instance, the Expressive Grammar and Visual Discrimination subtests showed statistically significant differences between the two groups at pre-test (see Appendix A for pre-test analyses results). Even though we matched treatment and control students across pre-program achievement (composite scores), we could not match them on every literacy subtest.

Limitations

This evaluation report marks the sixth year of the UPSTART evaluation. Each year we like to discuss the implication of the evaluation results on future research efforts. We used a nonequivalent group (pre-/post-test) design in years past, but for Cohort 6 we scaled-up our data collection to gather information from more students and used a matched group design. There are several benefits to balancing students using a one-to-one matching technique, but the method requires large groups of treatment and control students to find the matches, and many treatment students are removed from the analyses because they do not have an equivalent control student. Removing treatment students from our matched sample could reduce our statistical power to detect smaller treatment effects. Treatment students are randomly matched to control students with similar matching variables (see our method section for more information), but there is no way to determine if the students who were not matched would have influenced the results since they were not included in the analyses.

Even given the limitations of a smaller matched sample size than a nonequivalent group design, Coarsened Exact Matching (CEM) allowed us to make the treatment and control groups as similar as possible prior to running statistical models to determine differences in literacy between them. By reducing pre-existing differences across a set of predictor variables, using CEM provides a more accurate estimate of the impact of UPSTART compared to analyses done with nonequivalent treatment and control groups. Future evaluations should continue using matching methods to minimize pre-existing differences between nonequivalent treatment and control groups.

The largest barrier to matching treatment to control students is recruiting similar control students to participate in the evaluation. As UPSTART was initially intended to support low-income children who may be at risk for insufficient preparation for kindergarten, we similarly attempted to target low-income families in our control group. These families are difficult to locate with conventional recruiting strategies and it can be challenging to secure participation with both pre-testing and post-testing.

UPSTART and non-UPSTART (control) families are naturally occurring groups, devoid of random assignment, so it is important that they resemble each other as closely as possible to ensure that a balanced control group is present. Recruiting control families for the UPSTART evaluation has been a persistent challenge. As the UPSTART program expands its reach to include more families, the population of potential control families shrinks. In addition, some of our previous control family recruitment sites are no longer viable: due to unknown reasons, certain Head Start programs have chosen not to allow us to pass along information to parents (even when the program guarantees parents financial incentives for participation). We would like to emphasize that certain pre-K program providers, such as Centro de la Familia de Utah, and staff at the USOE have been great assets in helping the evaluators reach non-UPSTART control families. We hope to find other partners who serve similar populations, such as Women Infants and Children (WIC) and public preschool programs- all of which should support research and evaluation to improve the lives of their constituencies.

Recommendations

The UPSTART program shows continued success at helping preschool age children develop literacy skills and prepare for school. These outcomes would have specific benefits to at-risk children, whose families struggle with poverty and other issues, and often lack the resources to help their children develop the literacy skills needed to succeed in school. Given the success at improving literacy test scores, we recommend that the state continue providing the UPSTART program to children. The strong program effects support wide-scale implementation across at-risk preschool populations.

Program enrollment increased in the 2014-2015 program year, which resulted in more families benefitting from the computer-based instruction. However, slightly less C6 students were classified as graduates when compared to previous cohorts (92% vs 94% for C5, for example). In addition to a slight drop in graduation rates, average program usage dropped approximately 4 hours when compared to the previous year (C5 average use was 71 hours vs. 67 hours of average use for C6). While slight, these reductions need to be monitored to be sure it is not a trend due to the demands of increased enrollment. We recommend that the program vendor work with the evaluator and USOE staff to monitor program implementation carefully and to be sure that increased enrollment does not erode graduation or usage rates, two key areas for ensuring strong student literacy achievement and future program success.

We recommend that the matched treatment and control group design be used for future evaluations. This research design depends on collecting sufficient data from control students to allow high matching rates to treatment students. To accomplish these high match rates, we also recommend that the state work with the evaluators to strengthen relationships with other preschool providers, specifically Head Start organizations, WIC and public preschool programs to widen our ability to collect data from non-program control families. This strategy is a win-win for all involved: low-income families can help move the bar on research into early literacy (and receive financial incentives while doing it) and the state can review results across more students and have more data for evidence-based decision making about their pre-Kindergarten school readiness programs.

References

Bader, L. A., & Pearce, D. L. (2008). Bader Reading and Language Inventory (6th ed.). New York, NY: Pearson.

Brigance, A. H. (2004). Brigance Inventory of Early Development II (IED-II) (2nd ed.). N. Billerica, MA: Curriculum Associates.

Cohen, J. (1988) Statistical Power Analysis for the Behavioral Sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Davis-Kean, P. E. (2005). The influence of parent education and family income on child achievement: The indirect role of parental expectations and the home environment. *Journal of Family Psychology*, *19*(2), 294–304.

Evaluation and Training Institute. (2011, August). *Utah UPSTART education program evaluation kindergarten outcomes: Program impacts on reading proficiency* (Cohort 1/Year 1 Results Technical Report). Culver City, CA: Author.

Evaluation and Training Institute. (2012a, November). *Utah UPSTART program evaluation: Year 1/Cohort 1 First Grade* (Cohort 1 First Grade Technical Report). Culver City, CA: Author.

Evaluation and Training Institute. (2012b, March). *Utah UPSTART program evaluation kindergarten outcomes: Program impacts on reading proficiency* (Cohort 2 Results Technical Report). Culver City, CA: Author.

Evaluation and Training Institute. (2013, March). *Utah UPSTART program evaluation program impacts on early literacy: Third Year Results* (Cohort 3 Technical Report). Culver City, CA: Author.

Evaluation and Training Institute. (2014, February). *Utah UPSTART program evaluation program impacts on early literacy: Fourth Year Results* (Cohort 4 Technical Report). Culver City, CA: Author.

Evaluation and Training Institute. (2015, March). *Utah UPSTART program evaluation program impacts on early literacy: Year 5 Results* (Cohort 5 Technical Report). Culver City, CA: Author.

Guryan, J., Hurst, E., & Kearney, M. (2008). Parental education and parent time with children. *Journal of Economic Perspectives*, *22*(3), 23-46.

Neitzel, C., & Stright, A. D. (2004). Parenting behaviors during child problem solving: The role of child temperament, mother education and personality, and the problem-solving context. *International Journal of Behavioral Development*, *28* (2), 166 - 179.

Shadish, Cook, and Campbell (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton Mifflin Company.

Appendix A: Pre-test Analyses Results

We examined the treatment and control group differences in pre-test scores for the Brigance and Bader Total Composites and each individual subtest. The pre-test analyses were conducted to identify any pre-existing differences in scores between the treatment and control groups at pre-test that might affect subsequent analyses as well as to compare the matched and nonequivalent samples.

The results of the pre-test analyses are presented in this Appendix by each sample: Matched group and Nonequivalent group.

Pre-test Analyses Summary

Matched sample*.* The treatment and control groups were equivalent on all of the Bader measures at pre-test and on all but four of the Brigance subtests at pre-test. The magnitude of these differences ranged from small to medium.

Nonequivalent sample*.* The treatment and control groups were significantly different at pre-test on all of the Bader measures and on a majority of the Brigance measures, including the Total Brigance Composite.

Pre-test Conclusion. The matched sample did a better job of equating the treatment and control groups on the pre-tests.

Matched Groups

Bader Pre-test Analysis

There were no differences between the treatment and control group on the Bader pre-test.

Brigance Pre-test Analysis

In spite of using the matched sample (N=276), there were four statistically significant ($p<.05$) differences between the treatment and control group on the Brigance pre-test. These pre-test differences included the Expressive Vocabulary, Expressive Grammar, Recites Alphabet and the Visual Discrimination subtests and favored the treatment group with the exception of Expressive Vocabulary, which favored the control group. However, the treatment-control difference on the Brigance Total Pre-test Composite was not statistically significant, although the difference was in a direction favoring the control group. Comparison of the group pre-test mean differences between the treatment and the control group ("delta") along with the size and magnitude of the effects for these three Brigance pre-test measures are shown in **Table A.1**.

Table A.1
Brigance Pre-test Differences using the Matched Sample

Brigance Pre-test	Delta	ES	Magnitude of Effect
Expressive Vocabulary Subtest	-0.38	0.28	Significant small difference
Expressive Grammar Subtest	1.00	0.53	Significant medium difference
Visual Discrimination Subtest	1.57	0.35	Significant small difference
Recites Alphabet Subtest	2.59	0.30	Significant small difference
Total Brigance Pre-test Composite	-1.58	0.05	Nonsignificant difference

Nonequivalent Groups

Brigance Pre-test Analysis

In the nonequivalent sample, the treatment and control group were significantly different (p<.05) at pre-test on six of the ten Brigance subtests and on the total Brigance pre-test composite. Significant pre-test differences involved the following subtests: Expressive and Receptive Vocabulary, Expressive Grammar, Recites Alphabet, Lowercase Letter Knowledge and Lowercase Letter Sounds. Five of the six significant pre-test differences favored the control group over the treatment group. Initial group differences were in the small to moderate range. Comparison of the group pre-test mean differences (Delta) on the Brigance along with the size, significance[10] and magnitude of the pre-test differences are shown in **Table A.2**.

Table A.2
Brigance Pre-test Differences using the Nonequivalent Sample

Brigance Pre-test	Delta	ES	Significance and Magnitude
Expressive Vocabulary Subtest	-1.06	0.64	Significant moderate difference
Receptive Vocabulary Subtest	-0.46	0.45	Significant moderate difference
Expressive Grammar Subtest	0.44	0.26	Significant small difference
Visual Discrimination Subtest	0.02	0.00	Nonsignificant difference
Recites Alphabet Subtest	-2.61	0.20	Significant small difference
Letter Knowledge Subtest	-5.86	0.29	Significant small difference
Letter Sounds Subtest	-3.50	0.36	Significant small difference
Auditory Discrimination Subtest	-0.64	0.19	Nonsignificant difference
Survival Sight Words Subtest	-0.44	0.19	Nonsignificant difference
Pre-primer Vocabulary Subtest	-0.80	0.17	Nonsignificant difference
Total Brigance Composite	-24.84	0.58	Significant moderate difference

Bader Pre-test Analysis

In the nonequivalent sample (N=529), the treatment and control group means were significantly different (*p* ≤ .05) at pre-test on all of the Bader literacy measures (i.e., both subtests and the total test composite). In all cases, the pre-test differences were small and favored the control group over the treatment group. Comparison of the group pre-test mean differences (Delta) on the Bader along with the effect size (ES), and interpretations of the magnitude of the pre-test differences are shown in **Table A.3**.

Table A.3
Bader Pre-test Differences using the Nonequivalent Sample

Bader Pre-test	Delta	ES	Magnitude of Difference
Rhyme Recognition Subtest	-0.73	0.23	Significant small difference
Phonemic Blending Subtest	-0.87	0.30	Significant small difference
Segmentation Subtest	-0.51	0.24	Significant small difference
Total Bader Post-test Composite	-2.12	0.33	Significant small difference

[10] The interpretation of the significance of a between-group difference is influenced by the variability and degree of error associated with a given measure as well as the size of the difference.

Appendix B: Comparison of the Two Samples: UPSTART Outcomes

Comparison of Outcomes Summary

Matched sample. The matched sample size is considerably smaller than the nonequivalent sample (N= 276, N= 516, respectively), which could diminish the statistical power to detect differences between groups if the treatment effects are small. Growth rates from pre-test to post-test were significantly different between the treatment and control groups among all three Bader subtests and five of the ten Brigance subtests. The matched sample also produced medium to large effects in favor of the treatment group for all measures of the Bader and 8 out of 10 Brigance subtests.

Nonequivalent sample. Matching between treatment and control group students was not done, and the groups were included as they existed (i.e. unequal sizes, and unequal distributions of significant predictors). Growth rates between the treatment and control groups showed significant differences among all three Bader measures and on eight of the Brigance subtests, with the differences in favor of the treatment group. Effect sizes for the nonequivalent sample were typically small, with no large effects observed for any subtest.

Conclusion. Analyses using the matched sample produced larger effect sizes for all measures of the Bader and Brigance subtests, and, because these groups were matched across groups to balance significant predictors of literacy achievement, the ES calculations are more valid. However, the nonequivalent sample did show a greater number of statistically significant subtests in the growth rate analyses (8 of 10 subtests vs. 5 of 10 subtests), most likely due to the greater statistical power with the larger sample size (N=516).

Treatment Effect Size Estimates

Brigance

- Overall, the matched sample produced a greater number of significant subtests compared to the nonequivalent sample (8 out of 10 vs. 7 out of 10), and generated stronger effects across all significant subtests.
- In both the matched and nonequivalent group sample expressive vocabulary and receptive vocabulary were not affected by UPSTART participation.
- The recites alphabet subtest was shown to have no affect in the nonequivalent sample.
- The pre-primer vocabulary subtest was shown to have the greatest effect in both samples, with a large effect (1.10) among the matched sample and a medium effect (0.45) in the nonequivalent group sample.

Table B.1
Brigance Post-test Results: Matched vs. Nonequivalent Samples

Brigance Post-test	Matched Sample			Nonequivalent Sample		
	Delta	ES	Magnitude of Effect	Delta	ES	Magnitude of Effect
Expressive Vocabulary	0.11	0.05	Nonsignificant difference	0.20	0.06	Nonsignificant difference
Receptive Vocabulary	0.23	0.10	Nonsignificant difference	0.40	0.12	Nonsignificant difference
Expressive Grammar	0.93	0.49	Medium effect	0.45	0.23	Small effect
Visual Discrimination	2.23	0.58	Medium effect	1.28	0.34	Small effect
Recites Alphabet	4.60	0.49	Medium effect	0.02	0.00	Nonsignificant difference
Letter Knowledge	9.79	0.51	Medium effect	4.38	0.23	Small effect
Letter Sounds	6.47	0.63	Medium effect	3.02	0.29	Small effect
Auditory Discrimination	1.44	0.48	Medium effect	1.12	0.35	Small effect
Survival Sight Words	1.19	0.45	Medium effect	0.28	0.08	Small effect
Pre-primer Vocabulary	6.68	1.10	Large effect	3.58	0.45	Medium effect
Total Brigance Composite	34.66	0.81	Large effect	15.25	0.32	Small effect

Bader

- In the matched sample, UPSTART participation resulted in a medium impact on the rhyme recognition subtest, while UPSTART participation did not have any affect on rhyme recognition in the nonequivalent sample.
- Phonemic blending and phonemic segmentation produced significant effects in both samples. These subtests produced large effects among the matched sample and medium effects in the nonequivalent sample.
- Both samples show UPSTART had an impact on students' total composite scores, with a much stronger effect (0.95) among students in the matched sample compared to the nonequivalent sample (0.46).

Table B.2
Bader Post-test Results: Matched vs. Nonequivalent Samples

	Matched Sample			Nonequivalent Sample		
Bader Post-test	Delta	ES	Magnitude of Effect	Delta	ES	Magnitude of Effect
Rhyme Recognition	1.50	0.44	Medium effect	0.33	0.10	No effect
Phonemic Blending	3.11	0.99	Large effect	1.85	0.55	Medium effect
Segmentation	2.32	0.85	Large effect	1.43	0.47	Medium effect
Total Bader Post-test Composite	6.93	0.95	Large effect	3.60	0.46	Medium effect

Growth Rate Findings

Bader

- The matched and nonequivalent samples yielded the same results regarding literacy growth rate comparisons between the treatment and control groups as measured by the Bader. Specifically:
    - Both the matched and nonequivalent samples showed that the UPSTART treatment group had stronger growth rates relative to controls on the Bader rhyme recognition, phoneme blending, and phoneme segmenting subtests and on the Total Bader Composite.

Table B.3
Bader Growth Rate Comparisons using the Matched Sample

Bader Test	**Control Group** (N=138[11])	**Treatment Group** (N=138)	**T-C Significance p≤.05**
	Mean Growth	**Mean Growth**	
Rhyme Recognition	1.0902	2.4203	**
Phoneme Blending	1.2180	4.5000	**
Phoneme Segmenting	1.3233	3.5072	**
Total Bader	3.6316	10.4275	**

Table B.4
Bader Growth Rate Comparisons using the Nonequivalent Groups Sample

Bader Test	**Control Group** (N=316)	**Treatment Group** (N=200)	**T-C Significance p≤.05**
	Mean Growth	**Mean Growth**	
Rhyme Recognition	1.1139	-2.2050	**
Phoneme Blending	1.3513	-4.0850	**
Phoneme Segmenting	1.2975	-3.2050	**
Total Bader	3.7627	-9.4950	**

Brigance

- The matched and nonequivalent samples both found that the UPSTART treatment group showed strong literacy growth rates relative to a control group within the Letter Knowledge, Letter Sounds, Auditory Discrimination, Survival Sight Words, and Basic Vocabulary subtests as measured by the Brigance.

- Both samples also found that the UPSTART treatment group showed stronger literacy development relative to the controls on the Total Brigance Composite.

- Both samples showed that there was no significant difference in growth rates between the control and treatment groups within the Expressive Grammar and Recites Alphabet subtests.

[11] The sample size fluctuated between 133-138 because of missing data for subscales.

- In addition, the nonequivalent group sample showed strong growth rates within the Expressive Objects, Receptive Objects, and Visual Discrimination subtests.

Table B.5
Brigance Growth Rate Comparisons using the Matched Sample

Brigance Test	Control Group (N=138)	Treatment Group (N=138)	T-C Significance p≤.05
	Mean Growth	Mean Growth	
Expressive Vocab	0.414	0.906	NS
Receptive Vocab	0.075	0.080	NS
Expressive Grammar	0.474	0.406	NS
Visual Discrimination	2.857	3.565	NS
Recites Alphabet	5.691	7.717	NS
Letter Knowledge	15.451	22.174	*
Letter Sounds	6.316	13.217	*
Auditory Discrimination	1.457	2.935	*
Survival Sight Words	1.075	1.986	*
Basic Vocabulary	2.083	8.370	*
Total Brigance	39.429	76.159	*

Table B.6
Brigance Growth Rate Comparisons using the Nonequivalent Groups Sample

Brigance Test	Control Group (N=316)	Treatment Group (N=200)	*T-C* Significance p≤.05
	Mean Growth	Mean Growth	
Expressive Vocab	0.2310	1.5100	*
Receptive Vocab	0.2342	0.6200	*
Expressive Grammar	0.5918	0.5900	NS
Visual Discrimination	2.7342	4.0400	*
Recites Alphabet	4.0032	6.5300	NS
Letter Knowledge	10.9937	21.6600	*
Letter Sounds	5.2880	11.9700	*
Auditory Discrimination	1.0633	2.8100	*
Survival Sight Words	1.3987	2.1000	*
Basic Vocabulary	3.6962	8.0550	*
Total Brigance	33.5538	73.9500	*