EDUCATION

THE UTAH STATE BOARD OF EDUCATION

Report to the Education Interim Committee

Utah Preparing Students Today for a Rewarding Tomorrow (UPSTART) Report

September 2019

Todd Call
Coordinator for Digital Teaching and Learning
todd.call@schools.utah.gov

Jennifer Throndsen
Director of Teaching and Learning
jennifer.throndsen@schools.utah.gov

Darin Nielsen
Assistant Superintendent of Student Learning
darin.nielsen@schools.utah.gov

Patty Norman
Deputy Superintendent of Student Achievement
patty.norman@schools.utah.gov

STATUTORY REQUIREMENT

U.C.A. Section 53F-4-407 requires the State Board of Education to make a report on UPSTART to the Education Interim Committee by November 30 each year. The State Board is required to contract with an independent evaluator to evaluate the program. Reporting on the program shall include the (i) number of families participating in the program including the number of families requesting and furnished computers; (ii) number of private and public preschool providers participating in the program; (iii) frequency of software usage; (iv) obstacles encountered with software usage, hardware, or providing technical assistance to families; (v) student performance on assessments as detailed in statute; and (vi) any other information that is part of the independent evaluation.

Utah Preparing Students Today for a Rewarding Tomorrow (UPSTART) Report

EXECUTIVE SUMMARY

During the 2017-2018 school year, Cohort 9 participated in the Utah Preparing Students Today for a Rewarding Tomorrow (UPSTART) program. The UPSTART program uses a home-based educational technology approach to develop the school readiness of preschool children. The program is designed to give Utah four-year-olds an individualized reading, mathematics, and science curriculum with a focus on reading. Children participate in the program the year before they attend kindergarten. The UPSTART program is administered by the Waterford Institute. A total of 14,278 preschool students participated in Year 9 of the program. Students in Cohort 9 used the UPSTART program for an average of 56 hours during the program year. Students who were UPSTART graduates used the program for an average of 58 hours. The independent evaluation for Cohort 9 of the program is attached.

UPSTART Program Evaluation

Year 9 Program Results

Submitted to the Utah State Board of Education
May 2019

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

(This page intentionally left blank)

Table of Contents

Executive Summary 2

COHORT 9 EVALUATION 2
FIRST GRADE ANALYSIS 6
SUMMARY AND RECOMMENDATIONS 7

Introduction 8

UPSTART PROGRAM DESCRIPTION 8

Cohort 9 Evaluation 10

RESEARCH QUESTIONS 10
METHODS 11
UPSTART PROGRAM IMPLEMENTATION 19
UPSTART PROGRAM IMPACTS ON LITERACY 26
SUMMARY AND DISCUSSION 36

First Grade Analysis 37

RESEARCH QUESTIONS 38
METHODS 39
FINDINGS 43
SUMMARY AND DISCUSSION 48

Summary and Recommendations 49

References 51

Appendix A: Comparison of C9 Evaluation Samples 53

Appendix B: Determining UPSTART Effect Size Benchmark 54

(This page intentionally left blank)

Executive Summary

Utah Preparing Students Today for a Rewarding Tomorrow (UPSTART) is a home-based computer preschool program developed and provided by the Waterford Institute to prepare young children for school entry and future academic success. The Evaluation and Training Institute (ETI), has prepared this report for the Utah State Board of Education (USBE) to document UPSTART's impact on students in its ninth year of implementation (Cohort 9, with students enrolled during the 2017-2018 program year). ETI responded to feedback and guidance from the UPSTART Advisory Committee (UAC), and we have continued our revised research design to meet a higher level of accountability for the program and explore longer-term aspects of UPSTART by reporting on two different areas:

- The **Cohort 9 Evaluation** presents information on program student outcomes and implementation results for Cohort 9 using a pre-test/post-test design with a statistically matched control group in order to assess the program's impact on developing children's early literacy skills. Our research findings cover two areas: (1) how the program was implemented and (2) what types of impacts the program has on children's literacy.
- The **First Grade Analysis** presents findings on UPSTART's continued impact on students' literacy achievement once children enter the elementary school setting. Using statewide data, we analyzed whether achievement gains from UPSTART that occur prior to school entry are sustained through kindergarten and into first grade.

This Executive Summary presents a summary of findings for each reporting area, along with selected recommendations for improving the program and future evaluation efforts.

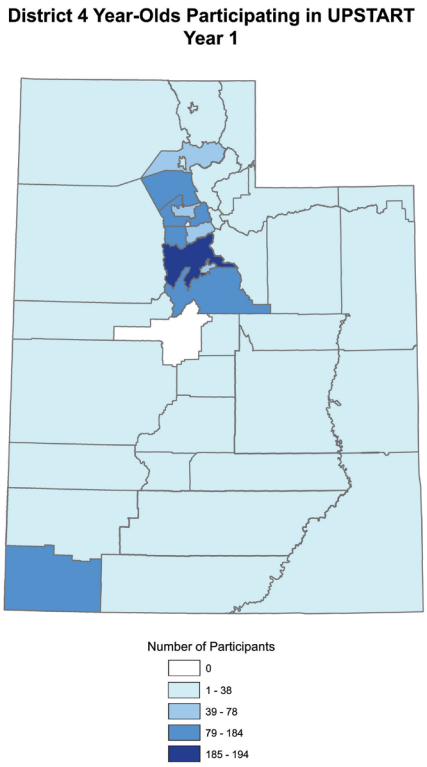
Cohort 9 Evaluation

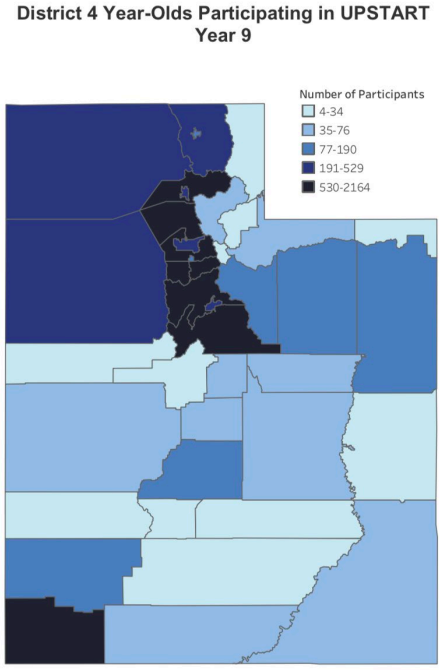
Program Implementation

The 2017-18 program year was a continued expansion of UPSTART enrollment, as the number of preschool students participating in the program in Year 9 (N = 14,278) grew by 3,533 students from the previous year (Year 8, N = 10,745), a 33 percent increase[1]. Over the past nine years, UPSTART program participation has increased, and the program has enrolled families in urban and rural areas throughout the state of Utah. The maps depicted below showcase UPSTART program participation by school district from the inception of the program (Year 1) to the most recent program year (Year 9).

[1] UPSTART participant enrollment and program usage data used to generate program implementation findings was provided to ETI by the Waterford Institute.

Maps of UPSTART program participation in Year 1 and Year 9 by School District





Forty-two percent of children enrolled in UPSTART Cohort 9 lived in families with incomes less than 200% of the federal poverty level and the majority of enrolled children were White (82%) and English speaking (92%). UPSTART enrollment increased from 10,745 children in Year 8 to 14,278 children in Year 9, an increase of 33 percent, while graduation rates remained constant at 89%.

Findings about UPSTART usage in Cohort 9 are summarized below:

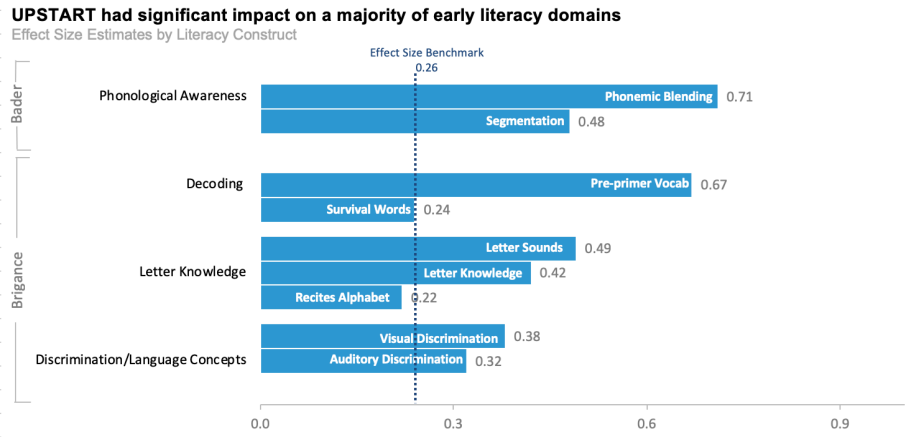
- Students who used the program for the recommended amount of time (or longer) had better reading outcomes than their matched counterparts who did not use the program.
- Students in Cohort 9 used the UPSTART program for an average of 54 hours during the program year. Students who were UPSTART graduates used the program for an average of 58 hours.
- Students in Cohort 9 had an 89% graduation rate, which reverses a trend of lower graduation rates year-to-year starting in Cohort 5 (which had a graduation rate of 94%) and continued in Cohort 6 (graduation rate of 92%) and Cohort 7 (graduation rate of 87%).

- Children who did not graduate were more likely to have parents with lower levels of education, speak a language other than English, be members of an underrepresented racial or minority group, have parents who were not married, and have higher levels of household poverty than children who graduated and completed the UPSTART program.
- A positive relationship was found between UPSTART curriculum use and evaluation outcomes: as program use increased, students' scores on literacy achievement measures increased.

Impacts on Literacy

We present effect sizes throughout our reporting to provide additional context for our findings. An effect size (ES) takes the difference between two group means on an outcome variable and represents it in standard deviation units. Effect sizes describe the magnitude of the difference between two groups, and essentially create a standardized scale to facilitate results interpretation. Following recommendations from the What Works Clearinghouse (WWC) (What Works Clearinghouse, 2017) and a meta-analysis of similar educational interventions and studies (Lipsey et al., 2012), we set an effect size threshold of .26 to denote effects that have practical significance and are substantively important.

UPSTART had a strong impact on children's emerging literacy skills based on results from effect size and growth score analyses. Children enrolled in UPSTART produced significant positive effects compared to control children on the Brigance composite, an instrument that measures decoding skills, letter knowledge, vocabulary and syntax, and pre-literacy discrimination (ES = .53). Similarly, UPSTART participants experienced large effects on the Bader composite, an instrument that assesses children's phonological awareness (ES = .56). The graph below presents effect sizes by literacy construct and provides a line marker to highlight effect sizes that fall above the predetermined threshold (.26 or higher) to showcase their practical significance.



Phonological awareness has been identified as one of the most important predictors of reading success and involves a child's facility with the sound structure of words (Phelps, 2003). Phonological skills include the ability to identify rhyming words, isolate a sound in a word, blend individual sounds, and detect word alliteration. Children's **phonological**

awareness abilities were significantly improved because of their UPSTART participation.

- UPSTART students had significantly higher phonemic blending skills (ES = .71) and phoneme segmenting skills (ES = .48).
- Compared to control children, students participating in UPSTART had significantly higher increases from the pre-test to the post-test on both phonological awareness subscales (blending and segmenting).

UPSTART had a significant impact on children's **word decoding** skills. Decoding, a core reading skill that is a precursor to reading fluency, is the ability to accurately identify individual printed words. Accurate decoding results from the successful acquisition of several key pre-literacy skills, including a child's ability to recognize written letters, discern letters that correspond to phonological sounds, and blend word sounds into the generation of a single word.

- Children participating in UPSTART had significantly higher post-test scores on decoding pre-primer vocabulary words (ES = .67) and on reading survival sight words (ES = .24).
- UPSTART children had stronger growth scores on reading pre-primer vocabulary words (e.g., "can", "and", "do") and survival sight words (e.g., "go", "stop", "out") compared to children who were not enrolled in the program.

Students who participated in UPSTART experienced a moderate improvement in their **letter knowledge** skills. The letter is the most basic unit of reading and familiarity with the letters of the alphabet has been shown to be a strong predictor of reading achievement. Additionally, understanding the connection between written letters and the sounds of speech is a precursor to decoding.

- UPSTART children had small to medium effects in their learning how to recite (ES = .22), identify (ES = .42), and sound out (ES = .49) letters of the alphabet.
- Compared to control students, UPSTART participants showed significantly stronger growth rates in learning how to pronounce letter sounds.

Before children can read, they need to be able to visually distinguish between shapes, letters, and words, even if they do not fully comprehend what letters represent. Similarly, children should be able to differentiate between spoken words (e.g., "fit" versus "fat") before comprehending written words. 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 = .38) as well as a small to medium effect on their ability to distinguish if two words sound the same (ES = .32).
- Children in UPSTART had stronger growth scores on their auditory discrimination of words when contrasted to children not enrolled in UPSTART.

The UPSTART program did not have a significant impact on children's **vocabulary**:

- UPSTART did not have significant effects on receptive vocabulary.
- Children enrolled in UPSTART did not have significantly different growth rates on vocabulary subscales when compared to control children.

First Grade Analysis

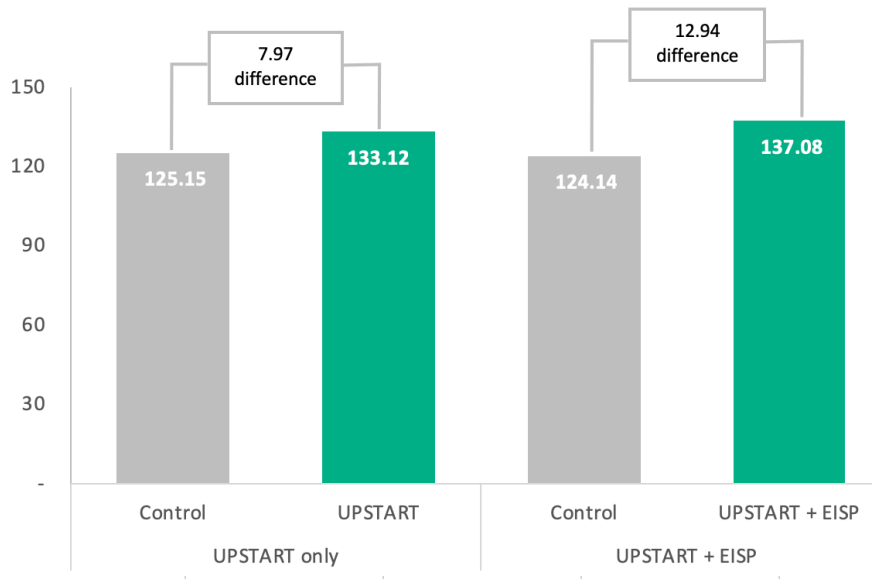
In order to determine whether or not UPSTART has a sustained benefit on children's literacy once they enter elementary school, UPSTART students and their counterparts who did not have any UPSTART experience were followed through kindergarten and first grade. To conduct this analysis, we had to address potentially confounding effects from the Early Software Intervention Program (EISP), a statewide computer-based literacy instruction software program available in grades K-3. To control for the impacts of EISP, we excluded any student who participated in EISP as a kindergartener from our control group. We utilized a post-test only design to determine if UPSTART participants had higher scores on the first grade DIBELS literacy assessment compared to similar comparison students. In an effort to isolate the effects of participating in EISP, we excluded any student who participated in EISP as a kindergartener from our control group. In addition, to give the state information about potential multiplier effects, we created two treatment groups: students who only participated in UPSTART during their preschool year (UPSTART only) and students who participated in UPSTART as preschoolers and who participated in the EISP program as kindergarteners (UPSTART + EISP).

Findings show that, on average, students who used the UPSTART program in preschool scored 7.97 points higher than comparison students on the DIBELS composite at the beginning of first grade. This difference was statistically significant and produced an effect size of .17.

Students who received continuous treatment from preschool through kindergarten did even better: Our findings show that the use of UPSTART + EISP had more of an impact on first grade reading than the UPSTART preschool program alone. Students with combined treatment in preschool and first grade scored 12.94 points higher than a group of matched comparison students, a statistically significant mean score advantage over their non-program peers that produced a .28 effect size (which is above our .26 effect size criteria to show practical significance for literacy achievement).

UPSTART students outperformed counterparts in first grade

1st Grade DIBELS Composite Scores by Treatment Group



Summary and Recommendations

The UPSTART program continues to show success in helping preschool aged children develop literacy skills in preparation for their entry into kindergarten, and new analyses suggest that UPSTART has benefits that last into elementary school.

Children who did not graduate from UPSTART were more likely than program graduates to reside in households below the poverty level, have parents with lower levels of education, and be English learners – the ideal target population for UPSTART and the children that stand to benefit most from the program. Cohort 9 maintained a graduation rate of 89%, which is noteworthy considering that the overall numbers of UPSTART participants increased by 33%**.** Monitoring of program use and graduation requirements needs to be continual to be sure that UPSTART is being administered with fidelity so that all children can receive the full program dosage as recommended by the vendor and obtain the full cognitive benefits of the program.

Due to the strong impact on early literacy development, we recommend that the state continue to provide the UPSTART program to children. Given the importance of graduation on literacy achievement outcomes, we recommend that the program vendor continue to work with the evaluator and USBE staff to monitor program implementation carefully and be sure the trend towards higher graduation rates continues. Specifically, we recommend that the program vendor consider the following recommendations:

- The program vendor could develop new strategies for addressing falling usage and graduation rates among the most at-risk students (i.e. those with high levels of poverty and with English as a second language). Some potential strategies might include:
    - Create peer support partnerships with similar community groups in high risk geographic locations to discuss strategies for increasing children's program use.
    - Increase communication with high risk families to evaluate potential barriers to program usage
    - Developing targeted incentives for families with the highest risk factors for not meeting program usage requirements, such as monthly awards (extrinsic), being highlighted in UPSTART communications to social networks as "Gold Star Families" (intrinsic).

The fact that UPSTART children maintained their advantage over their comparison counterparts through first grade can be construed as another important benefit of the UPSTART program. Findings from the first grade analysis indicate a continued effect of the UPSTART preschool program, and this effect is increased with continued individualized computer-based literacy instruction throughout kindergarten (i.e. in conjunction with EISP). As the UPSTART program expands to reach more Utah preschoolers across the state, we recommend that USBE continues the EISP program to provide individualized instruction that builds on the gains created by UPSTART.

Introduction

The Utah State Board of Education (USBE) 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.

The 2017-2018 UPSTART program year saw the program's scale increased to reach more families than in any previous cohort to date. As the program scaled-up, the evaluation had to be adapted to accommodate larger numbers of program students and the higher stakes as a result of the greater resource allocation for the program. While the scale and stakes increased, our research objectives remained constant: we continue to evaluate the program's *impact on developing children's early literacy skills* to help the state and stakeholders determine the benefits from participating in the program.

We enhanced the established evaluation design and reporting in three key ways to meet a higher level of accountability for the program, and to ensure that the program resources were having a positive impact on school readiness and beyond.

> In the **Cohort 9 Evaluation**, we present outcome results for UPSTART's ninth and largest year of implementation, hereafter referred to as Cohort 9 (C9). Additionally, we document the extent to which participants used the computerized curriculum as it was intended, establish the relationship between curriculum usage and literacy outcomes, and report the program's completion or "graduation" rate. As in our evaluations with recent cohorts, the Cohort 9 evaluation included a statistically balanced match of treatment and control students. While requiring a larger sample size, the matching process enhances our ability to detect treatment effects and, in general, improve the accuracy of the evaluation results.
>
> Second, in addition to determining the impact of the UPSTART program on students' school readiness prior to the beginning of kindergarten, this report will also present findings on UPSTART's continued impact on student literacy with kindergarten and first grade literacy scores in the **First Grade Analysis**. This longitudinal study meets the provision in state law to evaluate the long-term impacts of UPSTART on students and uses DIBELS literacy data collected in schools from over 38,000 students to determine whether or not UPSTART has a lasting impact on student literacy achievement.

Each of these analyses is presented in separate sections of the report, along with an overall summary and suggestions for program recommendations. We begin with a brief overview of the UPSTART preschool program.

UPSTART Program Description

Utah Preparing Students Today for a Rewarding Tomorrow (UPSTART) is a 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 ninth year of operation during the 2017-18 school year, the project's implementation contractor – the Waterford Institute – enrolled 14,278 preschool children and provided them with an adaptive program of computer-based early literacy instruction to prepare them academically for kindergarten. The 14,278 children enrolled in the ninth-year cohort,

hereafter referred to as Cohort 9 (C9), participated in UPSTART from September 2017 through May 2018. Cohort 9 is the largest group since the program's rollout.

The UPSTART software uses adaptive lessons, digital books, animated songs, and activities to deliver individualized early literacy content. The reading skills taught by the Waterford Early Learning Program at Level 1 of the curriculum[2] include:

- Phonological Awareness
- Phonics
- Comprehension and Vocabulary
- Language Concepts

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.

[2] 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.

Cohort 9 Evaluation

Research Questions

Our evaluation of the Cohort 9 of UPSTART users 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 C9 evaluation were as follows:

> **Research Question 1.1:** Do UPSTART students have better early literacy skills at kindergarten compared to control group students?
>
> **Research Question 1.2:** Do UPSTART students show stronger literacy growth rates from preschool to kindergarten compared to control group students?

In the impact analysis, the outcomes of interest were measures of early literacy skills relevant to emerging readers such as phonological awareness, letter recognition, letter sound knowledge, and vocabulary development. Results for research questions 1.1 and 1.2 are presented in the **UPSTART Program Impacts on Literacy** section of the report.

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

> **Research Question 1.3:** 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?
>
> **Research Question 1.4:** What percent of the participants completed the full implementation program (i.e., “graduated” as defined by the Waterford Institute)?
>
> **Research Question 1.5:** How does the level of UPSTART curriculum usage relate to reading readiness outcomes?

Data for research questions 1.3 and 1.4 were obtained from records maintained by the Waterford Institute and are answered in this report by descriptive statistics. The answer to research question 1.5 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 1.3 through 1.5 are presented in the **UPSTART Program Implementation** section of the report.

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 facilitated 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 began as nonequivalent by definition, and consequently selection bias could be assumed to operate to some degree in some manner. The pre-test allowed 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).

C9 Evaluation Samples

The C9 evaluation moved from using an unmatched group seen in previous years to a new approach first adopted in the C6 evaluation that uses 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.

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 the matched sample is described in more detail below.

Data Collection

We collected data from two groups of preschoolers, treatment children who had enrolled in UPSTART for Year 9 of the program (the 2017-18 school year) and nonparticipating control group children. The children were not randomly assigned to the treatment or control groups.

Treatment children. The UPSTART children came from an initial random sample of C9 UPSTART enrollees whose families were contacted about participating in the C9 evaluation[3]. Because the legislation extending the UPSTART program gave participation priority to low-income families and non-native English speakers (Utah Code: 53A-1a-1001), we similarly prioritized recruiting low-income families in our treatment sample. The recruited UPSTART children participated in pre-testing prior to entering the program over the summer of 2017 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, low-income housing units, and snowball sampling[4] from families who were UPSTART users.

Because the treatment and control groups were not created through random assignment, it was assumed that the two groups would be nonequivalent on factors that may influence literacy skills. Therefore, it is important to review the treatment and control demographics and pre-test scores carefully to statistically adjust for any imbalances so that accurate and fair comparisons can be made.

We created two analytic files for our data analysis. The Brigance data file consisted of 248 treatment children and 248 control children that were matched based on Brigance pre-test scores and other demographic characteristics. The Bader data file contained data from 430 preschool children (215 treatment children and 215 control children) matched on Bader pre-test performance. The inclusion of the two separate data files allowed us to better estimate the impact of the UPSTART program on early literacy and phonological awareness skills. **Tables 1.1** and **1.2** presents key demographic characteristics by the unmatched treatment and control samples by outcome of interest (i.e., the Brigance sample or Bader sample). As shown in **Tables 1.1** and **2.2**, control families were somewhat more advantaged compared to treatment families from the standpoint of parental education and household income level.

[3] C9 treatment families were screened based on location, parental education, income level, child language, and known disabilities.
[4] Snowball sampling is when existing participants recruit future participants among their personal network of acquaintances.

Table 1.1
Brigance Unmatched Treatment-Control Comparisons on Key Demographics

Demographic Categories		Treatment (N=276)	Control (N=611)
Gender	Female	48%	54%
	Male	52%	46%
Ethnicity	White	83%	82%
	Hispanic	16%	18%
Child Language	English	96%	98%
Parent Education Level	High School Diploma	21%	16%
	Some College	69%***	48%
	Bachelor's degree	6%	26%***
Parent Marital Status	Married	83%	80%
Household Income	Under $10,000	3%	5%
	$10k-$24,999	10%	16%**
	$25k-$49,999	35%	31%
	$50k-$74,999	42%**	27%
	$75k-$99,999	9%	15%*
	$100k or more	1%	7%**

*p<.05, **p≤.01, ***p≤.001

Table 1.2
Bader Unmatched Treatment-Control Comparisons on Key Demographics

Demographic Categories		Treatment (N=275)	Control (N=593)
Gender	Female	48%	54%
	Male	52%	46%
Ethnicity	White	83%	84%
	Hispanic	16%	16%
Child Language	English	98%	96%
Parent Education Level	High School Diploma	21%**	14%
	Some College	69%***	48%
	Bachelor's degree	6%	30%***
Parent Marital Status	Married	83%	83%
Household Income	Under $10,000	3%	3%
	$10k-$24,999	10%	15%
	$25k-$49,999	35%	28%
	$50k-$74,999	42%**	32%
	$75k-$99,999	9%	16%*
	$100k or more	1%	8%***

*p<.05, **p≤.01, ***p≤.001

There were significant differences between the two unmatched groups on household income and parent education level. 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). In light of these findings, it is important to ensure that the treatment and control groups are as comparable as possible with regard to parental

education before analysis or that statistical adjustments are performed to determine any impact of family characteristics on post-test literacy outcomes.

Significant differences between the treatment and control groups that favored the control group were found on both the Brigance and Bader pre-test literacy instruments. While the use of a pre-test and covariates with the full unmatched 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 that took into account Brigance and Bader pre-test performance along with key demographic characteristics 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 unmatched C9 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 were able to construct a comparison group of control children that resembled 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 C9 unmatched evaluation Brigance sample contained data from 276 treatment students from C9 and 611 comparison students who did not participate in the UPSTART program. The unmatched evaluation Bader sample contained data from 275 treatment children and 593 comparison students.
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:
    - Gender (Female/Male)
    - Ethnicity (White/Hispanic),
    - Language
    - Parent Education
    - Brigance Composite pre-test scores
3. Statistical tests assessed the balance between treatment and control group to ensure groups were as similar as possible at baseline (pre-test).

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. A similar procedure was performed using Bader pre-test scores to create a second analytic sample of matched treatment and comparison students for measuring impacts on the Bader assessment. **Table 2.1** displays the demographic breakdown of the matched treatment and control groups on the Brigance test. **Table 2.2** displays the demographic breakdown of the matched treatment and control groups on the Bader test. Note how the two groups in the matched sample are much more similar in terms of parental education and race than in the unmatched sample.

Table 2.1
Brigance Matched Treatment-Control Comparisons on Key Demographics

Demographic Categories		Treatment (N=248)	Control (N=248)
Child Gender	Female	49%	56%
	Male	51%	44%
Child Ethnicity	White	83%	81%
	Hispanic	17%	20%
Child Language	English	98%	97%
Parent Education Level	High School Diploma	20%	16%
	Some College	70%	70%
	Bachelor's degree	6%	6%
Parent Marital Status	Married	84%	81%
Household Income	Under $10,000	3%	6%
	$10k-$24,999	10%	10%
	$25k-$49,999	35%	33%
	$50k-$74,999	42%	42%
	$75k-$99,999	9%	9%
	$100k or more	1%	1%

*p<.05, **p≤.01, ***p≤.001

Table 2.2
Bader Matched Treatment-Control Comparisons on Key Demographics

Demographic Categories		Treatment (N=215)	Control (N=215)
Child Gender	Female	51%	51%
	Male	49%	49%
Child Ethnicity	White	82%	86%
	Hispanic	19%	16%
Child Language	English	99%	99%
Parent Education Level	High School Diploma	19%	19%
	Some College	72%	72%
	Bachelor's degree	7%	7%
Parent Marital Status	Married	84%	84%
Household Income	Under $10,000	3%	3%
	$10k-$24,999	10%	10%
	$25k-$49,999	38%	38%
	$50k-$74,999	39%	39%
	$75k-$99,999	9%	9%
	$100k or more	1%	1%

*p<.05, **p≤.01, ***p≤.001

Outcome Measures

The reading skills taught by the Waterford Early Learning Program at Level 1 of the curriculum[5] 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*, *pre-literacy discrimination*, *letter knowledge*, and *decoding* and are described in detail in **Table 3**. 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 phonological awareness subtests (rhyming, phonemic blending, and phoneme segmentation), along with a composite summary phonological awareness score that was calculated by adding the scores from the subtests.

Relevance of Outcome Measures. As stated previously, we selected our outcome measures based on their alignment to the UPSTART curriculum and on their ability to assess early literacy skills that are demonstrated predictors of reading success. Each outcome measure evaluates a key domain or construct of early literacy: pre-literacy discrimination, phonological awareness, letter knowledge decoding, and vocabulary. These five constructs are explained in further detail below.

> *Pre-Literacy Discrimination*. Before children can read or even comprehend the meaning of letters, they need to be able to visually discriminate between letter shapes. For example, if a child is unable to visually distinguish a "p" from a "b", she will incorrectly identify letters and their letter sounds. Similarly, children need to be able to discriminate between the sounds of words (e.g., "cat" from "can") to

[5] 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.

facilitate listening comprehension and to match letter and word sounds with their printed versions.

Phonological Awareness. Phonological awareness has been identified as one of the most important predictors of reading success and involves a child's facility with the sound structure of words (Phelps, 2003). Phonological skills include the ability to identify rhyming words, isolate a sound in a given word, blend individual sounds to produce a single, and detect word alliteration. We assessed the phonological awareness with two subscales from the Bader: phoneme segmentation and phoneme blending.

Letter Knowledge. Letters are the most basic unit of reading and familiarity with the alphabet and ability to recognize letters and their corresponding sounds is a prerequisite for decoding. Letter knowledge begins with being able to identify lower and uppercase letters in a variety of fonts, but also includes understanding the representational nature of letters and connecting printed letters with their phonemic sounds. Letter knowledge is evaluated in the current study by assessing children's ability to recite the alphabet, identify lowercase letters by name, and connect lowercase letters with their sounds.

Decoding. Decoding is the process of translating printed words into speech and is the precursor to reading fluency, the ability to read text accurately and quickly, either aloud or silent. Decoding relies on the successful acquisition of all the aforementioned reading skills, phonological awareness, letter knowledge, and pre-literacy discrimination. We measured decoding in the UPSTART study by asking children to read lists of simple pre-primer vocabulary (e.g., "and", "can", "go", "look") and presenting them with words they might have seen in their everyday lives (e.g., "stop", "in", "out").

Vocabulary. Vocabulary has been demonstrated to be a reliable predictor of later reading scores (Snow, Burns, & Griffin, 1998) and is necessary for making meaning of written and oral language. Children's vocabulary is measured by an expressive vocabulary test where they provide names to a series of pictures.

Table 3 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 3

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
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	Phonemic Blending	Blends separate word sounds into single word	0-8
		Phoneme Segmentation	Segments word into separate word sounds	0-8
		Rhyme Recognition	Identify rhyming words	0-10
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 Procedures

Data were collected for treatment group children who had enrolled in UPSTART for Year 9 of the program and 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 consisted of Bader data from 430 children, 215 treatment and 215 control, and Brigance data from 496 children, 248 treatment and 248 control, and 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.

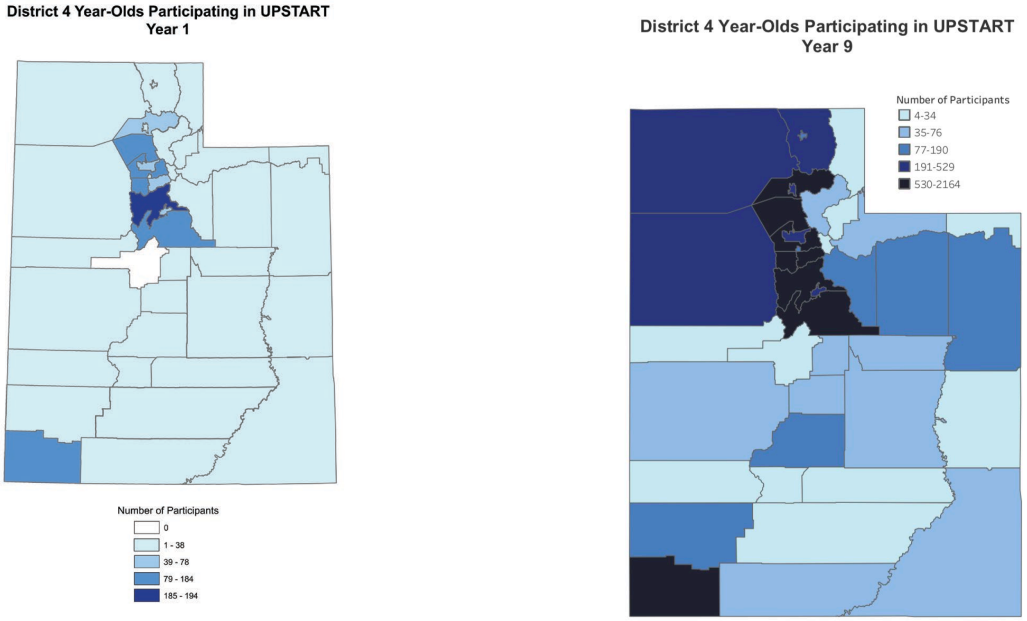
UPSTART Program Implementation

Findings reviewed in the UPSTART implementation section include ninth 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 2017-18 program year was a continued expansion of UPSTART enrollment, as the number of preschool students participating in the program in Year 9 (N=14,278) grew by 3,533 students from the previous year (Year 8, N=10,745), a 33 percent increase. Since the inception of the program, the number of students enrolled in the program rose from 1,248 children in Year 1 to 14,278 students in Year 9, an increase of over 1,000 percent. The maps depicted in **Figure 1** showcase UPSTART program participation by school district from the inception of the program (Year 1, N=1,248) to the most recent Year 9 (Year 9, N=14,278). As seen below in **Figure 1**, the UPSTART program has continued to further its reach over the past nine years and has increased enrollment in both urban and rural areas of the state.

Figure 1
Maps of UPSTART program participation in Year 1 and Year 9 by School District



The Waterford Institute provided a comprehensive dataset to ETI for the ninth-year UPSTART enrollment of 14,278 children, including demographic information, provisioned educational technology, UPSTART program usage, and whether or not children completed program requirements. This provisioned data was analyzed by ETI to generate the findings related to program implementation.

Some basic demographic characteristics of the C9 population are presented below in **Table 4**.

Table 4
Demographic Characteristics of C9 Population

Demographic Categories		All C9 UPSTART (N=14,278)
Child's Gender	Male	52%
	Female	48%
Child's Ethnicity	White	82%
	Hispanic	11%
	Asian/Pacific Islander	2%
	African American	1%
	Native American	<1%
	Other	3%
Child's Language	English	92%
	Spanish	7%
	Other	1%
Parent Educational Attainment	Some High School	13%
	High School Graduate	13%
	Some College	34%
	College Graduate	39%
	Advanced Degree	10%
Parent Marital Status	Married	91%
	Otherwise	9%
Household Poverty Level	Below 100%	13%
	Below 185%	37%
	Below 200%	42%

Note: Percentages may not add to 100% due to rounding. Data is from Waterford participant records.

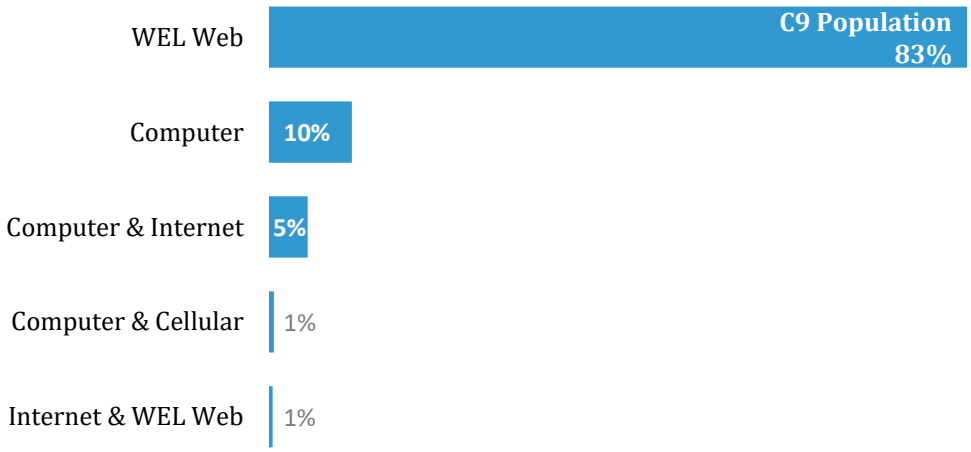
Slightly more C9 boys (52%) were enrolled than girls (48%) and in terms of ethnicity, the majority (82%) of the C9 enrollment was White, with 11% of the children being of Hispanic origin. Thirty seven percent of the C9 UPSTART participants lived in families with incomes less than 185% of the federal poverty level.[6]

[6] The federal poverty definition consists of a series of thresholds based on family size. In 2017, a 100% poverty threshold for a family of four was $24,600, while a 185% threshold for a family of four was $45,510 (see U.S. Department of Health and Human Services poverty guidelines at https://aspe.hhs.gov/2017-poverty-guidelines).

Provided UPSTART Equipment

The type of education technology provided to UPSTART children in Year 9 of the program is shown in **Figure 2** for all 14,278 children enrolled in the program. The majority of UPSTART children (83%) used the Waterford website to retrieve the UPSTART program. This allowed families to access the UPSTART curriculum from their home computers.

Figure 2. Equipment provided to C9 Participants by Waterford



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

Second most frequently, UPSTART provided free personal computers to 10% of the C9 children while they participated in the program. Another 5% of the C9 program participants were provided with internet access and personal computers. The remaining two percent of the C9 enrollment received computers and wireless access (1%), internet and access to the Waterford website (1%) or participated in a lending library program (less than 1%) 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 ninth year of UPSTART are summarized in **Table 5** and are compared to program "graduates". The average level of usage for all students enrolled in the ninth year of UPSTART (N=14,278) was approximately 54 hours of instruction. The C9 academic year covered 40 weeks of instruction, beginning the week of September 4, 2017 and ending May 28, 2018.

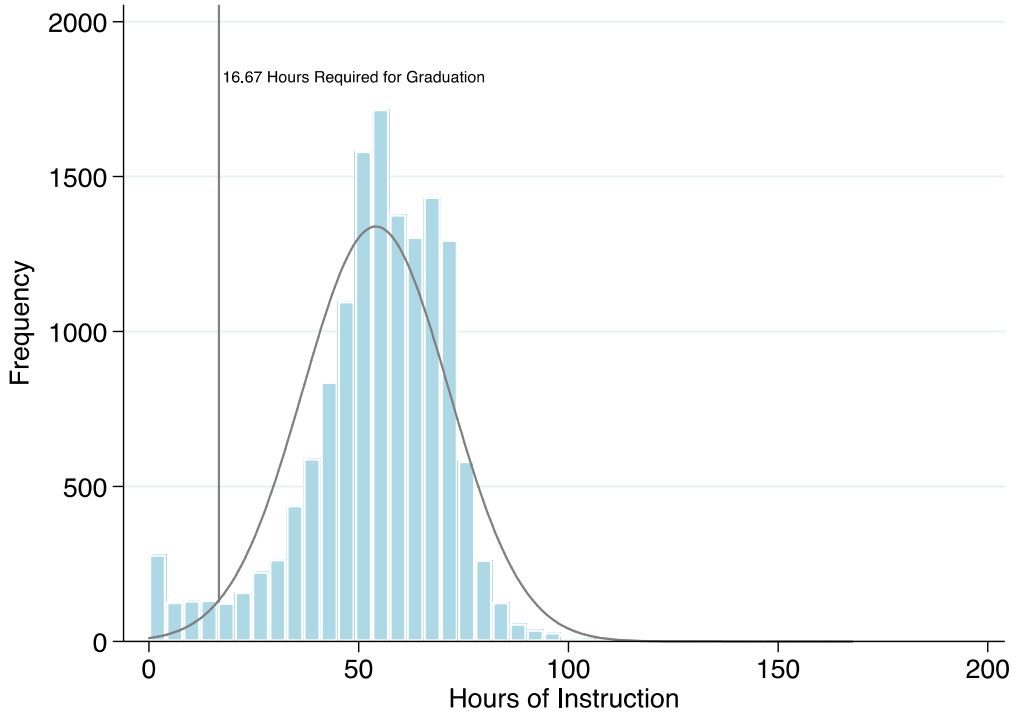
Table 5
C9 Hours of UPSTART Instruction

Group	N	Mean	SD	Range
All UPSTART	14,278	54.10	17.41	00.00 – 167.80
UPSTART Graduates	12,713	58.36	12.44	16.75 – 167.80
UPSTART Analysis Sample	248	57.08	14.32	4.95 - 88.32

Ninety-eight of the 14,278 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 9 of the program. 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 eighth year of the program (Evaluation and Training Institute, 2018). The children in the C9 evaluation analysis sample used the UPSTART curriculum for approximately 59 hours of instruction on the average (see **Table 6**).

The histogram in **Figure 3** shows the distribution of hours of instruction for the total C9 population (N=14,278). Ninety eight of the enrolled children logged zero hours of instruction during their time in UPSTART. At the other end of the spectrum, thirty-five children logged over 100 hours of instruction.

Figure 3. Hours of Instruction for C9 Families



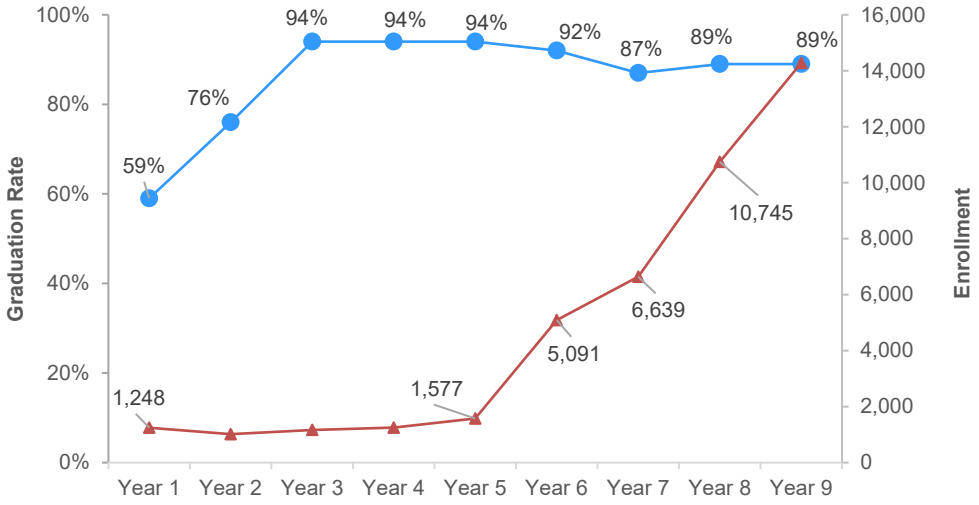
The bottom quartile of the C9 population completed 46.08 hours of instruction or less, the midpoint of the C9 distribution was 55.77 hours, and the top quartile completed in excess of 66.30 hours of instruction.

UPSTART Graduation Rate

Of the 14,278 children documented as enrolled in UPSTART in the ninth year of the program, the Waterford Institute classified 12,713 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. UPSTART graduate status was significantly correlated with hours of instruction (r = .69) and with the number of weeks in the program (r = .61).

By these usage requirements, Cohort 9 achieved a graduation rate of 89% (i.e., 12,713/14,278 = 0.89). As seen in **Figure 4**, this graduation rate is the same as the previous year, (89%) even in the face of increased enrollment, but slightly lower than the graduation rates that hovered between 92% and 94% in the initial pilot phase of the program that enrolled approximately 1,500 students in Years 3 through 5 and 5,000 students in Year 6.

Figure 4. UPSTART Graduation Rates and Enrollment



In order to further examine the features of program graduates and non-graduates, **Table 6** displays the demographic characteristics of UPSTART graduates and non-graduates in Cohort 9. Children who did not meet the program usage requirement were more likely than UPSTART graduates to speak a language other than English, be a member of an underrepresented racial or ethnic minority group, have parents with lower levels of education, reside in families with parents who were not married, and have higher levels of poverty.

Table 6
Demographic Characteristics of C9 Population

Demographic Categories		UPSTART Graduates (N=12,713)	UPSTART Non-Graduates (N=1,565)
Child's Gender	Male	51%	52%
	Female	48%	48%
Child's Ethnicity	White	83%	72%
	Hispanic	10%	17%
	Asian/Pacific Islander	2%	1%
	African American	1%	1%
	Native American	>1%	1%
	Other	2%	3%
Child's Language	English	92%	86%
	Spanish	6%	11%
	Other	1%	3%
Parent Educational Attainment	Some High School	3%	10%
	High School Graduate	12%	21%
	Some College	34%	41%
	College Graduate	41%	26%
	Advanced Degree	9%	5%
Parent Marital Status	Married	92%	80%
	Otherwise	8%	20%
Household Poverty Level	Below 100%	12%	27%
	Below 185%	35%	52%
	Below 200%	40%	56%

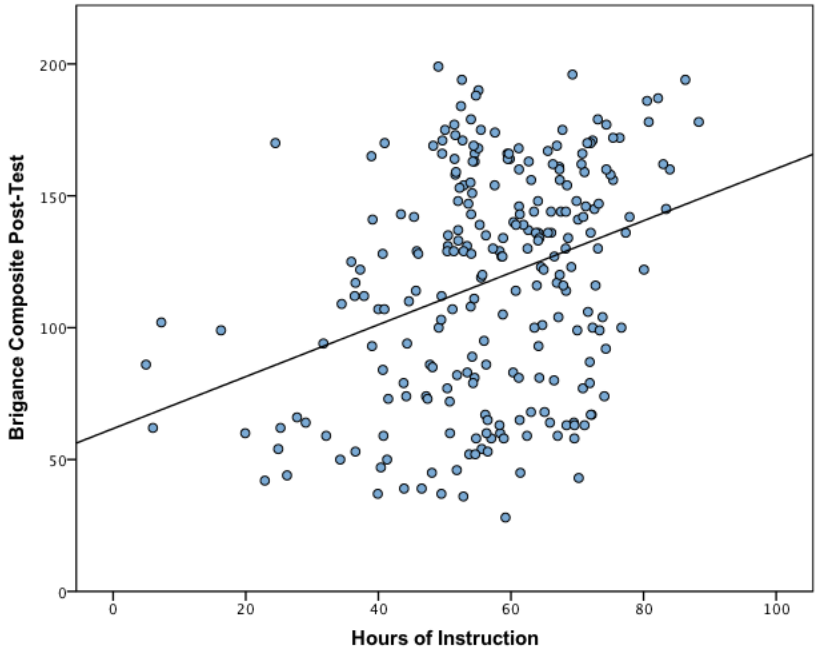
Note: Percentages may not add to 100% due to rounding. Data is from Waterford participant records.

UPSTART Usage and Literacy Outcomes

Similar to previous years, the ninth-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 small positive relationship between UPSTART usage (measured in hours of instruction) and Brigance post-test scores (r=.33). 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, a correlation analysis of the relationship between hours of UPSTART instruction and Bader composite post-test scores indicates a positive linear association between instruction time and scores on the Bader post-test (r = .23). This suggests that the acquisition of early phonological skills as measured by the Bader also tended to improve with increasing levels of exposure to UPSTART curriculum.

UPSTART Program Impacts on Literacy

This section includes results based on statistical comparisons of literacy achievement (test scores) for matched treatment and control groups during the ninth year of UPSTART implementation. 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.

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

> **Research Question 1.1:** *Do UPSTART students have better literacy skills at Kindergarten than control students?*
>
> **Research Question 1.2:** *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[7]. We conducted a series of models that explored the impact of household income level on the outcomes of interest and the results were not meaningfully different from our initial analysis.

[7] To create a concise report that highlights the most important findings for stakeholders, we did not present findings that were non-significant in figures.

Accordingly, we chose the simplest data analytic model to test for group differences because it offered ease of interpretation for multiple audiences and more complicated models were not needed to compare differences between the treatment and control group.

Effect Sizes: An Overview

We present effect sizes throughout our reporting to provide additional context for our findings. An effect size (ES) takes the difference between two group means on an outcome variable and represents it in standard deviation units. For example, an effect size of .30 would indicate that the difference between a treatment and control group is .30 standard deviation units. Effect sizes describe the magnitude of the difference between two groups, and essentially create a standardized scale so the results are easy to interpret and have meaning. In previous reports, we have interpreted effect sizes according to Cohen's (1988) general categorization of effect sizes as small (0.2), medium (0.5), and large (0.8) as a general rule of thumb.

However, it is important to note that Cohen's broad categories were designed for a range of effect sizes across a wide spectrum of social and behavioral research and are not specifically tailored for education interventions, studies, or samples. A more appropriate and meaningful benchmark for assessing the significance of an intervention's effect size is to compare it with the effects found for similar education interventions with comparable research samples and outcome measures (Lipsey et al., 2012). If an effect is larger than those of similar interventions, it has practical significance by virtue of being larger than previously reported effect sizes. Conversely, if an effect size is lower than comparable interventions and education research studies, then the impact may not be as impressive or significant.

How then, do we determine appropriate benchmarks for interventions similar to UPSTART? Researchers at the U.S. Department of Education's Institute of Education Sciences (IES) reviewed 829 effect sizes from 124 education research studies and determined that the average effect size for an evaluation that used a standardized subject outcome measure (like the Brigance/Bader) to assess a comprehensive educational intervention program that targeted individual students like UPSTART was .26 (Lipsey et. al, 2012). We provide this benchmark to contextualize the effect sizes presented in this report and to aid the reader in determining the practical significance of the effect of UPSTART – any effect size above .26 is higher than the average effect size seen in similar education evaluations. **Appendix B** provides greater detail on how the benchmark was determined. Our .26 threshold is similar to the benchmark specified by the What Works Clearinghouse (WWC), a federally funded initiative at IES that reviews educational research and interventions. The WWC considers effect sizes of .25 or larger to be "substantively important" and a qualified positive (or negative) effect, even if they do not reach statistical significance (What Works Clearinghouse, 2017).

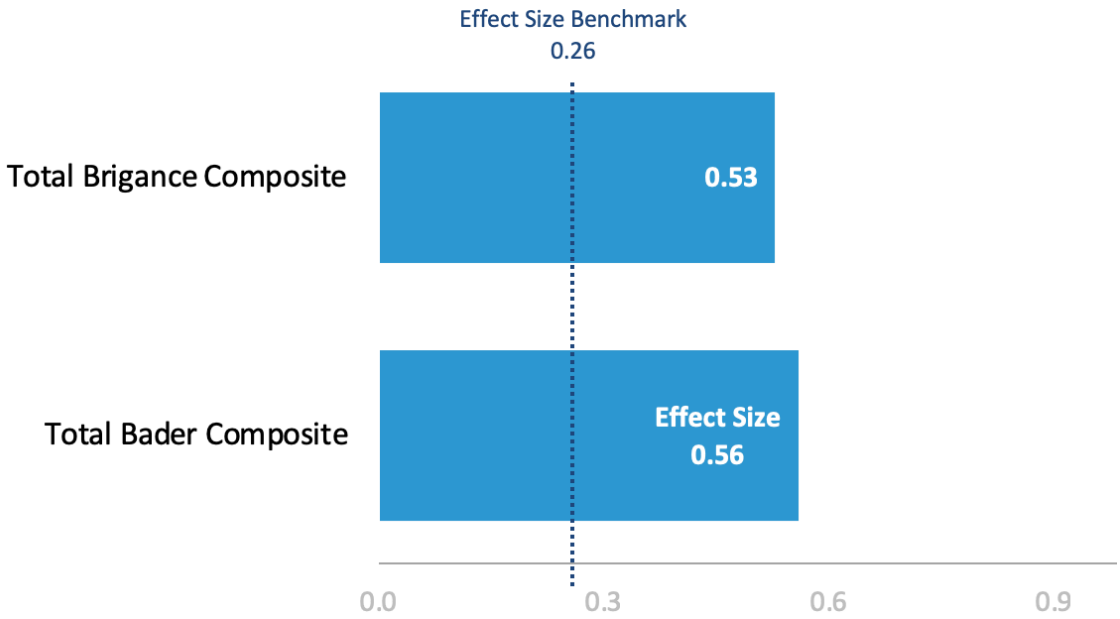
> **Research Question 1.1**: *Do UPSTART students have better literacy skills at entry to Kindergarten than control students?*

In order to demonstrate the impact of the UPSTART program, we present effect sizes that highlight the differences between UPSTART participants and a matched control group on post-test literacy measure.

Effect sizes[8] were calculated to show the magnitude of UPSTART's impact at post-test as measured by each of the 11 literacy subtests (8 Brigance subtests and 3 Bader subtests), and the Total Brigance and Bader Composites (composites include aggregated results of the subtests). **Graphs of effect sizes in this report provide a line marking the .26 benchmark to provide context and to showcase findings that have practical significance. Effect sizes with statistical significance ($p < .05$) are presented with blue bars.**

Combined post-test results showed that UPSTART participation had a medium impact on students' early literacy skill development. In the matched post-test sample[9] (N=496), UPSTART produced strong to medium effects (.56 and .53) as measured by the total Bader and Brigance composite scores that are well above the observed .26 effect size for similar interventions and evaluation studies (see **Figure 6**).

Figure 6. Brigance and Bader Posttest Analysis of Composite Scores



On average, children participating in UPSTART scored 60.19 points on the Brigance Composite before beginning the program and 118.05 points on the Brigance after the program was completed. Conversely, children who were not enrolled in UPSTART scored 60.58 points on the Brigance pre-test and 95.70 points on the Brigance post-test.

With regard to the Bader Composite, UPSTART children scored 4.34 points on the instrument at pre-test and 12.33 points at post-test, while their comparison counterparts scored 4.42 points on the Bader Composite pre-test and 8.56 points on the Bader post-test.

[8] 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.
[9] Brigance Treatment Group (N = 248); Control Group (N = 248)

UPSTART children scored significantly higher than control children on seven of the eight Brigance tests and two of three Bader subtests on the post-test, showing strong empirical evidence that UPSTART was successful in helping children develop key early literacy skills. The ES estimates for individual subtests on the Bader ranged from .48 (Segmentation) to .71 (Phonemic blending) and would be considered medium to large effects. Effect sizes on the three of eight Brigance subtests were below the Effect Size benchmark: Survival Site words (0.24), Recites Alphabet (0.22), and Expressive Vocabulary (0.16, not significant).

The effect size estimates for each statistically significant literacy subtest (9 out of 11), as measured by the Brigance and Bader instruments, are presented below in **Figure 7**. The results are organized according to the subtests' respective literacy constructs: decoding, phonological awareness, letter knowledge, and pre-literacy discrimination. Please refer to the **Outcome Measures** section beginning on page 17 for a discussion of the measurement constructs and **Table 3** for a list of all 11 subtests and their corresponding constructs.

Figure 7. Effect Size Estimates by Literacy Construct

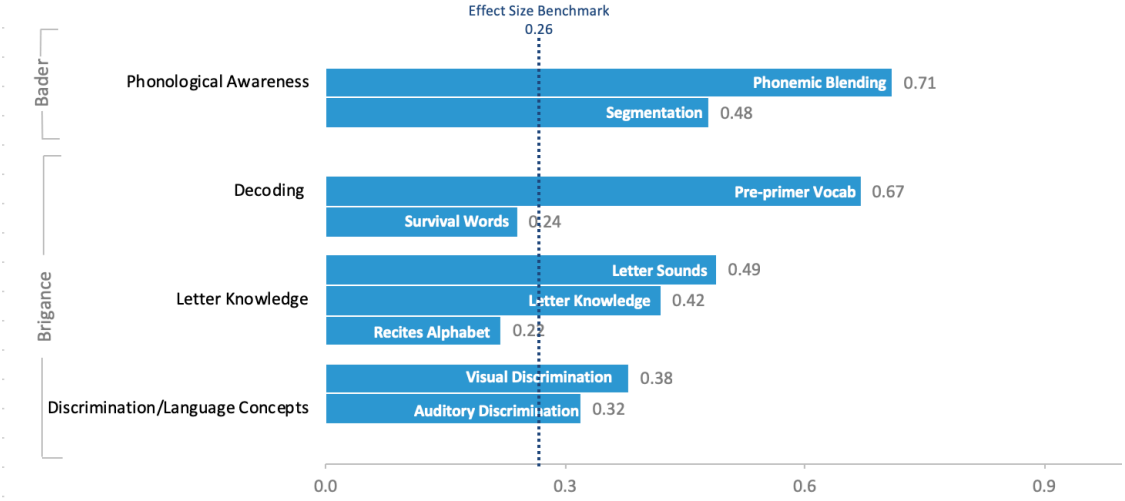
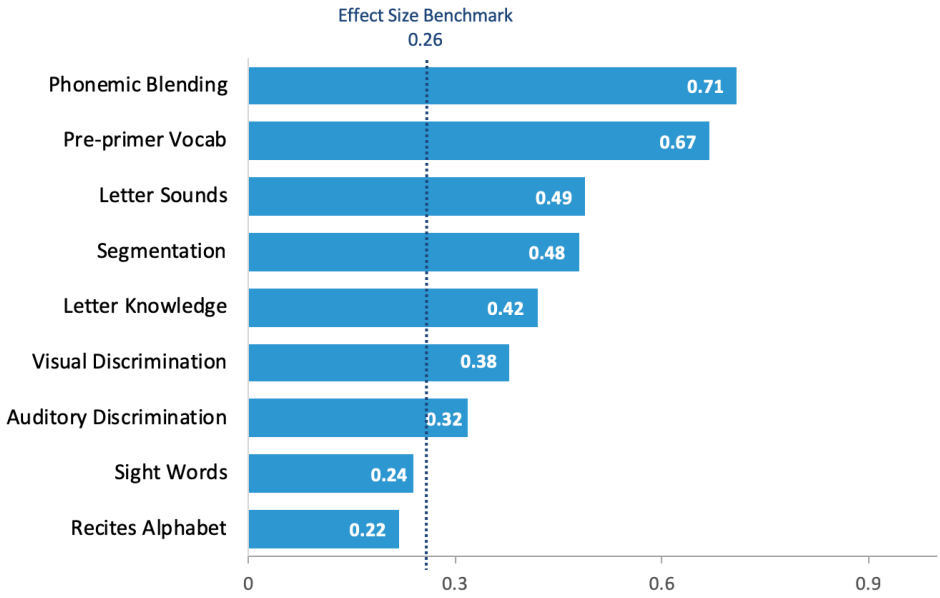


Figure 8 presents the ES of each literacy subtest by the size of their effects along with the .26 effect size benchmark from similar education intervention studies. UPSTART had the largest impact on phonemic blending (.71), pre-primer vocabulary (.67), and letter sounds (.49). Effect sizes from five of six subtests measuring decoding, phonological awareness, and visual/auditory discrimination were above the average .26 effect size benchmark from other similar education interventions and should be considered practically significant and consequential.

Figure 8. 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 and pre-test measures affected the results.

Using Brigance pre-test scores as covariates improved the estimate of UPSTART's overall impact to 22.70. The linear combination of UPSTART participation and the Brigance composite pre-test was significantly related to performance on the Brigance post-test, R^2 = .41, adjusted R^2 = .41, F = 168.92, p < .0001, and accounted for 41% of the explained variability in posttest outcomes.

Research Question 1.2: *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 sample t-tests were performed to examine growth rates as measured by the Brigance and the Bader total test composite scores for the treatment and control group children and each subtest (Phonemic Blending, Phonemic Segmenting, Visual Discrimination, Recites Alphabet, Letter Knowledge, Letter Sounds, Auditory Discrimination, Survival Sight Words, and Pre-Primer Vocabulary). 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 < .001) 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 4 points higher on the Bader and 23 points higher on the Brigance.

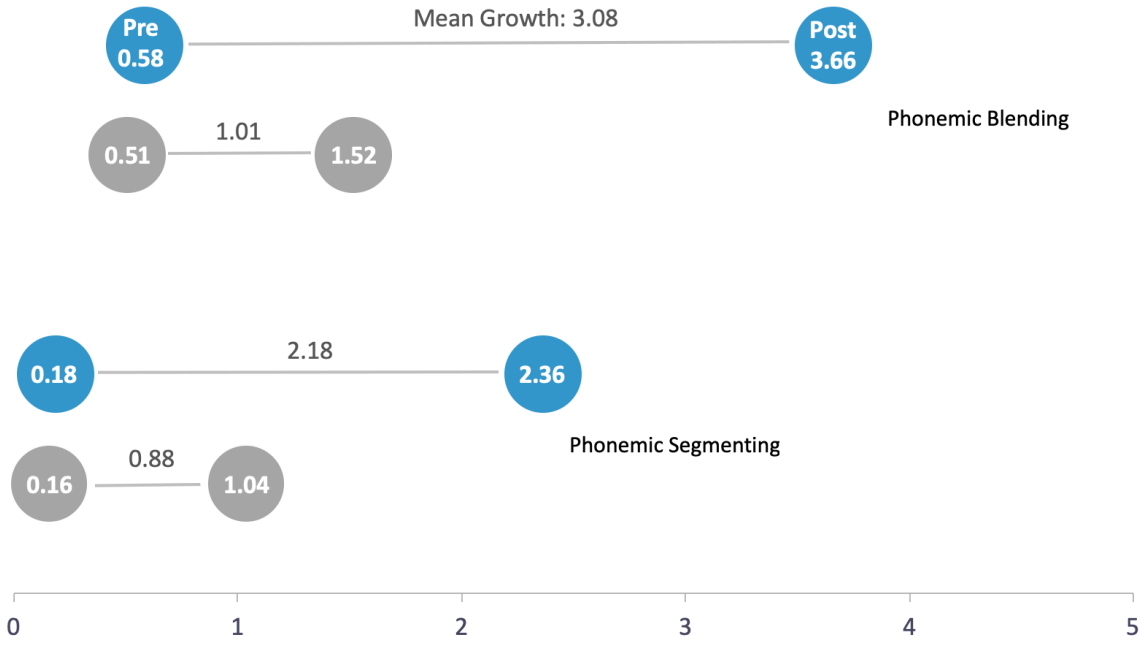
- The treatment group showed statistically stronger ($p < .01$) literacy growth rates compared to the control group on eight of the Brigance subtests (Expressive vocabulary, Visual Discrimination, recites alphabet, Letter Knowledge, Letter Sounds, Auditory Discrimination, Survival Sight Words, and Basic Pre-Primer Vocabulary) and two of three Bader subtests (Phonemic Blending and Segmentation).

- There was no statistically significant difference in mean growth rates between the treatment and control group on the Rhyming subtest.

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**, **letter knowledge, and pre-literacy discrimination**. UPSTART participants' scores are depicted in blue, while their control group counterparts are in grey.

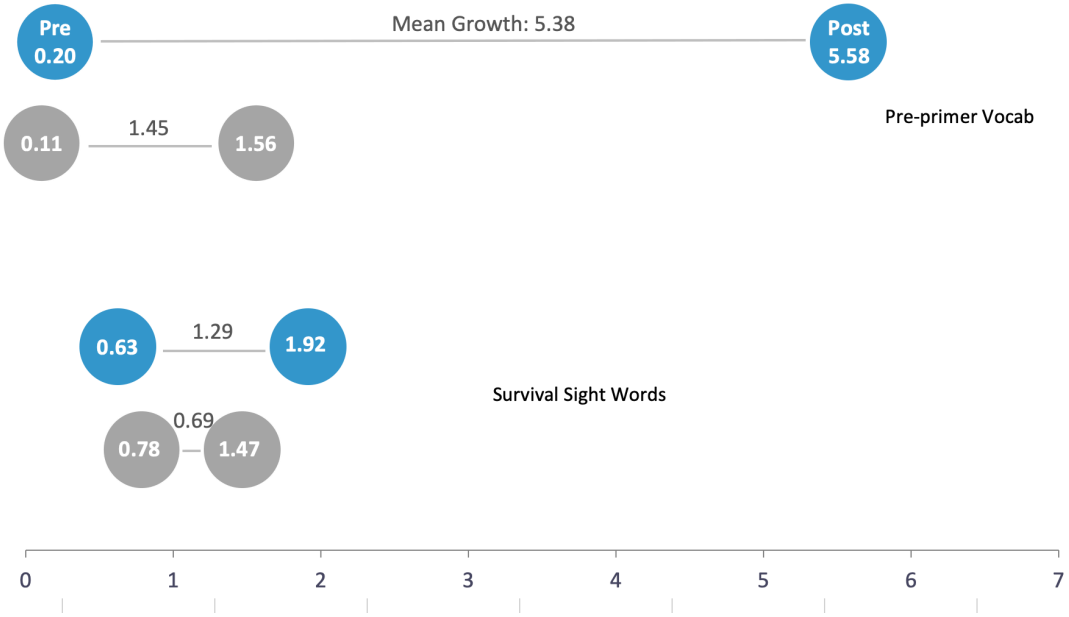
UPSTART children experienced significant, higher mean growth from pre-test to post-test compared to control children on all three subtests (phonemic blending and segmenting) that measure **Phonological Awareness**.

Figure 9. Phonological Awareness: Treatment and Control Group Pre-and-Posttest Mean Scores



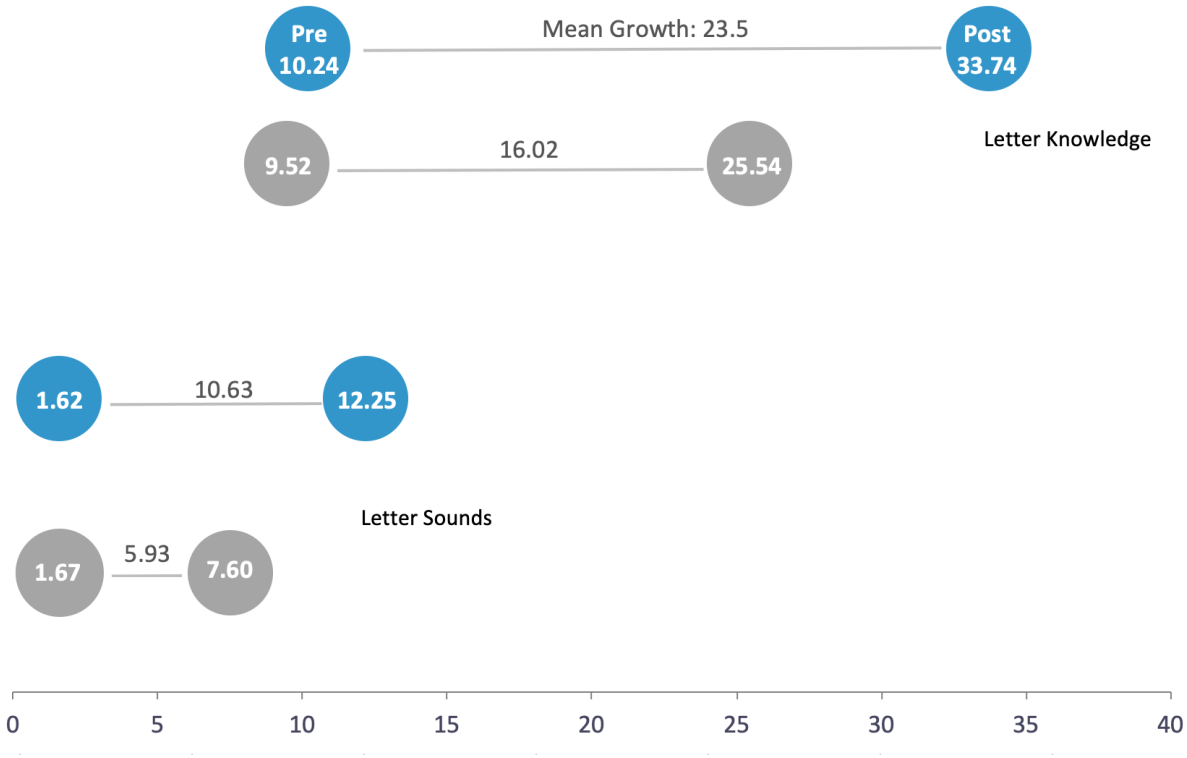
UPSTART students experienced significant, higher mean growth compared to the control group on both subtests used to measure children's **Decoding** ability, including pre-primer vocabulary and survival sight words.

Figure 10. Decoding: Treatment and Control Group Pre-and-Posttest Mean Scores



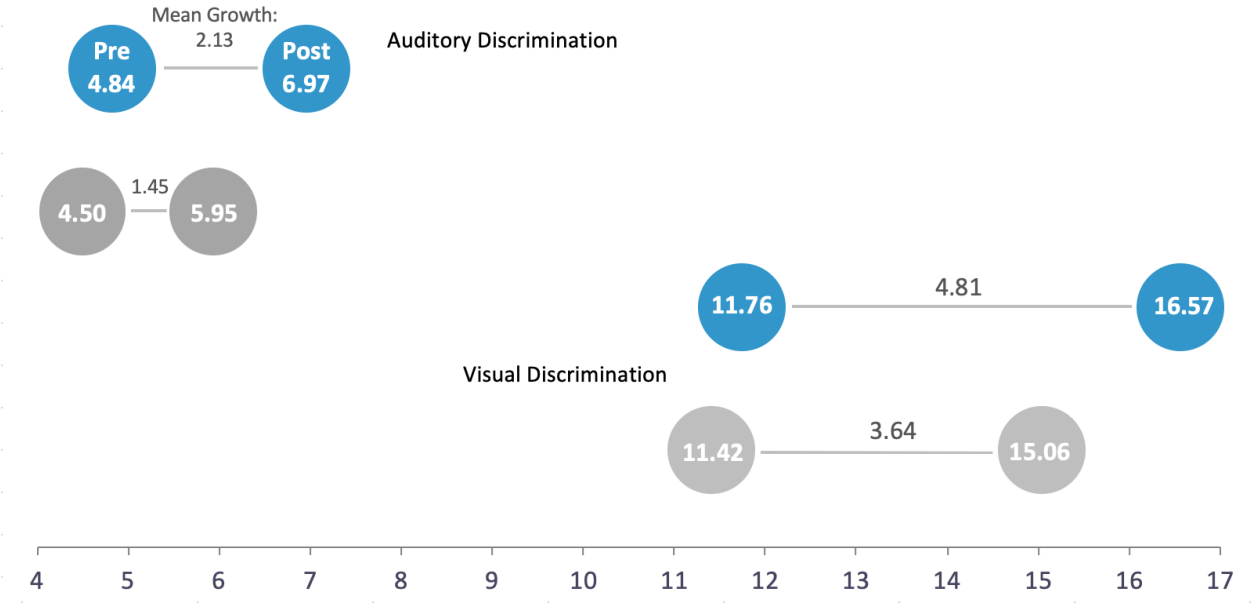
UPSTART children experienced significantly higher growth, compared to non-UPSTART children, in measuring **Letter Knowledge**. UPSTART children showed stronger growth in producing sounds of lower-case letters (letter sounds). A significant difference in the growth rates of treatment and control students was not observed for the Identifying Uppercase Letters or Reciting the Alphabet subtests.

Figure 11. Letter Knowledge: Treatment and Control Group Pre-and-Posttest Mean Scores



Students who were enrolled in UPSTART had significantly higher levels of growth on one subscale measuring **Pre-Literacy Discrimination**, compared to students who did not participate in UPSTART. UPSTART children were more likely to improve on this subtest which involved visually identifying the similarities and differences between forms, letters and words. A significant difference in the growth rates between the two groups was not observed for the auditory discrimination subtest, where children identify similarities and differences between word sounds.

Figure 12. Pre-literacy Discrimination: Treatment and Control Group Pre-and-Posttest Mean Scores



Summary and Discussion

This section of the Cohort 9 (C9) evaluation report summarizes findings and trends for UPSTART implementation and impacts on early literacy skills.

Program Implementation

Based on the program enrollment demographic and usage data provided by UPSTART program officers at the Waterford Institute, the program was implemented with great success. UPSTART enrollment increased from 10,745 to 14,278 children in Year 9, an increase of 33 percent over the past year. Enrollment increased in areas across the state of Utah and UPSTART has reached families in both rural and urban communities. Forty-two percent of the children enrolled in Year 9 lived in families with incomes less than 200% of the federal poverty level and the majority of children were White (82%) and English speaking (92%).

Most of the C9 children accessed the UPSTART curriculum through the Waterford website (83%). Approximately 10% of the ninth-year participants received a computer loan and 5% were provided with a computer and Internet. Despite increased enrollment across the state, graduation rates at 89% were consistent with the previous year, but slightly lower than the 92%-94% graduation rate that characterized earlier cohorts. Families with children who did not graduate from UPSTART tended to have lower levels of parental education, higher levels of poverty, and be members of underrepresented racial, ethnic, and linguistic groups.

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 statistical effects (Bader ES = .56; Brigance ES = .53) on learning compared to non-program children that are stronger, on average, than other educational evaluation studies on similar interventions with comparable outcomes and participants. 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 small to large effects (effect sizes that surpassed our .26 threshold ranged from .26 to .71). 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 phonemic blending (measures phonological awareness), pre-primer vocabulary (measures decoding skills) and letter sounds (measures letter knowledge).

First Grade Analysis

Evaluations of the UPSTART program have consistently shown a medium to strong impact on improving children's early literacy skills prior to entering kindergarten. For example, as reported in our recent evaluation, students enrolled in Cohort 8 during the 2016-17 academic year experienced significant positive effects (ES = .50) compared to control children on the Brigance composite, an instrument that measures decoding skills, letter knowledge, vocabulary and syntax, and pre-literacy discrimination (Evaluation and Training Institute, 2018). Other evaluations of preschool programs conducted after program completion show similar evidence of increased skills in both early literacy and mathematics (Weiland & Yoshikawa, 2013), suggesting that high-quality preschool programs can foster school readiness and prepare children to meet the demands of kindergarten.

Looking at the long-term impact of preschool participation, while some research points to continuing benefits of high-quality preschool experiences on cognitive outcomes into adolescence (Vandell, Belsky, Burchinal, Vandergrift, & Steinberg, 2010), other researchers have found evidence of a "preschool fadeout" (Smith et al., 2016), with the benefits of preschool diminishing in elementary school, and in some cases as soon as by kindergarten or first grade (Puma, Bell, Cook, & Heid, 2010). A variety of factors may be involved in the convergence of preschool attendees' and non-attendees' test scores, including as schooling that fails to build on the gains created by early childhood education or teachers who focus their attention on catching non-attendees up to the level of their preschool attendee counterparts (Yoshikawa et al., 2013)

As part of the UPSTART program expansion, stakeholders were interested in the long-term impact of UPSTART on students and whether program benefits present upon entry to kindergarten can be sustained once children begin elementary school. The First Grade Analysis examines whether the achievement gains from UPSTART that occurred prior to school entry were sustained through kindergarten and first grade.

Kindergarten EISP Exposure

Education initiatives such as the UPSTART program do not operate in isolation, and there are often multiple efforts or programs to foster student achievement in young learners. During the 2017-2018 school year, statewide legislation through the Early Intervention Software Program (EISP) provided funding to districts to supplement kindergarten students' classroom learning with computer-based adaptive reading software programs. The goal of EISP is to provide additional individualized instruction for students in order to increase the number of students reading at grade level and to ensure students are meeting literacy achievement benchmarks. Schools interested in participating in the program submitted applications to the USBE and selected their reading software of choice from among seven vendors. Software vendors provided training and support to schools throughout the year and their programs were used in 403 schools and by 23,090 kindergarten students in 2017-18.

Consequently, it is possible that a student who was enrolled UPSTART preschool program in 2016-17, matriculated into a kindergarten classroom that was also participating in the EISP program during the 2017-18 school year. Participating in the EISP program would be major confound for the purposes of our study – both UPSTART and EISP software programs are computer-based, adaptive, and provide individualized instruction on a consistent and prescribed basis in early literacy. A student who did not participate in UPSTART but who was enrolled in a school receiving EISP services might outperform students who did not participate in either program. Additionally, because both the UPSTART preschool and EISP program involve the use of computer-based early literacy software, it is important to determine the unique impact of UPSTART preschool from participation in EISP kindergarten instruction and the possibility of potential multiple effects from participating in both programs. As the evaluators for both the UPSTART and the EISP programs, we are in a unique position to be able to determine which program (if any) a student participated in and create independent and mutually exclusive groups to ascertain the distinct impact of UPSTART on children's literacy outcomes, and the impact of the combination of UPSTART and EISP.

Research Questions

The research questions used to guide the direction of our first-grade analysis are as follows:

> **Research Question 3.1:** *Does the use of a home-based, computer-supported literacy skills training program in preschool result in stronger school-based literacy outcomes at the beginning of first grade compared to a group of peers matched in terms of demographic characteristics who did not receive the preschool program?*

We hypothesized that if UPSTART has no effect on sustaining students' literacy skills through the first grade, then the children who participated in UPSTART (the treatment group) would perform at the same level as a comparison control group (children who were not exposed to UPSTART or EISP) on measures of literacy development at the beginning of first grade. If UPSTART does have a continued impact on students' literacy achievement, then the treatment group should perform significantly better than the control group on literacy measures at the beginning of first grade.

Additionally, in light of calls for investigation of aligned preschool-elementary school curricular approaches in sustaining preschool benefits (Jenkins et al., 2016), we conducted an explorative analysis of the impact of participating in both the UPSTART and EISP programs. Would participation in UPSTART during the preschool year, coupled with participation in EISP during the kindergarten year, lead to stronger literacy outcomes compared to students who did not participate in either program? Our second research question addresses this line of inquiry:

> **Research Question 3.2:** *Does the use of a home-based, computer-supported literacy skills training program in preschool* ***along with a computer-based kindergarten program*** *result in stronger school-based literacy outcomes at the beginning of first grade compared to a group of peers matched in terms of demographic characteristics who did not receive the preschool or kindergarten program?*

If UPSTART and participation in the EISP program has a continued impact on students' literacy achievement, then we would expect children who were enrolled in UPSTART preschool and participated in EISP to have significantly stronger performance on first grade literacy measures when compared to comparison students who did not participate in either program.

Methods

This section describes the research methods used to answer our research questions, including the research design, outcome measures, data sources, and procedures utilized to create the analytic sample.

Research Design

Due to the fact that we do not have pre-program data for the complete sample of participating students, we elected to implement the first-grade evaluation of the UPSTART preschool program as a nonequivalent groups post-program only design. The evaluation design is diagrammed below in **Table 13**.

Treatment children participated in UPSTART during the eighth year of implementation (Cohort 8) the 2016-17 preschool year. While the control group remains constant (children with no UPSTART exposure or participation in EISP), the treatment group varies based on our specific analytic goals. When answering **Research Question 3.1** and exploring the unique impact of UPSTART on children's first grade literacy achievement, the treatment group consists of students who only used UPSTART. The UPSTART + EISP group is used as the treatment group to answer **Research Question 3.2** and investigate the combined effects of enrolling in UPSTART preschool program and participating in the EISP program.

Table 13
First Grade Analysis Evaluation Design

		Preschool 2016-17	**Kindergarten 2017-18**	**First Grade 2018-19**
Treatment	**UPSTART only**	UPSTART	No Program	
	UPSTART + EISP	UPSTART	EISP Program	
Control	**Control (no program use)**	No Program	No Program	
Measure				Post-Test Only DIBELS BOY 1st Grade

Because the first grade analysis necessitates a quasi-experimental design in which the treatment and control groups are not completely equivalent on factors that may influence reading achievement outcomes, we utilized statistical match techniques (CEM) to equate the two groups and minimize the presence of preexisting differences. We matched treatment and control groups on the demographic variables of ethnicity, language, low income status, Title 1 enrollment, and English Learner and special education status. We did not, however, equate the groups on the basis of Beginning of the Year (BOY) Kindergarten DIBELS scores. It has been demonstrated that UPSTART students enter the school setting with higher literacy scores than comparison students and negating that effect through statistical controls would not be an accurate representation of the short-term impact of UPSTART.

Measures

Our outcome measure consisted of the DIBELS, a standardized measure of literacy achievement for elementary school students. The DIBELS is administered to students in Grades K-3 in schools throughout the state. At the beginning of the year of kindergarten (BOY), the DIBELS measures children's competency with the alphabetic principle and basic phonics with the Letter Naming Fluency and First Sound Fluency subtests. The subtests administered at the second half of kindergarten (middle of year - MOY and end of year - EOY) and beginning of first grade (BOY) assess children's letter knowledge, phonics and word attack skills with the following measures: Letter Naming Fluency, Phoneme Segmenting Fluency, and Nonsense Word Fluency (see **Table 14**).

Table 14
DIBELS Next Subscales by Administration Period

	Kindergarten BOY	Kindergarten MOY	Kindergarten EOY	First Grade BOY
First Sound Fluency	X	X		
Letter Naming Fluency	X	X	X	X
Phoneme Segmentation Fluency		X	X	X
Nonsense Word Fluency		X	X	X

The DIBELS Composite score is an overall measure of children's early literacy ability and is calculated by summing the subtest scores associated with each test administration period. The DIBELS First Grade Composite score serves as our outcome measure.

Data Sources

We relied on data from four different sources to create our final dataset and complete our analyses, including demographic data, literacy achievement scores, UPSTART usage information, and participation in the EISP educational software program.

- The USBE provided demographic data for students enrolled in first grade during the 2018-19 academic year. The demographic data consisted of student-level information such as gender, race, socioeconomic status, English language learner status, primary language, and Title 1 school status.
- DIBELS Next data was provided by the USBE for grades and years under study.
- Student-level data detailing UPSTART preschool software usage for children enrolled in Cohort 8 during 2016-17 was provided by the Waterford Institute. All

students who were enrolled in UPSTART were included in our analysis, regardless of the amount of time they used the program.

- We used archival data from the EISP evaluation to identify and flag kindergarteners who participated in the EISP program during the 2017-18 program year. All students who participated in the EISP program were included in our analysis, irrespective of use.

Merged Data File

We removed instances of duplicate cases and records with missing SSIDs, baseline scores (DIBELS Kindergarten BOY) or outcome scores (DIBELS First Grade BOY) and systematically merged the data files together, using state provided identification numbers (SSIDs). Cases may have failed to merge due to students skipping or repeating grades, having incorrect SSIDs entered into the data file, or leaving the public school system (e.g., moving out of state, enrolling in home school). The complete merged data file consisted of a total of 38,234 cases, broken out into the following independent groups:

Table 15
Group Sizes for Unmatched First Grade Analysis File

Group	N=
UPSTART only	3,503
UPSTART + EISP	3,310
EISP only	15,271
Control	16,150
Total	**38,234**

One of the shortcomings of post-test only designs is selection bias, or that it is difficult to determine if any observed post-test differences between the treatment and treatment group are due to preexisting differences. In an effort to address this issue, we utilized CEM to create balanced matched samples to statistically control for significant differences between our treatment and control groups. (For a detailed discussion of CEM, please see the **Cohort 9 Evaluation**.) Our final analytic samples consisted of two data files: (1) one data file containing UPSTART only students (N = 3,503) and a matched comparison sample (N = 3,503) of students did not have UPSTART or EISP program experience and (2) a second data file containing UPSTART plus EISP students (N = 3,307) and a matched comparison sample (N = 3,307) of students who did participate in either program.

Table 16 presents key demographic characteristics for the matched analytic sample of students who only participated in the UPSTART preschool program (UPSTART Only) and their matched comparison students.

Table 16
UPSTART Only and Control Student Comparisons on Key Demographics

Demographic Categories		Treatment (N=3,503)	Control (N=3,503)
Child Gender	Female	51%	51%
	Male	49%	49%
Child Race	White	84%	84%
	Hispanic	9%	9%
Child Language	English Language Learner	5%	5%
Title 1 School	Yes	24%	24%
	Targeted for Individual Students	12%	12%
Household Income	Low Income	25%	25%

The demographic characteristics of students who participated in UPSTART as preschoolers and were enrolled in a kindergarten classroom that received EISP program software (UPSTART + EISP) and their similarly matched comparison students are displayed in **Table 17**.

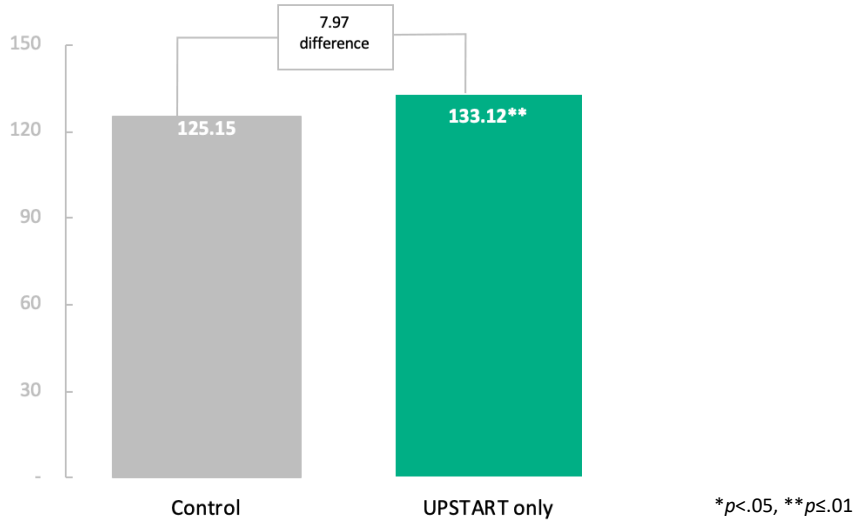
Table 17
UPSTART Only + EISP Program – Control Comparisons on Key Demographics

Demographic Categories		Treatment (N=3,307)	Control (N=3,307)
Child Gender	Female	51%	49%
	Male	49%	51%
Child Race	White	81%	81%
	Hispanic	12%	12%
Child Language	English Language Learner	6%	6%
Title 1 School	Yes	26%	26%
	Targeted for Individual Students	14%	14%
Household Income	Low Income	30%	30%

Findings

Our first set of analyses looks at the impact of enrolling only in the UPSTART preschool program on first grade literacy outcomes. When compared to a group of comparison students matched on demographic characteristics, we find evidence that first grade beginning of year (BOY) DIBELS scores are statistically significantly higher for children who were enrolled in the UPSTART preschool program. Specifically, as seen in **Figure 15**, UPSTART students had an average BOY first grade DIBELS composite score of 133.12 compared to the average score of 125.15 for control students, a 7.97-point difference.

Figure 15. First Grade DIBELS Composite Scores UPSTART and Control students

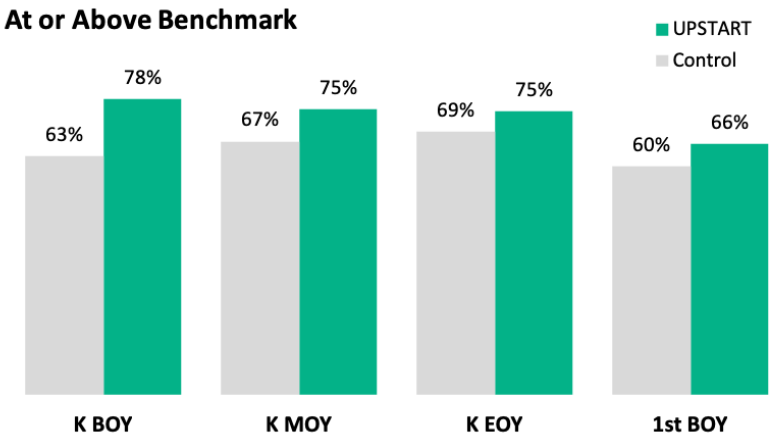


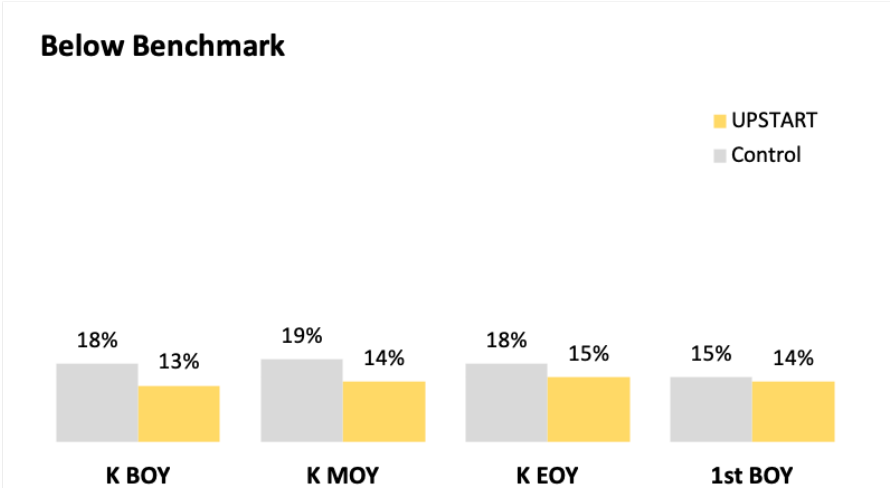
The difference between UPSTART participants and their non-program comparison counterparts on the first grade BOY DIBELS composite produced an effect size of .17, which is less than the .26 effect size benchmark for similar interventions and evaluation studies. (For a more detailed discussion of effect size, please see the **Cohort 9 Evaluation**). An analysis of DIBELS composite scores at testing periods at the beginning, middle, and end of kindergarten and at the beginning of first grade using independent t-tests indicate that UPSTART children performed significantly higher on the DIBELS composite throughout kindergarten and at the beginning of first grade when compared to a group of control children who did not participate in UPSTART.

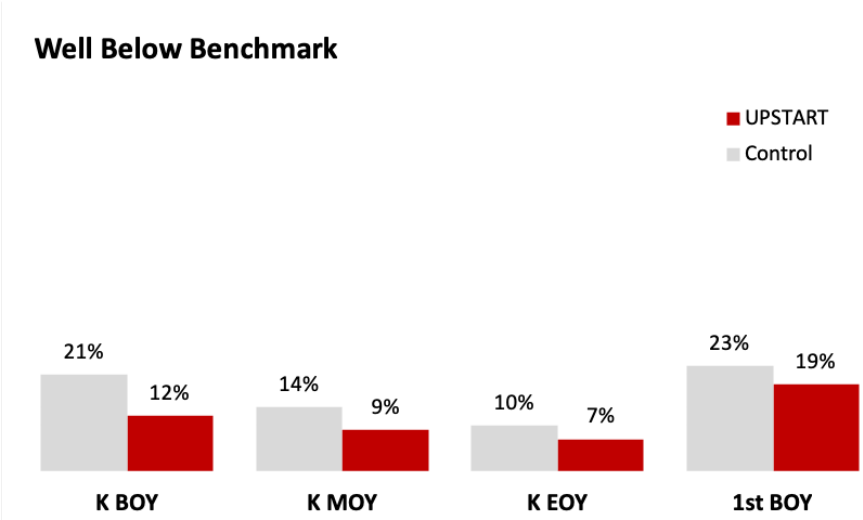
The bar graphs in **Figure 16** show the performance of children who participated in the UPSTART program with children who were not UPSTART participants on the DIBELS composite benchmark classifications that are measured at multiple time points in kindergarten and the beginning of first grade. DIBELS benchmarks are empirically derived cut points that indicate adequate reading skill for a particular grade and time of year and are categorized as at or above benchmark, below benchmark, and well below benchmark. Children who received instruction from UPSTART outperformed similar comparison students throughout kindergarten and into first grade. As seen in the **Figure 16** bar graphs, UPSTART children were more likely to be classified as at or above benchmark at each assessment period than comparison students who did not participate in UPSTART and were less likely to be classified as below or well below literacy

benchmarks. Interestingly, both UPSTART and comparison students had lower levels of literacy achievement at the beginning of first grade (66% of UPSTART children and 60% of comparison children categorized at or above benchmark) compared to the end of kindergarten (75% of UPSTART children and 69% of comparison children categorized at or above benchmark).

Figure 16. Literacy Benchmarks Over Time: UPSTART only and Comparison Students

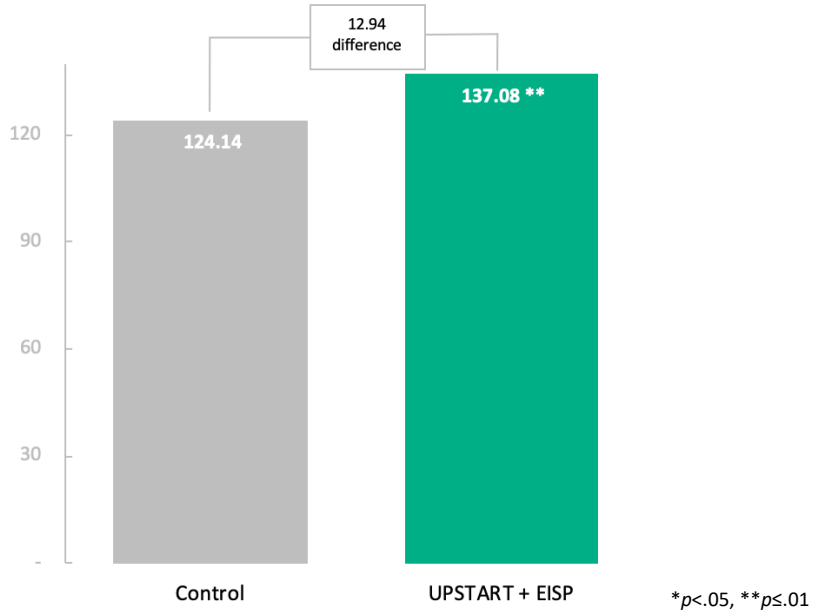






The second set of analyses takes into account the presence of the statewide EISP software program initiative and evaluates the impact of participating in UPSTART and a similar adaptive computer-based program that provides individualized literacy instruction throughout kindergarten. We found that students who participated in UPSTART during preschool and EISP during the kindergarten academic year had statistically significantly higher scores on the first grade DIBELS composite than students who did not participate in either program. As seen in **Figure 17**, mean scores on the first grade DIBELS composite were 137.08 for the UPSTART + EISP treatment group and 124.14 for students who did not receive any literacy software, a 12.94 difference. This difference produced an effect size of .28, which is above the .26 effect size benchmark for similar studies reported in the literature and should be considered a practically significant result.

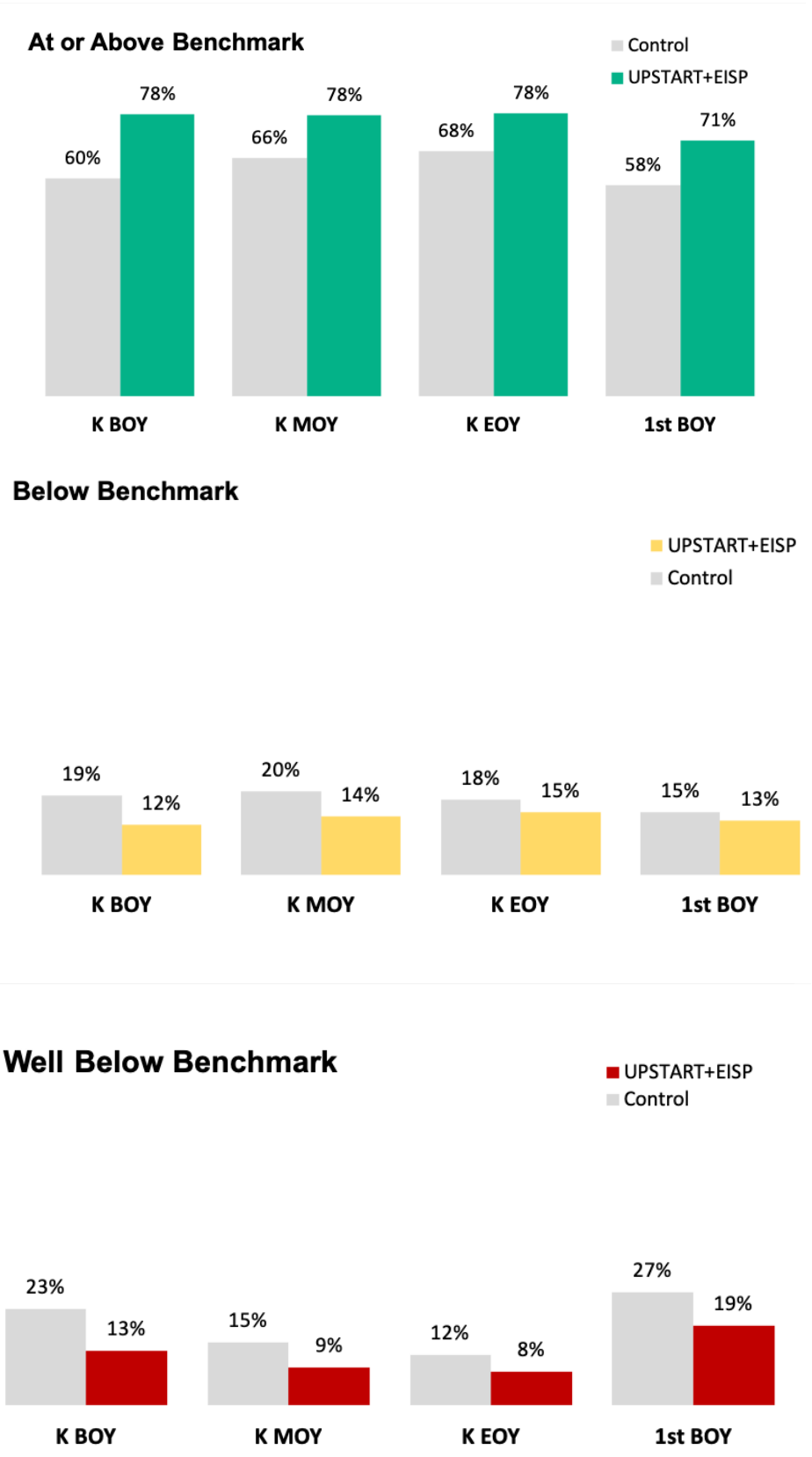
Figure 17. First Grade DIBELS Composite Scores UPSTART + EISP and Control students



The bar graphs in **Figure 18** show the overall performance of children who participated in the UPSTART and EISP programs with children who participated in neither program on the DIBELS composite benchmark classifications measured throughout kindergarten and at the beginning of first grade. Similar to the results in the UPSTART only analysis, children who received instruction from UPSTART and EISP programs outperformed comparison students throughout kindergarten and into first grade. Children who participated in UPSTART and EISP were more likely to be classified as at or above literacy benchmarks at each assessment period, and less likely to be classified as below or well below benchmarks (see **Figure 18**).

There is also an analogous pattern to the UPSTART only analysis of lower levels of literacy achievement at the beginning of first grade, with 71% of UPSTART + EISP children and 58% of comparison children categorized at or above benchmark in first grade, compared to 78% of UPSTART children and 68% of comparison children categorized at or above benchmark at the end of kindergarten.

Figure 18.
Literacy Benchmarks Over Time: UPSTART/EISP and Comparison Students



Summary and Discussion

Our first grade analysis moved beyond evaluating the immediate impact of the UPSTART preschool program on preparing children for entry into traditional school environments to assess whether or not UPSTART has a sustained benefit on children's literacy achievement once children are in elementary school. Specifically, we followed Cohort 8 students through kindergarten and first grade and utilized a post-test only design to determine if UPSTART participants had higher scores on the first grade DIBELS assessment compared to students who were not enrolled in UPSTART. In an effort to isolate the effects of participating in the EISP program, a statewide computer-based literacy instruction software program for grades K-3, we excluded any student who participated in EISP as a kindergartener from our control group. We also created two treatment groups to examine potential multiplier effects from participating in both programs: students who only participated in UPSTART during their preschool year (UPSTART only) and students who participated in UPSTART as preschoolers and who participated in the EISP program as kindergarteners (UPSTART + EISP).

We found significant small effects for the sustained benefit of UPSTART that were consistent with previous cohort findings. UPSTART has a positive impact on students without additional curricular support (the effect size of the UPSTART only group was .17) and an even larger impact on students who receive further individualized computer-based instruction (the effect size of the UPSTART + EISP group was .28). The effect of participating in both UPSTART and the EISP program was larger than the average effect size reported in similar evaluations with comparable interventions, measures, and students.

Because we used all students who participated in the UPSTART or EISP programs, regardless of the amount students actually used the programs, our treatment samples are considered "intent to treat" (ITT) samples. ITT samples represent the most conservative estimate of the long-term impact of UPSTART because it includes students who met vendors' requirements for program use as well as students who may have only used the program sporadically or not at all (Montori & Guyatt, 2001). However, other researchers argue that if a participant is included in the treatment group, but did not actually receive treatment, it indicates little about the treatment's efficacy (Gupta, 2011). To that end, we recommend that future analysis of the long-term effects of UPSTART include a subsample of UPSTART users who fulfilled program requirements for usage.

Summary and Recommendations

The UPSTART program shows continued success at helping preschool age children develop literacy skills and prepare for entry into kindergarten. There is also evidence that UPSTART program students' literacy achievement is sustained throughout kindergarten and into first grade. Given the success at improving literacy test scores, we recommend that the state continue providing the UPSTART program to children.

During the 2017-2018 program year, a consistent group of C9 students were classified as graduates when compared to the previous cohorts (89% graduation rate in C8 and C9) even in the face of a 33% enrollment increase. It is important to continually monitor program usage as previous reports indicated that children who failed to meet the program requirements for graduation had, on average, significantly lower literacy outcome scores when compared to UPSTART graduates (Evaluation and Training Institute, 2018). Moreover, Cohort 9 families that did not meet usage requirements were more likely to have other indicators of risk, such as lower levels of parental education, lower household incomes, and being non-native English speakers. Graduation rates need to be carefully monitored because a significant decline might erode literacy outcomes for the most at-risk students.

Program Recommendations. Although the graduation rates for C9 students were the same as the previous year, as UPSTART continues its expansion it is important to continually monitor program implementation 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. Specifically, we recommend that the program vendor consider the following recommendations:

- The program vendor could develop new strategies for addressing falling usage and graduation rates among the most at-risk students (i.e. those with high levels of poverty and with English as a second language). Some potential strategies might include:
    - Establishing peer support systems among similar groups to discuss strategies for supporting children's program use.
    - Highlighting evaluation information that links graduation with higher literacy outcomes.
    - Developing targeted incentives for families with the highest risk factors for not meeting program usage requirements, such as monthly awards (extrinsic), being highlighted in UPSTART communications to social networks as "Gold Star Families" (intrinsic).

Results from the first-grade analysis indicate that UPSTART children were able to maintain their advantage in literacy outcomes through the beginning of first grade, and that these effects were greater for UPSTART children who participated in the EISP program. Because the EISP program also provides students with individualized adaptive computer-based literacy instruction, it provides a logical support to build on the gains created by UPSTART during students' preschool year.

Evaluation Method Recommendations & Future Research. 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 that serve low-income families, 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. (2018, February). *Utah UPSTART program evaluation program impacts on early literacy: Year 8 Results* (Cohort 8 Technical Report). Culver City, CA: Author.

Gupta, S. K. (2011). Intention-to-treat concept: A review. *Perspectives in Clinical Research*, *2*(3), 109-112. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3159210/

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

Jenkins, J. M., Watts, T. W., Magnuson, K., Gershoff, E., Clements, D., Sarama, J., Duncan, G. J. (2016). *Do high quality kindergarten and first grade classrooms mitigate preschool fadeout?* Irvine Network on Interventions in Development.

Lipsey, M., Puzio, K., Yun, C., Hebert, M., Steinka-Fry, K., Cole, M., Roberts, M., Anthony, K. and Busick, M. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms*. Washington DC: Institute of Education Sciences.

Lipsey, M., Weiland, C., Yoshikawa, H., Wilson, S., & Hofer, K. (2015). Prekindergarten age cutoff regression-discontinuity design: Methodological issues and implications for application. *Educational Evaluation and Policy Analysis*, *37*, 296-313.

Mistry, R. S., Benner, A. D., Biezanz, J. C., Clark, S. L., & Howes, C. Family and social risk, and parental investments during the early childhood years as predictors of low-income children's school readiness outcomes *Early Childhood Research Quarterly*, *25*, 432-449.

Montori V.M. &, Guyatt G. H. (2001) Intention-to-treat principle. *Canadian Medical Association Journal*, *165*(10), 1339-1341. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC81628/

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.

Phelps, S. (2003). *Phonological Awareness Training in a Preschool Classroom of Typically Developing Children*. Electronic Theses and Dissertations. Paper 772. http://dc.etsu.edu/etd/772

Puma, M., Bell, S., Cook, R., Heid, C. (2010). *Head Start Impact Study. Final Report*. U.S. Department of Health and Human Services, Administration for Children & Families. Washington, DC.

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

Smith, T. M., Cobb, P., Farran, D. C., Cordray, D. S., & Munter, C. (2013). Evaluating math recovery: Assessing the causal impact of a diagnostic tutoring program on student achievement. *American Educational Research Journal*, *50*(2), 1–32.

Snow, C.E., Burns, M., S., & Griffin, P. (1998). *Preventing Reading Difficulties in Young Children*. Washington, DC: National Academy Press.

Vandell, D. L., Belsky, J., Burchinal, M., Vandergrift, N., & Steinberg, L. (2010). Do effects of early child care extend to age 15 years? Results from the NICHD Study of Early Child Care and Youth Development. Child Development, 81(3), 737-756.

Weiland, C., & Yoshikawa, H. (2013). Impacts of a Prekindergarten Program on Children's Mathematics, Language, Literacy, Executive Function, and Emotional Skills. *Child Development*, *84*(6), 2112–2130.

What Works Clearinghouse. (2017). Procedures handbook (version 4.0). Retrieved from https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_procedures_handbook_v4.pdf

Appendix A: Comparison of C9 Evaluation Treatment Samples

The matched and unmatched treatment samples are compared with the C9 population on key demographic characteristics reported by the program vendor in **Table A.1**. Both of the unmatched and matched treatment samples are more homogenous than the full population of preschoolers who were enrolled in Cohort 9, with 80% of both unmatched and matched children, being White and 99% classified as English speakers.[10]

Table A.1
Sample Treatment Comparisons on Key Waterford Demographics

Demographic Categories		C9 Population (N = 14,278)	Unmatched Sample (N=276)	Matched Sample (N=248)
Gender	Male	52%	52%	52%
	Female	48%	48%	48%
Ethnicity	White	82%	80%	80%
	Hispanic	11%	11%	12%
Child Language	English	92%	99%	99%
Parent Education Level	Some College	34%	71%	72%
	Bachelor's Degree	39%	0%	0%
Parent Marital Status	Married	91%	82%	83%
Poverty Status	Under 185%	37%	72%	73%

The C9 population had parents with higher college graduation levels and lower levels of poverty. Whereas 39% of the parents in the overall C9 population have a college degree, the modal level of parent education in the unmatched and matched treatment sample was some college (71% and 72%, respectively). Additionally, 37% of families in the C9 sample were under the 185% federal poverty rate compared to 72% of families in the unmatched sample and 73% of families in the matched sample. As mentioned in the main body of the report, we focused on recruiting low-income families for our treatment sample to reflect the prioritization of these families by the state in the recent legislative extension of the UPSTART program.

The matched treatment 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.

[10] The testing protocol tests all children in English and requires children to understand directions in English and give verbal assent to proceed with testing. Moreover, parents need to have sufficient understanding of English to give informed consent for their participation.

Appendix B: Determining UPSTART Effect Size Benchmark

One way to assess the practical significance of an intervention is to compare its impact with effect sizes from similar evaluation studies – those that use analogous outcome measures, are evaluating a comparable intervention, or are evaluating interventions that target similar groups. Researchers at the Institute of Education Sciences (IES) reviewed 829 effect sizes from 124 education research studies conducted on K-12 students and reported an array of different effect size distributions that can provide insight into what constitutes a large or small effect relative to similar education evaluation studies (Lipsey et. al, 2012). They provide the following benchmarks to be used as normative comparisons:

- *Benchmark by outcome measure*. IES researchers looked at the type outcome measures (i.e., did researchers use a self-developed outcome measure, a general standardized outcome measure like an IQ test, or a subject-specific standardized outcome measure like a reading or math test) by grade level and found that the average effect size for education research studies evaluating elementary students with a standardized subject test (like the Brigance and Bader literacy tests) was .25. Average effect sizes were slightly higher for middle school students, but lower for high school students (.32 and .03, respectively)
- *Benchmark by intervention type*. Another metric for evaluating effect size was based on the type of intervention under investigation. Researchers sorted the interventions of reviewed studies into several broad categories (e.g., a whole school program, a teaching technique, a new instructional format, skill training, or an instructional program). The UPSTART program was closest to an instructional program, or "a relatively complete and comprehensive package for instruction in a content area like a curriculum or a more or less free standing program (e.g., science or math curriculum; reading programs for younger students; broad name brand programs like Reading Recovery; organized multisession tutoring program in a general subject area." (p. 35) The average effect size for research studies that evaluated a comprehensive instructional program such as UPSTART was .13. Larger effect sizes were found for interventions in the instructional component/skill training and teaching techniques and categories (.36 and .35, respectively).
- *Benchmark by intervention target*. A final yardstick to contextualize effect sizes focused on the targeted group of the intervention (e.g., individual students, small group, classroom, whole school, mixed.) that targeted individual students had average effect sizes of .40. Interventions that targeted individual students had the highest observed effect sizes, on average.

To determine a single benchmark, we took an average of the three different benchmarks (i.e., benchmark by outcome measure = .35; benchmark by intervention type = .13; and benchmark by intervention target = .40) and the resulting benchmark value was .26. This benchmark will be used to contextualize the effect sizes presented in this report and to aid the reader in determining the practical significance of the effect of UPSTART.