UPSTART Program Evaluation

Year 10 Program Results

Submitted to the Utah State Board of Education
January 2020

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

(This page intentionally left blank)

Executive Summary

The Waterford Institute recognizes that nearly half of our country's 4-years-olds face socioeconomic barriers that keep them from early education. Utah Preparing Students Today for a Rewarding Tomorrow (UPSTART) is a home-based computer preschool program developed and provided by Waterford to prepare young children for school entry and future academic success. The program offers in-home, early education access to unserved and often lower income families by providing support, technology, and internet where needed. 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 tenth year of implementation (Cohort 10, with students enrolled during the 2018-2019 program year). In partnership with the UPSTART Advisory Committee (UAC), ETI continues to enhance the research design to meet a higher level of accountability for the program and additionally explore the longer-term impact of UPSTART. This report addresses the following two areas:

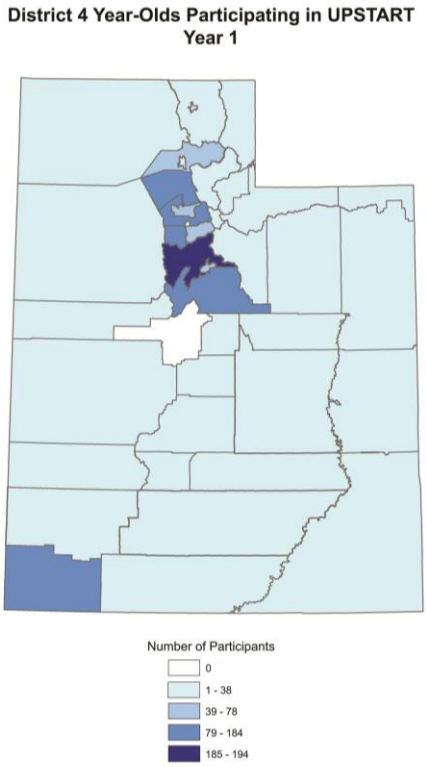
- The **Cohort 10 Evaluation** presents information on program implementation and its impact on developing children's early literacy skills. Our research findings cover two areas: (1) how the program was implemented and (2) what types of impacts the program had on children's literacy.
- The **First Grade 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 occurring prior to school entry were sustained through kindergarten and into the beginning of first grade.

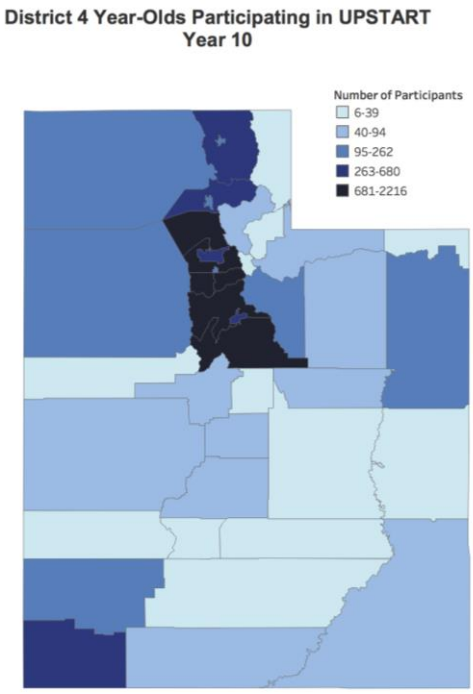
This Executive Summary presents a summary of findings for each reporting area, along with selected recommendations for improving the program and future evaluation efforts.

Cohort 10: Program Implementation

The 2018-19 program year has continued to maintain high numbers of UPSTART enrollment, as the number of preschool students participating in the program in Year 10 (N = 14,125) sustained similar numbers from the previous year (Year 9, N = 14,278). Over the past ten 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 10).

Maps of UPSTART program participation in Year 1 and Year 10 by School District





Thirty-six percent of children enrolled in UPSTART Cohort 10 were from families with incomes less than 185% of the federal poverty level and the majority of enrolled children were White (82%) and English speaking (92%). UPSTART maintained high enrollment numbers in Year 10, with graduation rates relatively consistent with recent years at 87%.

Findings about UPSTART usage in Cohort 10 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 10 used the UPSTART program for an average of 53 hours during the program year. Students who were UPSTART graduates used the program for an average of 57 hours.
- Students in Cohort 10 had an 87% graduation rate, which continued 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%), Cohort 7 (graduation rate of 87%), Cohort 8 (graduation rate of 89%), and Cohort 9 (graduation rate 89%).
- 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.

Cohort 10: UPSTART's Impact on Preschool Literacy

In our analysis of UPSTART impact on early literacy, we examine results in two different ways, (1) the effect size between the two groups at post-test and (2) the different rate of growth from pre to post-tests for each of the two groups. 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. Effect sizes that represent meaningful differences in a given context are referred to as having a practical or substantive significance. 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 the difference between the groups is large enough to be meaningful in this context.

UPSTART had a strong impact on children's emerging literacy skills based on results from both the effect size and growth score analyses. Children enrolled in UPSTART outperformed 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 outperformed their control counterparts on the Bader composite, an instrument that assesses children's phonological awareness (ES = .35).

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 = .68) and phoneme segmenting skills (ES = .71).
- 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 = .59) and on reading survival sight words (ES = .36).

- 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 also improved 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 substantive treatment effects in their learning how to recite (ES = .32), identify (ES = .33), and sound out (ES = .47) 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 improvement on **pre-literacy discrimination and language concepts**.

- UPSTART had a meaningful impact on children’s ability to discriminate between different shapes, letters, and words (ES = .54), but fell under the threshold in their ability to distinguish if two words sound the same (ES = .20).
- Children in UPSTART had stronger growth scores on their auditory discrimination of words when contrasted to children not enrolled in UPSTART.

The UPSTART program had a measurable impact on children’s **vocabulary**:

- Just above our benchmark for substantive effects on learning, UPSTART positively affected receptive vocabulary (ES=.27).
- Children enrolled in UPSTART had significantly higher growth rates on vocabulary subscales when compared to control children.

UPSTART’s Long-Term Impact: First Grade Analysis

In order to determine if 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 to the beginning of 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.

The DIBELS is a standardized measure of literacy achievement for elementary school students and is administered to students in Grades K-3 throughout the state. At the beginning of the year of kindergarten, the DIBELS measures children’s competency with the alphabetic principle and basic phonics. The subtests administered at the second half

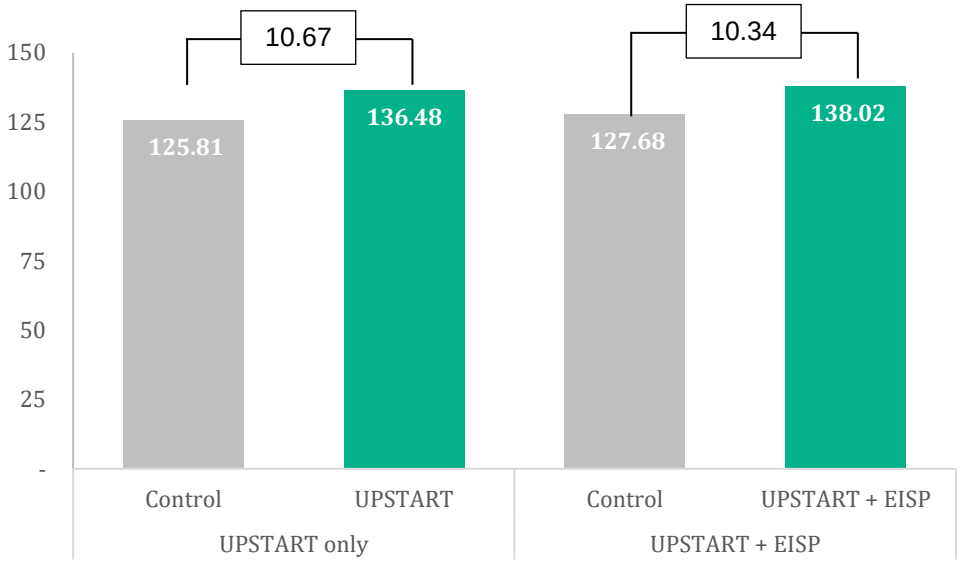
of kindergarten and beginning of first grade assess children's letter knowledge, phonics and word attack skills. The DIBELS Composite score is an overall measure of children's early literacy ability and is calculated by summing the subtest scores associated with each test administration period. The DIBELS First Grade Composite score serves as our outcome measure.

Findings show that, on average, students who used the UPSTART program in preschool scored 10.67 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 .22.

Students who received continuous treatment from preschool through kindergarten (UPSTART plus EISP) did equally as well at the beginning of first grade. Students with combined treatment in preschool and kindergarten scored 10.34 points higher than a group of matched comparison students, a statistically significant mean score advantage over their non-program peers that produced a .22 effect size.

UPSTART students outperformed counterparts in first grade

1st Grade DIBELS Composite Scores by Treatment Group



Top Level Summary

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. The fact that UPSTART children maintained their advantage over their comparison counterparts up to the beginning of first grade is further evidence of another important benefit of the UPSTART program.

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, which are the children that stand to benefit most from the program. Cohort 10 maintained a graduation rate of 87%, slightly less than the previous graduation rate of Year 9 (89%). Program monitoring for use and graduation requirements should be continual to ensure UPSTART is administered with fidelity so that all children can receive the full benefit of the program.

Introduction to the UPSTART Evaluation

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.

We evaluated 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 two 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 10 Evaluation**, we present outcome results for UPSTART's tenth year of implementation, hereafter referred to as Cohort 10 (C10). Additionally, we documented the extent to which participants used the computerized curriculum as it was intended, established the relationship between curriculum usage and literacy outcomes, and reported the program's completion or "graduation" rate. As in our evaluations with recent cohorts, the Cohort 10 evaluation included a statistically balanced match of treatment and control students. While requiring a larger sample size, the matching process enhanced our ability to detect treatment effects and, in general, improved 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, the First Grade Analysis presents findings on UPSTART's continued impact on student literacy through kindergarten and up to the beginning of first grade. This longitudinal study met the provision in state law to evaluate the long-term impacts of UPSTART on students and used DIBELS literacy data collected in schools to determine whether or not UPSTART had 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 tenth year of operation during the 2018-19 school year, the project's implementation contractor – the Waterford Institute – enrolled 14,125 preschool children and provided them with an adaptive program of computer-based early literacy instruction to prepare them academically for kindergarten. The 14,125 children enrolled in the tenth-year cohort, hereafter referred to as Cohort 10 (C10), participated in UPSTART from September 2018 through May 2019.

The UPSTART software used 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[1] include:

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

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

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

Cohort 10 Evaluation

Research Questions

Our evaluation of the Cohort 10 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 C10 evaluation were as follows:

> **Research Question 1.1:** Do UPSTART students have better early literacy skills at entry to 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.

Method Summary

The following section presents a summary of 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. A comprehensive methods section can be found in **Appendix C.**

Research Design and Data Collection

To evaluate the impact of the UPSTART program, data were collected for treatment group children who had enrolled in UPSTART for Year 10 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 using the general literacy and phonemic manipulation instruments a year later, before entering kindergarten. 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.

The final sample consisted of phonemic manipulation data from 474 children, 237 treatment and 237 control, and general literacy data from 450 children, 225 treatment and 225 control, and was based on the subset of children with valid matched pre-test and post-test data, who had not previously used the UPSTART computerized learning program as documented through the pre-screening interview. We used a statistical process called "Coarsened Exact Matching" (CEM) to match control students to treatment students to make two equal or "balanced" groups across a set of important predictor variables.

Outcome Measures

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. The Brigance composite scale of early literacy achievement was created by combining the scores from the eight subtests, which 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 1**. We consider the Brigance composite to be a measure of general literacy, and we refer to it as such for the remainder of this report.

The Bader. The Bader Reading and Language Inventory (Bader, 2008) was selected as a measure of *phonological awareness* and focused on *phoneme manipulation*. Phoneme manipulation involves the child's ability to detect the sound structure of spoken words at three levels: rhyming, syllables, and phonemes. The Bader composite scale is a measure of phonemic manipulation, and we refer to it as such for the remainder of this report.

Table 1
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/Phoneme Manipulation	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

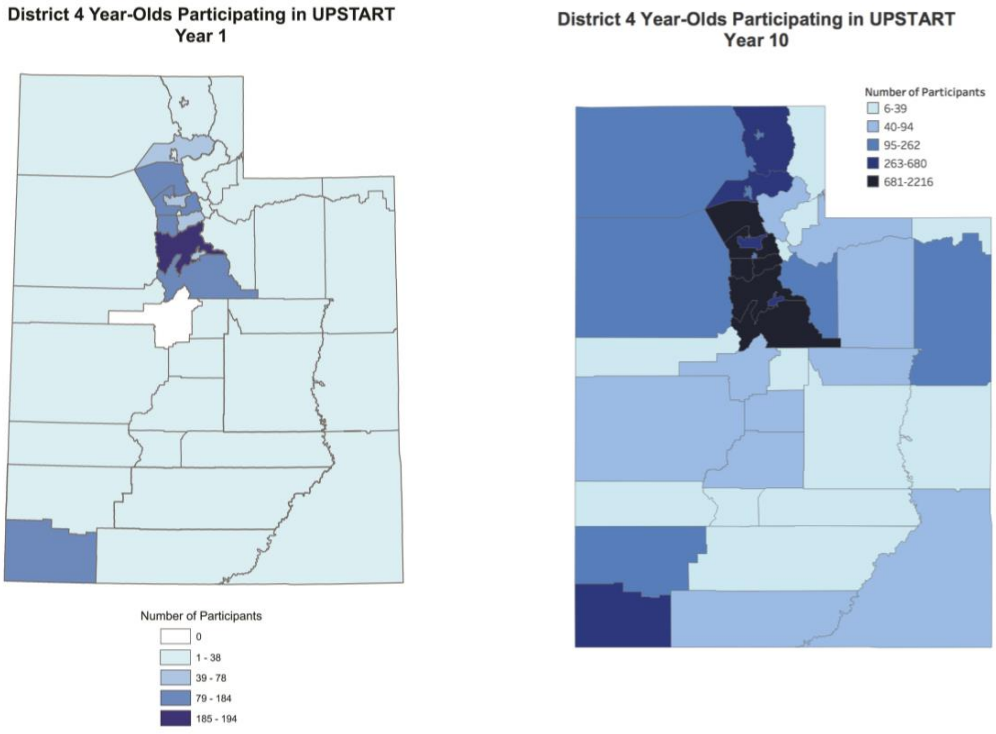
UPSTART Program Implementation

Findings reviewed in the UPSTART implementation section include tenth 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 2018-19 program year was a continued expansion of UPSTART enrollment, as the number of preschool students participating in the program in Year 10 (N = 14,125) slightly decreased in participation from the previous year (Year 9, N=14,278), however maintained relatively high numbers since the inception of the program. The number of students enrolled in the program rose from 1,248 children in Year 1 to 14,125 students in Year 10, 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 10 (N=14,125). As seen below in **Figure 1**, the UPSTART program has continued to further its reach over the past ten 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 10 by School District



The Waterford Institute provided a comprehensive dataset to ETI for the tenth-year UPSTART enrollment of 14,125 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 C10 population are presented below in **Table 2**.

Table 2
Demographic Characteristics of C10 Population

Demographic Categories		All C10 UPSTART (N=14,125)
Child's Gender	Male	51%
	Female	49%
Child's Ethnicity	White	80%
	Hispanic	12%
	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	3%
	High School Graduate	13%
	Some College	33%
	College Graduate	40%
	Advanced Degree	11%
Parent Marital Status	Married	91%
	Otherwise	9%
Household Poverty Level	Below 100%	9%
	Below 185%	36%
	Below 200%	37%

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

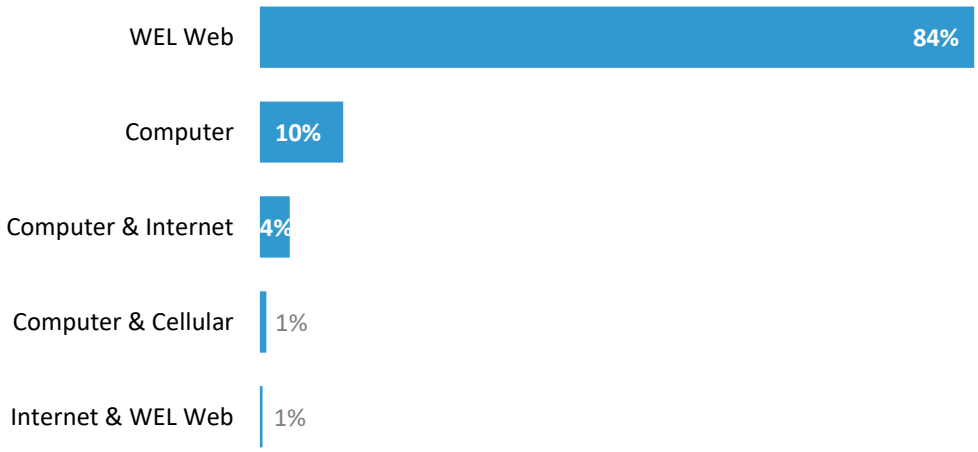
Slightly more C10 boys (51%) than girls (49%) were enrolled and in terms of ethnicity, the majority (80%) of the C10 enrollment was White, with 12% of the children being of Hispanic origin. Thirty six percent of the C10 UPSTART participants lived in families with incomes less than 185% of the federal poverty level.[2]

[2] The federal poverty definition consists of a series of thresholds based on family size. In 2018, a 100% poverty threshold for a family of four was $25,100, while a 185% threshold for a family of four was $46,435(see U.S. Department of Health and Human Services poverty guidelines at https://aspe.hhs.gov/2018-poverty-guidelines).

Provided UPSTART Equipment

The type of education technology provided to UPSTART children in Year 10 of the program is shown in **Figure 2** for all 14,125 children enrolled in the program. The majority of UPSTART children (84%) used the Waterford website to retrieve the UPSTART program. This allowed families to access the UPSTART curriculum from their home computers.

Figure 2. Equipment provided to C10 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 C10 children while they participated in the program. Another 4% of the C10 program participants were provided with internet access and personal computers. The remaining two percent of the C10 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 UPSTART participants who also participated in the ETI evaluation (i.e. UPSTART evaluation sample). The hours of instruction observed for all children documented as enrolled in the tenth year of UPSTART are summarized in **Table 3** and are compared to program "graduates". The average level of usage for all students enrolled in the tenth year of UPSTART (N=14,125) was approximately 53 hours of instruction. The C10 academic year covered 40 weeks of instruction, beginning the week of September 3, 2018 and ending May 27, 2019.

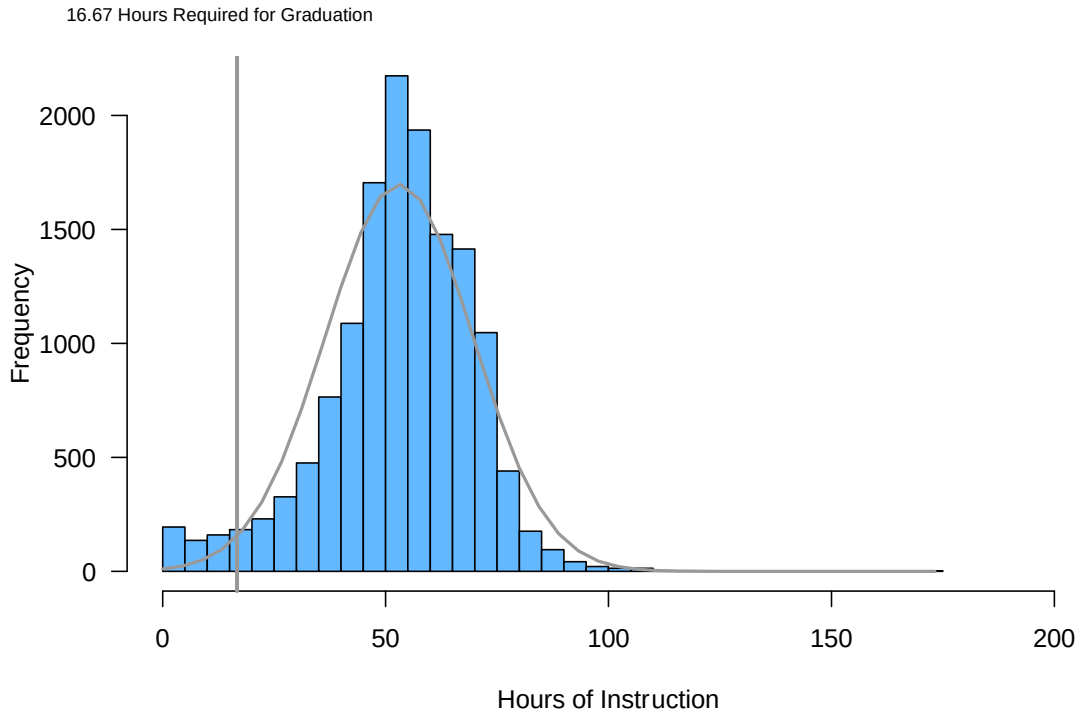
Table 3
C10 Hours of UPSTART Instruction

Group	N	Mean	SD	Range
All UPSTART	14,125	53.01	16.59	00.00 – 173.20
UPSTART Graduates	12,354	57.28	12.25	16.74 – 173.20
UPSTART Analysis Sample	246	53.24	15.74	2.42 – 90.72

Twenty-two of the 14,125 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 10 of the program. For enrolled families whose children did use the curriculum, the average duration in the program was approximately 37 weeks. This usage pattern is similar to that observed in the ninth year of the program (Evaluation and Training Institute, 2018). The children in the C10 evaluation analysis sample used the UPSTART curriculum for approximately 53 hours of instruction on the average. (**see Figure 3)**

The histogram in **Figure 3** shows the distribution of hours of instruction for the total C10 population (N=14,125). Twenty-two of the enrolled children logged zero hours of instruction during their time in UPSTART. At the other end of the spectrum, forty-two children logged over 100 hours of instruction.

Figure 3. Hours of Instruction for C10 Families



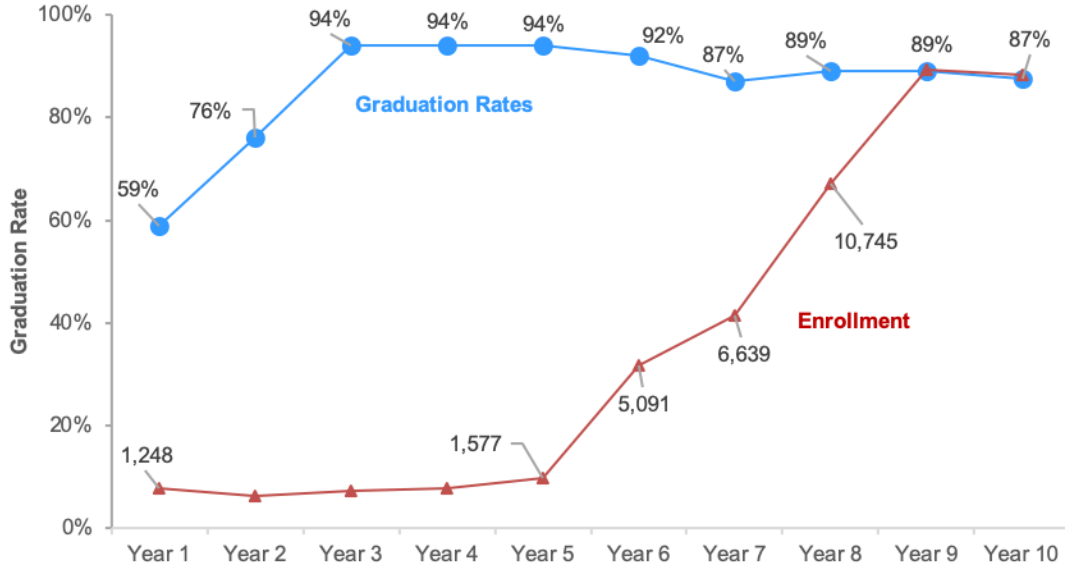
The bottom quartile of the C10 population completed 44.94 hours of instruction or less, the midpoint of the C10 distribution was 54.20 hours, and the top quartile completed in excess of 64.08 hours of instruction.

UPSTART Graduation Rate

Of the 14,125 children documented as enrolled in UPSTART in the tenth year of the program, the Waterford Institute classified 12,354 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 = .68) and with the number of weeks in the program (r = .53).

By these usage requirements, Cohort 10 achieved a graduation rate of 87% (i.e., 12,354/14,125 = 0.87). As seen in **Figure 4**, this graduation rate is similar, but slightly lower than 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 4** displays the demographic characteristics of UPSTART graduates and non-graduates in Cohort 10. 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 4
Demographic Characteristics of C10 Population

Demographic Categories		UPSTART Graduates (N=12,354)	UPSTART Non-Graduates (N=1,771)
Child's Gender	Male	52%	50%
	Female	48%	50%
Child's Ethnicity	White	82%	68%
	Hispanic	11%	19%
	Asian/Pacific Islander	2%	4%
	African American	1%	2%
	Native American	<1%	1%
	Other	2%	4%
Child's Language	English	92%	87%
	Spanish	6%	10%
	Other	2%	1%
Parent Educational Attainment	Some High School	3%	6%
	High School Graduate	11%	20%
	Some College	32%	39%
	College Graduate	41%	27%
	Advanced Degree	13%	8%
Parent Marital Status	Married	93%	81%
	Otherwise	7%	19%
Household Poverty Level	Below 100%	8%	20%
	Below 185%	33%	55%
	Below 200%	34%	56%

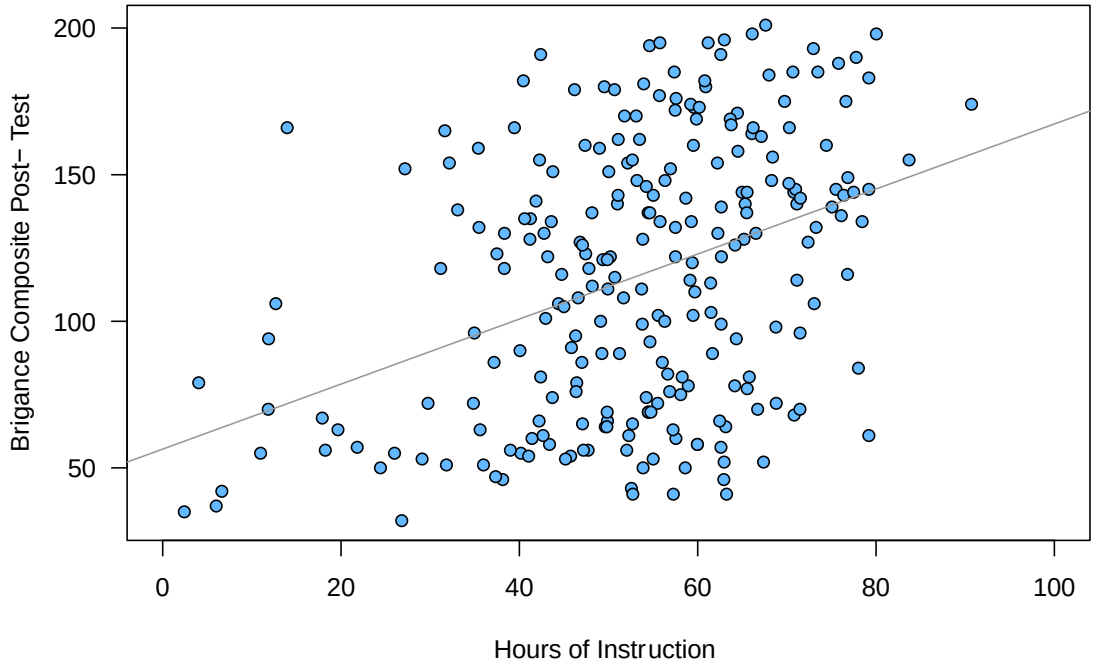
Note: Percentages may not add to 100% due to rounding. Data is from Waterford participant records.

UPSTART Usage and Literacy Outcomes

Similar to previous years, the tenth-year evaluation of UPSTART found curriculum usage to be significantly and positively related to literacy outcomes as measured by composite scores on the general literacy and phonemic manipulation 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=.38). That is, general literacy post-test scores tend to increase with increasing hours of UPSTART usage. The Brigance Post-Test is administered a year following the pre-test and prior to entering Kindergarten. Trained test administrators conduct the testing (see Appendix C for a detailed description of the research methods).

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 phonemic manipulation composite post-test scores indicates a positive linear association between instruction time and scores on the phonemic manipulation post-test (r = .30). 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 tenth 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[3]. We conducted a series of models that explored the impact of household income level, a potentially important predictor of early literacy, on the outcomes of interest and the results were not meaningfully different from our initial analysis. 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.

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

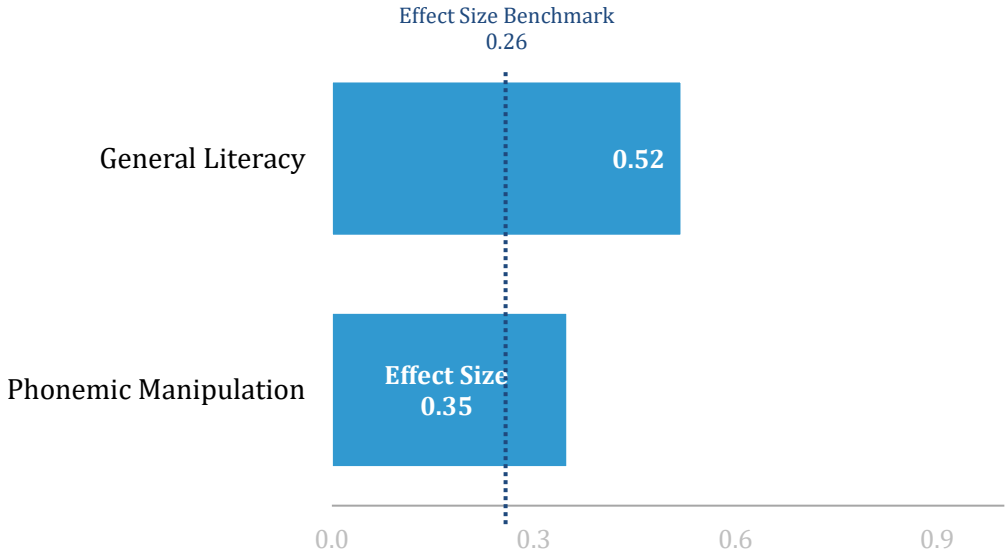
In order to demonstrate the impact of the UPSTART program, we present effect sizes that highlight the differences between UPSTART participants and a matched control group on post-test literacy measures.

Effect sizes[4] were calculated to show the magnitude of UPSTART's impact at post-test as measured by each of the 11 literacy subtests (8 general literacy subtests and 3 phonemic manipulation 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. Appendix B** provides greater detail on how the benchmark was determined.

Combined post-test results showed that UPSTART participation had a significant impact on students' early literacy skill development. In the matched post-test sample[5] for the Brigance and Bader Composite Scores (N=450 and N=474), UPSTART exceeded the threshold for meaningful impact (.52 and .35), well above the observed .26 effect size for similar interventions and evaluation studies (see **Figure 6**).

[3] To create a concise report that highlights the most important findings for stakeholders, we did not present findings that were non-significant in figures.
[4] 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.
[5] Brigance Treatment Group (N = 225); Control Group (N = 225)

Figure 6. General Literacy and Phonemic Manipulation Posttest Analysis of Composite Scores



On average, children participating in UPSTART scored 53.53 points on the general literacy measure before beginning the program and 112.09 points on the general literacy measure after the program was completed. Conversely, control children who were not enrolled in UPSTART scored 53.84 points on the general literacy pre-test and 91.13 points on the general literacy post-test.

With regard to the phonemic manipulation, UPSTART children scored 0.49 points on the instrument at pre-test and 5.91 points at post-test, while their control counterparts scored 0.58 points on the phonemic manipulation pre-test and 3.94 points on the phonemic manipulation post-test.

UPSTART children scored significantly higher than control children on all eight general literacy tests and two of three phonemic manipulation 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 phonemic manipulation ranged from .71 (Segmentation) to .68 (Phonemic Blending) and would be considered meaningful using conventional guidelines, however, Auditory Discrimination was below our threshold for substantive impacts (0.20).

The effect size estimates for each statistically significant literacy subtest (9 out of 11), as measured by the general literacy and phonemic manipulation 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 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

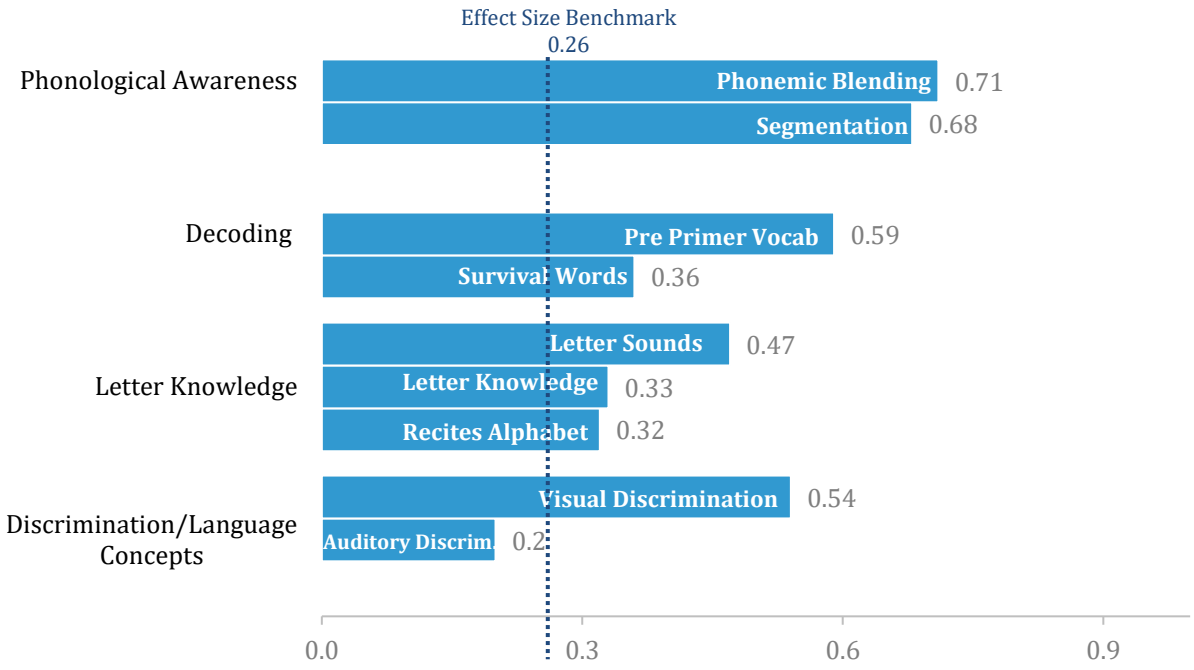
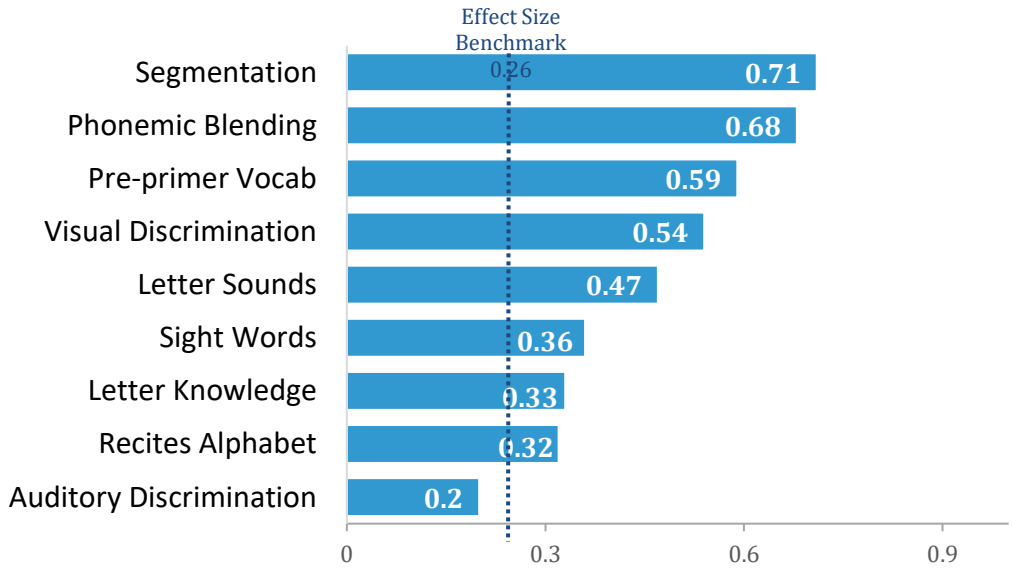


Figure 8 presents the ES of each literacy subtest by the size of their effects along with the .26 effect size benchmark from similar education intervention studies. UPSTART had the largest impact on segmentation (.71), phonemic blending (.68), and pre-primer vocabulary (.59). Effect sizes measuring letter knowledge, decoding, and visual discrimination were above the benchmark of .26 effect size and should be considered practically significant.

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.

Simple linear regression analysis using only the treatment as a predictor produced an estimated effect of 21.51 on the general literacy posttest. Using pretest as a covariate slightly improved the estimate of UPSTART's impact and the linear combination of the treatment and the general literacy composite pretest was significantly related to the general literacy posttest, $R^2 = .33$, adjusted $R^2 = .33$, $F(2, 446) = 110.87$, $p < .0001$ and accounted for 33% 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 general literacy and phonemic manipulation 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 general literacy and phonemic manipulation measures, with the treatment group scoring an average of 4 points higher on phonemic manipulation and 23 points higher on general literacy.

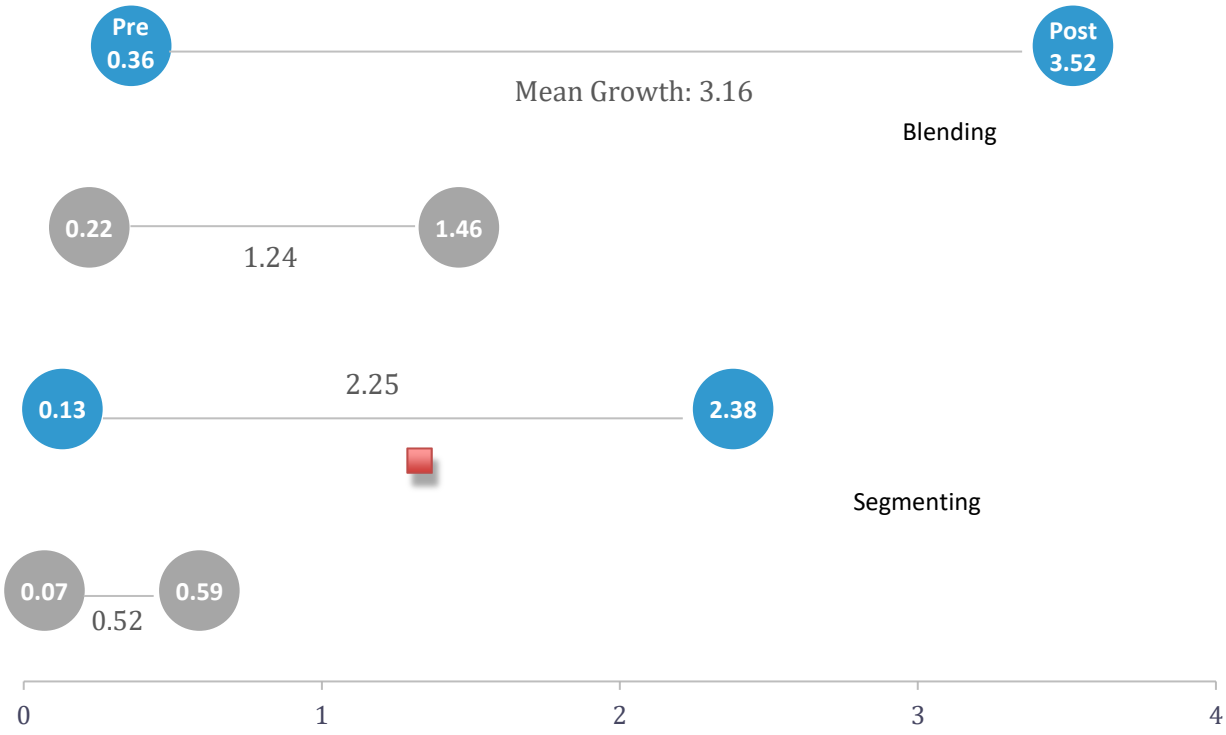
- The treatment group showed statistically stronger ($p < .01$) literacy growth rates compared to the control group on eight of the Brigance subtests (Expressive vocabulary, Visual Discrimination, Recites alphabet, Letter Knowledge, Letter Sounds, Auditory Discrimination, Survival Sight Words, and Basic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

Growth Rates (change over time). Growth rates from pre-test to post-test are shown in the figures below. Each figure categorizes the general literacy and phonemic manipulation 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 two of the three subtests (phonemic blending and segmenting) that measure **Phonological Awareness**. However, for the last subtest,

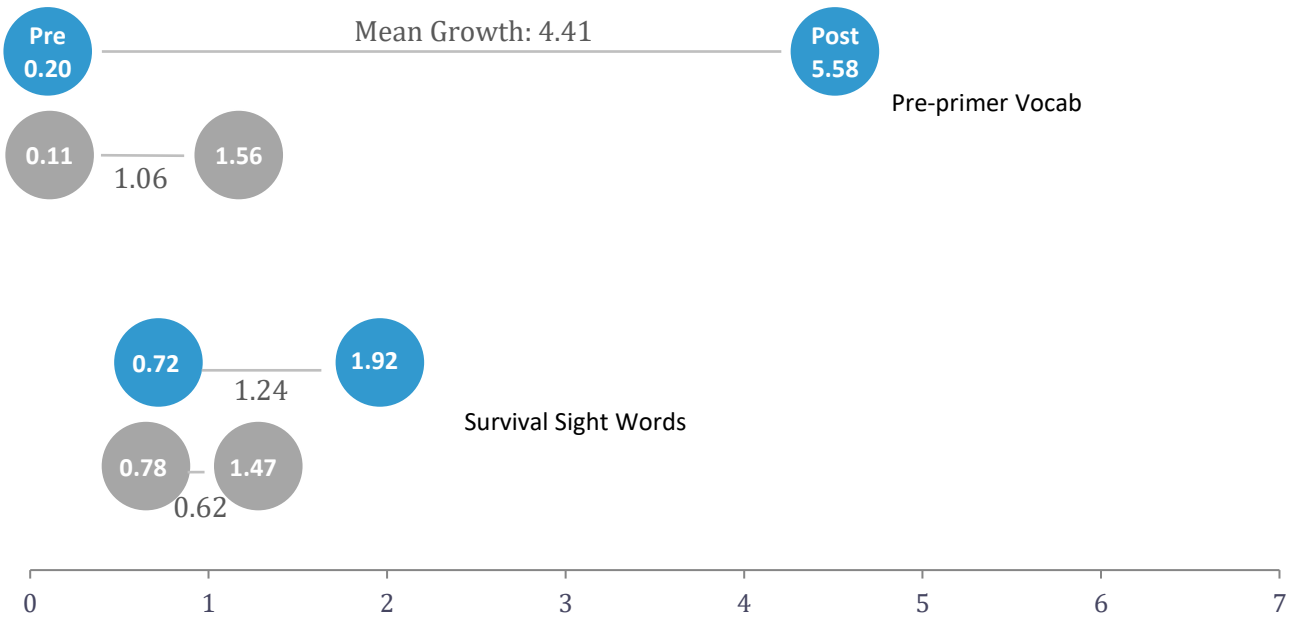
rhyme recognition, control children had a significantly higher growth rate than UPSTART children.

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



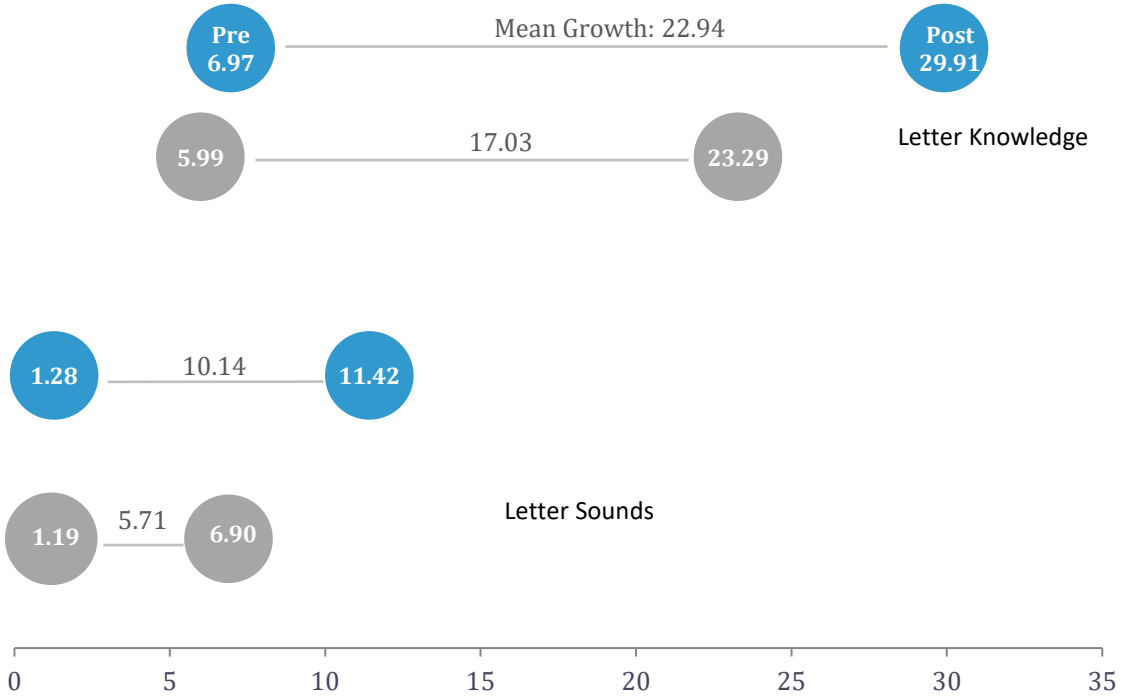
UPSTART students experienced significant, higher mean growth compared to the control group on both subtests used to measure children's **Decoding** ability, including pre-primer vocabulary and survival sight words.

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



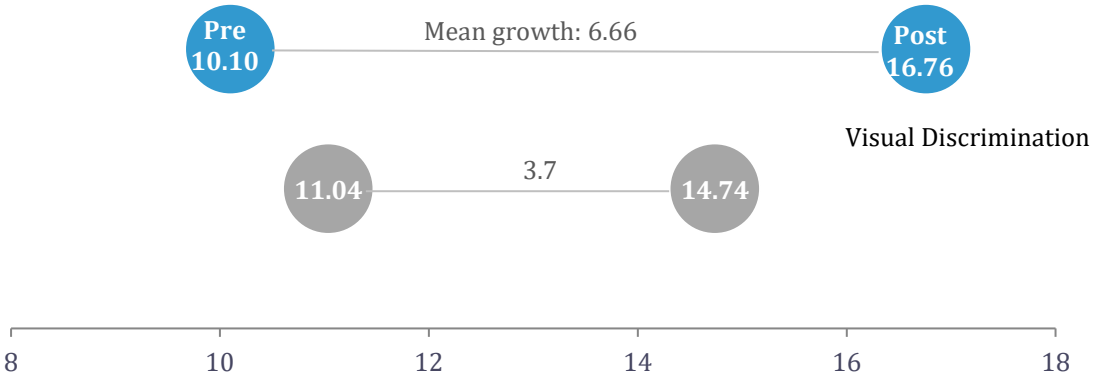
UPSTART children experienced significantly higher growth, compared to non-UPSTART children, in measuring **Letter Knowledge**. UPSTART children showed stronger growth in producing sounds of lower-case letters (letter sounds) and for the Identifying Uppercase Letters or Reciting the Alphabet subtests.

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



Preschool Results Summary and Discussion

This section of the Cohort 10 (C10) 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 maintained enrollment with 14,125 children in Year 10 compared to 14,278 in Year 9. Enrollment continued in areas across the state of Utah and UPSTART has reached families in both rural and urban communities. Thirty-six percent of the children enrolled in Year 10 lived in families with incomes less than 185% of the federal poverty level and the majority of children were White (80%) and English speaking (92%).

Most of the C10 children accessed the UPSTART curriculum through the Waterford website (84%). Approximately 10% of the Year 10 participants received a computer loan and 5% were provided with a computer and Internet. Despite increased enrollment across the state, graduation rates at 87% were relatively consistent with the previous year, though 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

Based on the results from both the effect size and growth score analyses, UPSTART participation had a strong impact on children's emerging literacy skills. Children participating in the program significantly outperformed their non-participating counterparts on literacy measures including 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.

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.

First Grade Analysis

As part of the UPSTART program expansion, stakeholders were interested in the long-term impact of UPSTART on students and whether program benefits can be sustained once children enter elementary school. The First Grade Analysis specifically examines whether the recognized achievement gains from UPSTART are maintained through kindergarten and up until the start of 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 2018-2019 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 was 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 four vendors. Software vendors provided training and support to schools throughout the year and their programs were used in 603 schools and by 38,898 kindergarten students in 2018-19.

Consequently, it is possible that a student who was enrolled in the UPSTART preschool program in 2017-18, matriculated into a kindergarten classroom that was also participating in the EISP program during the 2018-19 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 involved 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 able to create independent and mutually exclusive groups to ascertain the distinct impact of UPSTART on children's literacy outcomes, as well as the combined impact of UPSTART and EISP.

Research Questions

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

> **Research Question 2.1:** *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 up to the beginning of 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 2.2:** *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.

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 9 students through kindergarten and up to the beginning of 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. (see **Table 5**). 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).

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, therefore we utilized statistical match techniques (CEM) to equate the two groups and minimize the presence of preexisting differences.

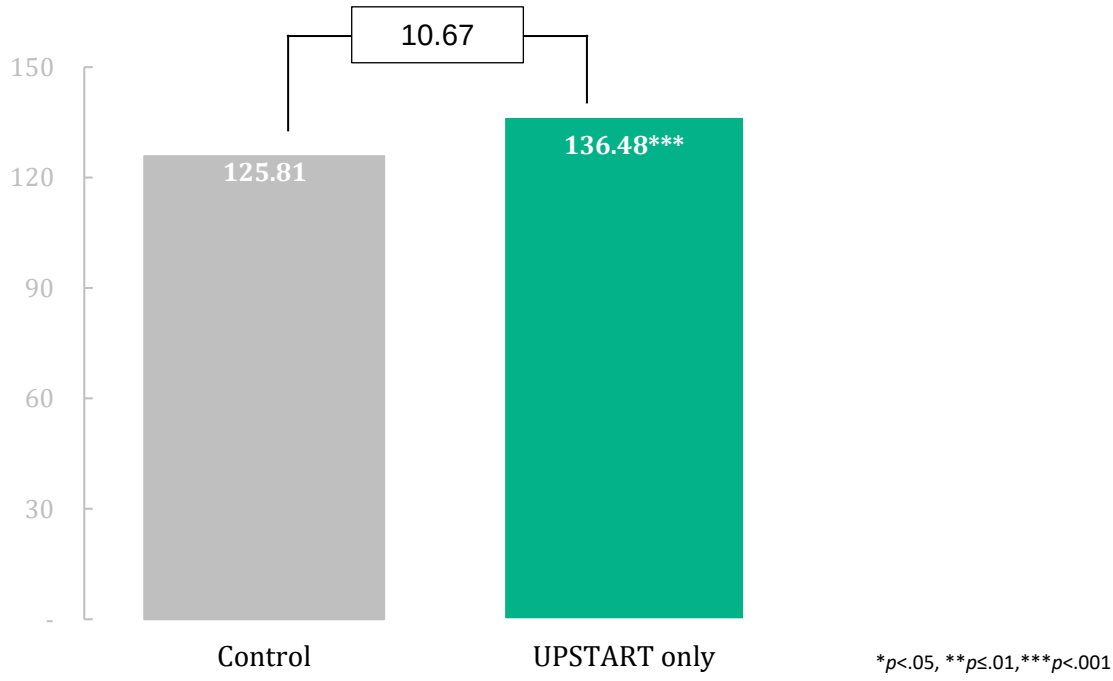
Table 5
First Grade Analysis Evaluation Design

		Preschool 2017-18	Kindergarten 2018-19	First Grade 2019-20
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

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 13**, UPSTART students had an average BOY first grade DIBELS composite score of 136.48 compared to the average score of 125.81 for control students, a 10.67-point difference.

Figure 13. First Grade DIBELS Composite Scores UPSTART and Control students

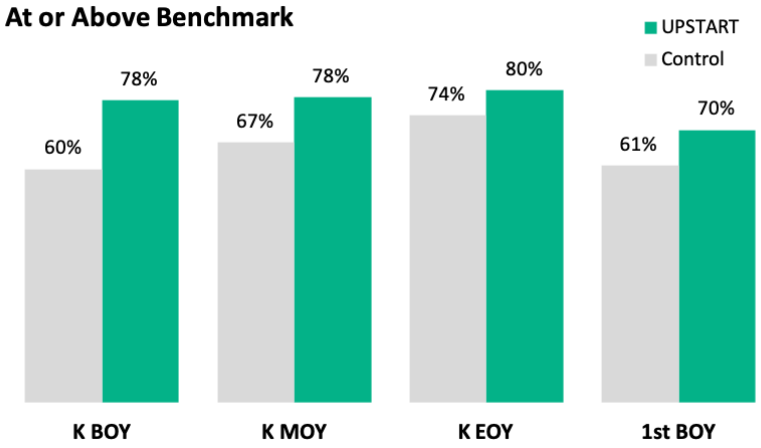


The difference between UPSTART participants and their non-program comparison counterparts on the first grade BOY DIBELS composite produced an effect size of .22,

which is less than the .26 effect size benchmark for similar interventions and evaluation studies. 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 14** 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 (we focus on the first category in our results). Children who received instruction from UPSTART outperformed similar comparison students throughout kindergarten and into first grade. As seen in the **Figure 14** 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 (70% of UPSTART children and 61% of comparison children categorized at or above benchmark) compared to the end of kindergarten (80% of UPSTART children and 74% of comparison children categorized at or above benchmark).

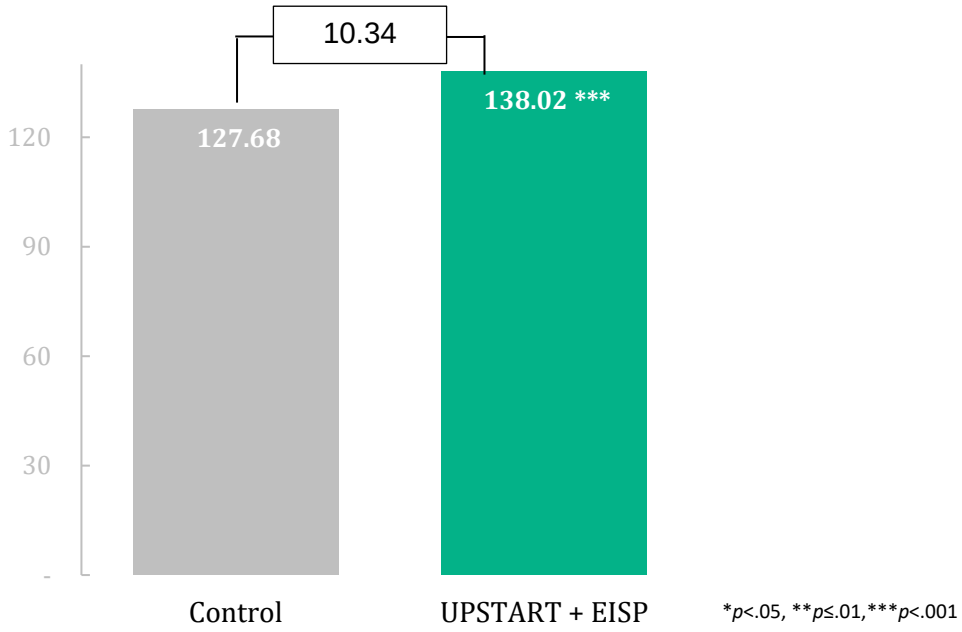
Figure 14. Literacy Benchmarks Over Time: UPSTART only and Comparison Students



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 15**, mean scores on the first grade DIBELS

composite were 138.02 for the UPSTART + EISP treatment group and 127.68 for students who did not receive any literacy software, a 10.34 difference. This difference produced an effect size of .22, which is below the .26 effect size benchmark for similar studies reported in the literature.

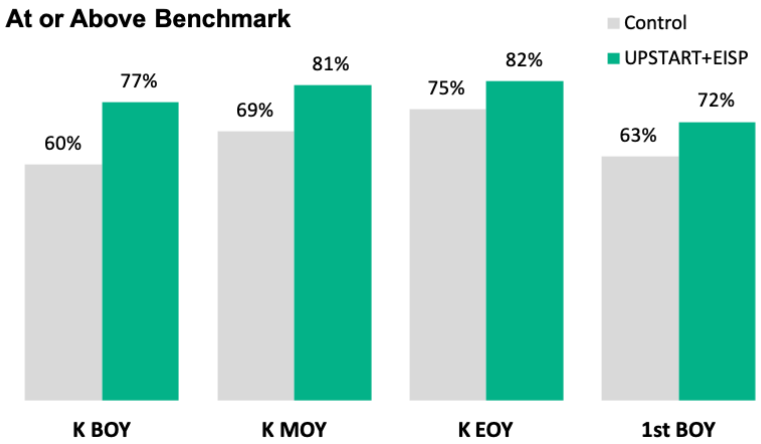
Figure 15. First Grade DIBELS Composite Scores UPSTART + EISP and Control students



The bar graphs in **Figure 16** 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 16**).

There is also an analogous pattern to the UPSTART only analysis of lower levels of literacy achievement at the beginning of first grade, with 72% of UPSTART + EISP children and 63% of comparison children categorized at or above benchmark in first grade, compared to 82% of UPSTART children and 75% of comparison children categorized at or above benchmark at the end of kindergarten.

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



First Grade Results Summary and Discussion

We found meaningful 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 .22) and a similar small impact on students who receive further individualized computer-based instruction (the effect size of the UPSTART + EISP group was .22).

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.

Final 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 the start of first grade. Given the success at improving literacy test scores, we recommend that the state continue providing the UPSTART program to children.

During the 2018-2019 program year, 87% of C10 students were classified as UPSTART graduates, similar to previous graduation rates of C8 and C9 (both 89%). The Cohort 10 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. Given the differences in results for children who meet the graduation requirement compared to those that do not, maintaining optimal program participation levels is important for supporting stronger literacy outcomes, particularly to benefit our most at-risk populations (Evaluation and Training Institute, 2018).

Program Recommendations. Although the graduation rates for C10 students were the similar to 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 equal 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. In order to increase and maintain access to non-program control group families, it is recommended that the state partner with the evaluation team in developing relationships with preschool providers serving low-income families, specifically Head Start organizations, WIC and public preschool programs. The benefits are two-fold where 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 C10 Evaluation Treatment Samples

The matched and unmatched treatment samples are compared with the C10 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 10, with 80% of both unmatched and matched children, being White and 98% classified as English speakers.[6]

Table A.1
Sample Treatment Comparisons on Key Waterford Demographics

Demographic Categories		C10 Population (N = 14,125)	Unmatched Sample (N=246)	Bader Matched Sample (N=237)	Brig Matched Sample (N=225)
Gender	Male	51%	51%	52%	51%
	Female	49%	49%	48%	49%
Ethnicity	White	80%	80%	79%	78%
	Hispanic	12%	19%	19%	20%
Child Language	English	92%	98%	98%	96%
Parent Education Level	Some College	33%	67%	68%	68%
	Bachelor's Degree	40%	4%	3%	4%
Parent Marital Status	Married	91%	82%	81%	81%
Poverty Status	Under 185%	36%	80%	81%	82%

The C10 population had parents with higher college graduation levels and lower levels of poverty. Whereas 33% of the parents in the overall C10 population have a college degree, the modal level of parent education in the unmatched and matched treatment sample was some college (68% respectively). Additionally, 36% of families in the C10 sample were under the 185% federal poverty rate compared to 80% of families in the unmatched sample and 81 and 82% of families in the matched sample for the Bader and Brigance, respectively. 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.

[6] 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 Brigance and Bader 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.

Appendix C: Detailed C10 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).

C10 Evaluation Samples

The C10 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 10 of the program (the 2018-19 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 C10 UPSTART enrollees whose families were contacted about participating in the C10 evaluation[7]. 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 2018 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[8] 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 general literacy data file consisted of 225 treatment children and 225 control children who were matched based on Brigance pre-test scores and other demographic characteristics. The phonemic manipulation file contained data from 474 preschool children (237 treatment children and 237 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 C.1** presents key demographic characteristics by the unmatched treatment and control samples by outcome of interest. As shown in **Tables C.1**, control families were somewhat more advantaged compared to treatment families from the standpoint of parental education and household income level.

[7] C10 treatment families were screened based on location, parental education, income level, child language, and known disabilities.
[8] Snowball sampling is when existing participants recruit future participants among their personal network of acquaintances.

Table C.1
Unmatched Treatment-Control Comparisons on Key Demographics

Demographic Categories		Treatment (N=246)	Control (N=1,162)
Gender	Male	51%	47%
	Female	49%	53%
Ethnicity	Hispanic	19%	19%
	White	80%	82%
Child Language	English	98%	96%
Parent Education Level	HS Diploma	24%**	16%
	Some College	67%***	47%
	Bachelor's Degree	4%	27%***
Parent Marital Status	Married	82%	80%
Household Income	under $10,000	2%	4%
	$10,000-$24,999	13%	15%
	$25,000-$49,999	42%*	35%
	$50,000-$74,999	35%*	28%
	$75,000-$99,999	6%	13%**
	$100k or more	2%	6%*
Literacy Skills	Bader Composite (Pre-Test)	.61	4.65***
	Brigance Composite (Pre-Test)	56.27	72.96***

*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 general literacy and phonemic manipulation 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 general literacy and phonemic manipulation 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 C10 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 C10 unmatched evaluation general literacy sample contained data from 246 treatment students from C10 and 1,162 comparison students who did not participate in the UPSTART program. The unmatched evaluation Bader sample contained data from 246 treatment children and 1,162 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 C. 2.1** displays the demographic breakdown of the matched treatment and control groups on the general literacy test. **Table C. 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 C 2.1
General Literacy Matched Treatment-Control Comparisons on Key Demographics

Demographic Categories		Treatment (N=225)	Control (N=225)
Child Gender	Male	51%	50%
	Female	49%	50%
Child Ethnicity	Hispanic	20%	27%
	White	78%	76%
Child Language	English	96%	97%
Parent Education Level	HS Diploma	24%	24%
	Some College	68%	68%
	Bachelor's Degree	4%	4%
Parent Marital Status	Married	81%	76%
Household Income	under $10,000	2%	3%
	$10,000-$24,999	14%	13%
	$25,000-$49,999	43%	43%
	$50,000-$74,999	36%	36%
	$75,000-$99,999	4%	4%
	$100k or more	1%	1%
Literacy Skills	Brigance Composite (Pre-Test)	53.53	53.84

*p<.05, **p≤.01, ***p≤.001

Table C 2.2
Phonemic Manipulation Matched Treatment-Control Comparisons on Key Demographics

Demographic Categories		Treatment (N=237)	Control (N=237)
Child Gender	Male	52%	46%
	Female	48%	54%
Child Ethnicity	Hispanic	19%	24%
	White	79%	79%
Child Language	English	98%	97%
Parent Education Level	HS Diploma	24%	24%
	Some College	68%	68%
	Bachelor's Degree	3%	3%
Parent Marital Status	Married	81%	75%
Household Income	under $10,000	2%	2%
	$10,000-$24,999	13%	13%
	$25,000-$49,999	43%	43%
	$50,000-$74,999	36%	36%
	$75,000-$99,999	4%	4%
	$100k or more	2%	2%
Literacy Skills	Bader Composite (Pre-Test)	.49	.58

*p<.05, **p≤.01, ***p≤.001

Outcome Measures

The reading skills taught by the Waterford Early Learning Program at Level 1 of the curriculum[9] 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. Eight 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 eight subtests. Possible scores on the Brigance composite range from a low of 0 points to a high of 201 points. We consider the Brigance composite to be a measure of general literacy, and we refer to it as such for the remainder of this report.

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 (phonemic blending, and phoneme segmentation), along with a composite summary phonological awareness score that was calculated by adding the scores from the subtests. We consider the Bader composite to be a measure of phonemic manipulation, and we refer to it as such for the remainder of this report.

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.

[9] 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.

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/he 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 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 sound, 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 C.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 C.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 10 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 general literacy and phonemic manipulation subscales 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 general literacy and phonemic manipulation subscales. The final analysis file consisted of phonemic manipulation data from 474 children, 237 treatment and 237 control, and general literacy data from 450 children, 225 treatment and 225 control, and was based on the subset of children with valid matched pre-test and post-test data, who had not previously used the UPSTART computerized learning program as documented through the pre-screening interview.

Appendix D: Detailed First Grade Analysis Methods

This section expands upon 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 group post-program only design. The evaluation design is diagrammed below in **Table D.1**.

Treatment children participated in UPSTART during the eighth year of implementation (Cohort 9) the 2017-18 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 D. 1
First Grade Analysis Evaluation Design

		Preschool 2017-18	**Kindergarten 2018-19**	**First Grade 2019-20**
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 D.2**).

Table D.2
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 2019-20 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 9 during 2017-18 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 2018-19 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 D.3
Group Sizes for Unmatched First Grade Analysis File

Group	N=
UPSTART only	4,020
UPSTART + EISP	4,810
EISP only	17,024
Control	13,261
Total	**39,118**

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 10 Evaluation**.) Our final analytic samples consisted of two data files: (1) one data file containing UPSTART only students (N = 4,012) and a matched comparison sample (N = 4,012) of students did not have UPSTART or EISP program experience and (2) a second data file containing UPSTART plus EISP students (N = 4,800) and a matched comparison sample (N = 4,800) of students who did participate in either program.

Table D.4 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 D.4
UPSTART Only and Control Student Comparisons on Key Demographics

Demographic Categories		Treatment (N=4,012)	Control (N=4,012)
Child Gender	Female	50%	49%
	Male	50%	51%
Child Race	White	83%	83%
	Hispanic	9%	9%
Child Language	English Language Learner	2%	2%
Title 1 School	Yes	11%	11%
	Targeted for Individual Students	7%	7%
Household Income	Low Income	12%	12%

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 D.5**.

Table D.5
UPSTART Only + EISP Program – Control Comparisons on Key Demographics

Demographic Categories		Treatment (N=4,800)	Control (N=4,800)
Child Gender	Female	49%	48%
	Male	51%	52%
Child Race	White	84%	84%
	Hispanic	10%	10%
Child Language	English Language Learner	2%	2%
Title 1 School	Yes	12%	12%
	Targeted for Individual Students	6%	6%
Household Income	Low Income	10%	10%