

UPSTART Program Evaluation

Year 7 Program Results

Submitted to the Utah State Board of Education
April, 2017



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

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

Table of Contents

Executive Summary	3
Program Implementation	3
Impacts on Literacy	4
Recommendations	5
Preface	7
Introduction	8
UPSTART Program Description	8
Evaluation Research Questions	8
Research Methods	10
Research Design	10
C7 Evaluation Samples	10
Sample Data Collection	11
Matched Treatment-Control Group Sample	13
Outcome Measures	14
Data Collection Procedures	16
UPSTART Program Implementation	18
UPSTART Enrollment	18
Provided UPSTART Equipment	20
UPSTART Usage	20
UPSTART Graduates	22
UPSTART Usage and Literacy Outcomes	23
UPSTART Program Impacts on Literacy	25
Do UPSTART students have better literacy skills at entry to Kindergarten than control students?	25
Do UPSTART students show stronger literacy growth rates from preschool to Kindergarten than control students?	29
Discussion and Recommendations	32
Program Implementation	32
Program Impacts on Literacy Development	33
Limitations	33
Recommendations	34
References	37
Appendix A: Comparison of Evaluation Samples	39
Appendix B: Determining UPSTART Effect Size Benchmark	40

Executive Summary

Utah Preparing Students Today for a Rewarding Tomorrow (UPSTART) is a home-based computer preschool program developed and provided by the Waterford Institute to prepare young children for school entry and future academic success. The Evaluation and Training Institute (ETI), the external evaluator of UPSTART since 2009, has prepared this report for the Utah State Board of Education (USBE) to document UPSTART's impact in its seventh year of implementation (Cohort 7, enrolled during the 2015-2016 program year).

The evaluation of UPSTART's seventh cohort moved from using a nonequivalent control group seen in previous years to a pre-test/post-test design with a statistically matched control group for assessing the program's impact on developing children's early literacy skills in preschool. This increased our ability to control for other variables that influence literacy growth over time. 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.

Program Implementation

Over the past seven years, UPSTART program participation has increased and the program has enrolled families in urban and rural areas throughout the state of Utah. Sixty-four percent of children enrolled in UPSTART Cohort 7 lived in families with incomes less than 200% of the federal poverty level and the majority of enrolled children were White (77%) and English speaking (88%). UPSTART enrollment increased from 5,091 children in Year 6 to 6,639 children in Year 7, an increase of 30 percent, while graduation rates declined from 92% to 87%.

Findings about UPSTART usage are summarized below:

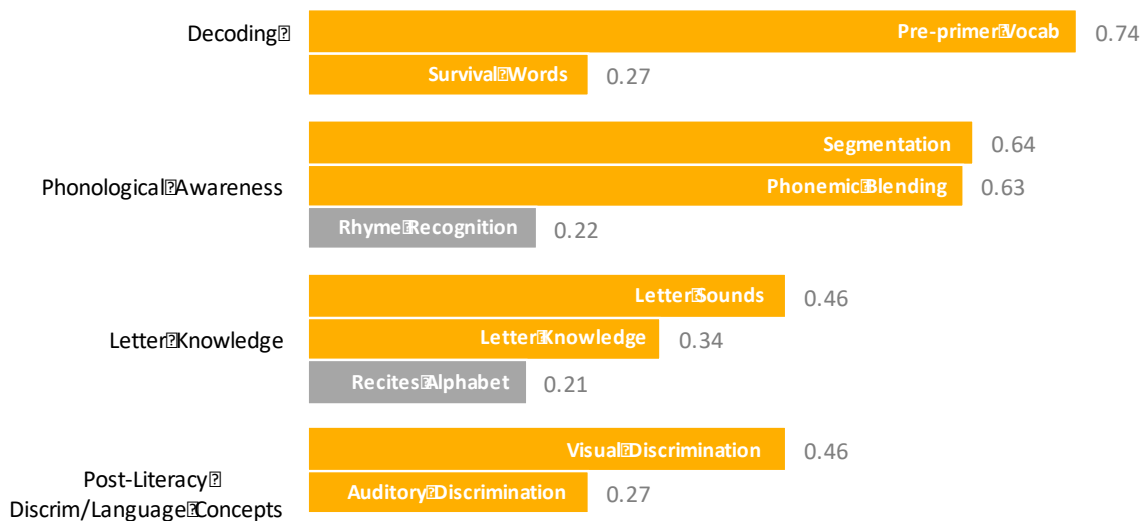
- Students who used the program for the recommended amount of time (or longer) had better reading outcomes than their matched counterparts who did not use the program.
- Students in Cohort 7 used the UPSTART program for an average of 60 hours during the program year, which is down from the average use seen in the previous Cohort (Cohort 6; who used the program for an average of 67 hours in the 2014-2015 program year).
- Students in Cohort 7 had an 87% graduation rate, which is five percent lower than the graduation rate for Cohort 6 (92%). This follows a trend of lower graduation rates year-to-year starting in Cohort 5 (which had a graduation rate of 94%)
- Families who did not graduate were more likely to have parents with lower levels of education, have children who spoke a language other than English, and have higher levels of household poverty than families who graduated and completed the UPSTART program.
- Cohort 7 UPSTART graduates had an average program use of 66 hours.
- A positive relationship was found between UPSTART curriculum use: as program use increased, students' scores on literacy outcome measures increased.

Impacts on Literacy

UPSTART had a strong impact on children's emerging literacy skills based on results from effect size and growth score analyses. Children enrolled in UPSTART produced significant positive effects (ES = .52) compared to control children on the Brigance composite, an instrument that measures decoding skills, letter knowledge, vocabulary and syntax, and pre-literacy discrimination. Similarly, UPSTART participants experienced medium effects (ES = .62) on the Bader composite, an instrument that assesses children's phonological awareness. Effect sizes that fall above a predetermined threshold of similar education evaluation studies (.26 or higher) are highlighted in gold in the graph below to showcase their practical significance.

UPSTART had a significant impact on a majority of early literacy domains

Effect Size Estimates by Literacy Construct



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 (medium ES = .74) and reading survival sight words (small ES = .27).
- 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.

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 phoneme segmenting skills (medium ES = .64), phonemic blending skills (medium ES = .63), and facility with rhyme recognition (small ES = .22).
- Compared to control children, students participating in UPSTART had significantly higher increases from the pre-test to the post-test on all three phonological awareness subscales (rhyming, blending, and segmenting).

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

- UPSTART children had small to medium effects in their learning how to recite (ES = .21), identify (ES = .34), and sound out (ES = .46) letters of the alphabet.
- Compared to control students, UPSTART participants showed significantly stronger growth rates in learning how to pronounce letter sounds.

Before children can read, they need to be able to visually distinguish between shapes, letters, and words, even if they do not fully comprehend what letters represent. Similarly, children should be able to differentiate between spoken words (e.g., "fit" versus "fat") before comprehending written words. UPSTART participants showed a moderate impact on **pre-literacy discrimination and language concepts**.

- UPSTART had a medium effect on children's ability to discriminate between different shapes, letters, and words (ES = .46) as well as a small to medium effect on their ability to distinguish if two words sound the same (ES = .27).
- Children in UPSTART had stronger growth scores on their auditory discrimination of words when contrasted to children not enrolled in UPSTART.

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

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

Summary and Recommendations

The UPSTART program continues to show success in helping preschool aged children develop literacy skills in preparation for their entry into kindergarten. These outcomes would have specific benefits to at-risk children, whose families struggle with poverty and other issues, and often lack the resources to help their children develop the literacy skills needed to succeed in school.

Enrollment continued to increase in the 2015-2016 program year, which resulted in the program reaching more families with at-risk children. However, less Cohort 7 students were classified as graduates when compared to previous cohorts, and the lower graduation rates were more prominent in the most at-risk populations. In addition to the

decline in graduation rates, average program usage dropped approximately 7 hours when compared to the previous year, and evidence shows how important graduation status is for achieving strong program outcomes.

Program graduates had significantly higher scores on literacy outcome measures compared to children who did not meet program usage requirements. Additionally, the children who did not graduate from UPSTART were more likely than program graduates to reside in households below the poverty level, have parents with lower levels of education, and be English learners – the ideal target population for UPSTART and the children that stand to benefit most from the program. These reductions in program use and declines in graduation need to be monitored to be sure that UPSTART is being administered with fidelity so that all children can receive the full program dosage as recommended by the vendor.

Given the strong impact on early literacy development, we recommend that the state continue to provide the UPSTART program to children. Given the decline in graduation rates, we recommend that the program vendor work with the evaluator and USBE staff to monitor program implementation carefully and be sure the current trend towards lower graduation rates and lower program use does not continue. 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:
 - The use of social networks and/or social media to reach parents who are difficult to communicate with.
 - Developing targeted incentives for families with the highest risk factors for not meeting program usage requirements, such as monthly awards (extrinsic), being highlighted in UPSTART communications to social networks as “Gold Star Families” (intrinsic).
- The program vendor could highlight the benefits of meeting usage guidelines by showing parents evaluation information that links graduation status with reading outcomes (e.g., through slides during training, brief informational guides, e-mails, etc.). The evaluation data can provide evidence about program effects and motivate parents to meet UPSTART usage recommendations.

Preface

The Utah State Office of Education (USOE) hired the Evaluation and Training Institute (ETI), a non-profit research and consulting firm, to conduct a multi-year evaluation of the UPSTART program to determine the effectiveness of the home-based preschool program in academically preparing children for school success. This report includes evaluation results for UPSTART's seventh year of implementation during the 2015-2016 program year, hereafter referred to as Cohort 7 (C7).

The 2015-2016 program year was the continuation of the UPSTART expansion seen during the 2014-2015 program year with Cohort 6, which saw the program's use increased to reach more families than in any previous cohort. The UPSTART expansion was due in part to empirical evidence from previous positive program evaluation findings (Evaluation and Training Institute 2011, 2012a, 2012b, 2013, 2014). As the program scaled-up, the evaluation had to be adapted to accommodate larger numbers of program students, and higher stakes related to greater resource allocation for the program. While the scale and stakes increased, our research objectives remained constant: we continued to evaluate the program's *impact on developing children's early literacy skills in preschool* to help the state and stakeholders determine the benefits from participating in the program.

We enhanced the established evaluation design to meet a higher level of accountability for the Cohort 7 students, and to ensure that the program resources were having a positive impact on school readiness. The Cohort 7 evaluation included a balanced one-to-one match of treatment and control students. While requiring a larger sample size, the matching process enhanced our ability to detect treatment effects and, in general, improve the accuracy of the evaluation results.

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

We would like to acknowledge and thank reviewers from the USBE, the UPSTART Advisory Committee, and the Waterford Research Institute for their efforts and expertise that they contributed to reviewing a draft of this report.

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

Introduction

UPSTART Program Description

Utah Preparing Students Today for a Rewarding Tomorrow (UPSTART) is a pilot project established by the Utah state legislature that uses a home-based education technology approach to develop the school readiness skills of preschool children. In its seventh year of operation during the 2015-16 school year, the project's implementation contractor – the Waterford Institute – enrolled 6,639 preschool children and provided them with an adaptive program of computer-based early literacy instruction to prepare them academically for kindergarten. The 6,639 children enrolled in the seventh year cohort, hereafter referred to as Cohort 7 (C7), participated in UPSTART from September 2015 through June 2016. Cohort 7 is the largest group since the program's rollout.

The UPSTART software uses adaptive lessons, digital books, songs, and activities to deliver early literacy content. The reading skills taught by the Waterford Early Learning Program at Level 1 of the curriculum¹ include:

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

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

Evaluation Research Questions

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

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.

With respect to concerns for school readiness, our research questions for the C7 evaluation study were as follows:

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

3. What was the extent of UPSTART curriculum usage in terms of the amount of exposure per participant, as measured in minutes or hours of instruction per week?
4. What percent of the participants completed the full implementation program (i.e., “graduated” as defined by the Waterford Institute)?
5. How does the level of UPSTART curriculum usage relate to reading readiness outcomes?

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

Research Methods

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

Research Design

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

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

The use of both a pre-test and a comparison group facilitates our ability to examine potential threats to validity, which could jeopardize a clear interpretation of the results (Shadish, Cook, & Campbell, 2002). Because students could not be randomly assigned to treatment or control groups, the groups 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).

C7 Evaluation Samples

The C7 evaluation moved from using an unmatched group seen in previous years (Evaluation and Training Institute 2011, 2012a, 2012b, 2013, 2014, 2015), to a new approach first adopted in the C6 evaluation (Evaluation and Training Institute, 2016) 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.

Sample Data Collection

The C7 study consists of data from 581 preschool children: 230 treatment group children who had enrolled in UPSTART for Year 7 of the program (the 2015-16 school year) and 351 nonparticipating control group children. The children were not randomly assigned to the treatment or control groups.

Treatment children. The 230 UPSTART children came from an initial random sample of C7 UPSTART enrollees whose families were contacted about participating in the C7 evaluation². 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 230 UPSTART children participated in pre-testing prior to entering the program over the summer of 2015 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 sampling³ 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.

Table 1 presents key demographic characteristics for the unmatched treatment and control sample. As shown in **Table 1**, control families were somewhat more advantaged compared to treatment families from the standpoint of parental education and household income level.

² C7 treatment families were screened based on location, parental education, income level, child language, and known disabilities.

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

Table 1
Unmatched Treatment-Control Comparisons on Key Demographics

Demographic Categories		Treatment (N=230)	Control (N=351)
Gender	Female	47%	52%
	Male	51%	48%
Ethnicity	Caucasian	89%	81%
	Hispanic	7%	11%
Child Language	English	94%	92%
Parent Education Level	High School Diploma	14%	11%
	Some College	49%	45%
	Bachelor's degree	31%	33%
	Graduate degree	4%	5%
Parent Marital Status	Married	88%	84%
Household Income	Under \$10,000	6%	5%
	\$10k-\$24,999	14%	13%
	\$25k-\$49,999	51%	26%
	\$50k-\$74,999	24%	28%
	\$75k-\$99,999	4%	17%
	\$100k or more	2%	11%

Studies of child development have found that parents with higher levels of education spend more time with their children in ways likely to enhance their development, hold higher expectations for their children, and use varied and complex language and speech patterns (Davis-Kean, 2005; Guryan et al., 2008; Neitzel & Stright, 2004). Similarly, children residing in households with financial hardship experience lower levels of language/literacy stimulation which impact school readiness outcomes (Mistry et al., 2010). Thus it is important to ensure that the treatment and control groups are as comparable as possible with regard to parental education and income before analysis or that statistical adjustments are performed to determine any impact of family characteristics on post-test literacy outcomes.

Significant differences between the treatment and control groups that favored the control group were found on both the Brigance and Bader pre-test literacy instruments. While the use of a pre-test and covariates with the full unmatched sample allows us to examine and statistically control for pre-existing literacy skills and demographic differences between the treatment and control groups, using these control methods can reduce our ability to detect treatment effects and to estimate their size. We determined that using a matched treatment and control group strategy that took into account pre-test performance and 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 C7 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 C7 unmatched evaluation sample contained data from 230 treatment students from C7 and 351 comparison students who did not participate in the UPSTART program.
2. Students from the pool of potential controls were then matched to treatment students using CEM, which found an exact match—or twin—for treatment students from the group of control students in terms of:
 - Sex (Female/Male)
 - Ethnicity (White or Hispanic),
 - Language
 - Parent Education
 - Household income
 - 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. **Table 2** displays the demographic breakdown of the matched treatment and control groups. Note how the two groups in the matched sample are much more similar in terms of parental education than in the unmatched sample. There is, however, still a slight difference in income levels between the two samples, with the control group having the advantage of a larger proportion of families with higher household incomes. Using household income in the matching process resulted in a drastically smaller sample size which would limit our ability to test for significant effects, so we elected to retain the slight imbalance and evaluate the impact of household income during the statistical analysis process.

Table 2
Matched Treatment-Control Comparisons on Key Demographics

Demographic Categories		Treatment (N=208)	Control (N=208)
Child Gender	Female	48%	51%
	Male	52%	49%
Child Ethnicity	Caucasian	91%	91%
	Hispanic	6%	6%
Child Language	English	96%	98%
Parent Education Level	High School Diploma	14%	12%
	Some College	50%	50%
	Bachelor's degree	30%	30%
	Graduate degree	3%	5%
Parent Marital Status	Married	88%	87%
Household Income	Under \$10,000	5%	2%
	\$10k-\$24,999	15%	10%
	\$25k-\$49,999	50%	29%
	\$50k-\$74,999	24%	30%
	\$75k-\$99,999	4%	17%
	\$100k or more	2%	12%

Outcome Measures

The reading skills taught by the Waterford Early Learning Program at Level 1 of the curriculum⁴ include:

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

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

The Brigance. The Brigance Inventory of Educational Development (Brigance, 2014) was selected as an early literacy measure of phonics and vocabulary knowledge and as a measure of pre-Kindergarten academic and cognitive skills. Ten scales were administered from the language development and academic/cognitive domains of the Brigance. Brigance subscales measured the literacy constructs of *vocabulary*, *pre-literacy discrimination*, *letter knowledge*, and *decoding* and are described in detail in

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

Table 3. A composite Brigance score to create a comprehensive score of early literacy achievement was created by adding the scores from the ten subtests. Possible scores on the Brigance composite range from a low of 0 points to a high of 240 points.

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

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, letter knowledge, phonological awareness, decoding, and vocabulary. These five constructs are explained in further detail below.

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

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

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

Decoding. Decoding is the process of translating printed words into speech and is the precursor to reading fluency, the ability to read text accurately and quickly, either aloud or silent. Decoding relies on the successful acquisition of all the aforementioned reading skills, phonological awareness, letter knowledge, and pre-literacy discrimination. We measured decoding in the UPSTART study by asking children to read lists of simple pre-primer vocabulary (e.g., "and", "can",

“go”, “look”) and presenting them with words they might have seen in their everyday lives (e.g., “stop”, “in”, “out”).

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

Table 3 summarizes the alignment between the UPSTART curriculum and the literacy constructs measured by the Brigance and Bader, and also contains information about specific skills assessed by the Brigance and Bader subscales, along with possible scale ranges.

Table 3
Alignment of Outcome Measures with UPSTART Curriculum

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

Note: An asterisk indicates that the scale was administered during the pre-test and post-test data collection periods.

Data Collection Procedures

Data were collected for 230 treatment group children who had enrolled in UPSTART for Year 7 of the program and 351 control group children who had not enrolled in the UPSTART program. The children’s parents were given an intake questionnaire during the pre-test session that collected demographic information from children, parents, and the household. The children were post-tested on the Brigance and Bader a year later before entering kindergarten.

A student data file was developed based on data collected from the intake questionnaire and from the pre-test and post-test administrations of the Brigance and Bader. The final

analysis file was based on the subset of children with valid matched pre-test and post-test data, and who had not previously used the UPSTART computerized learning program as documented through the pre-screening interview.

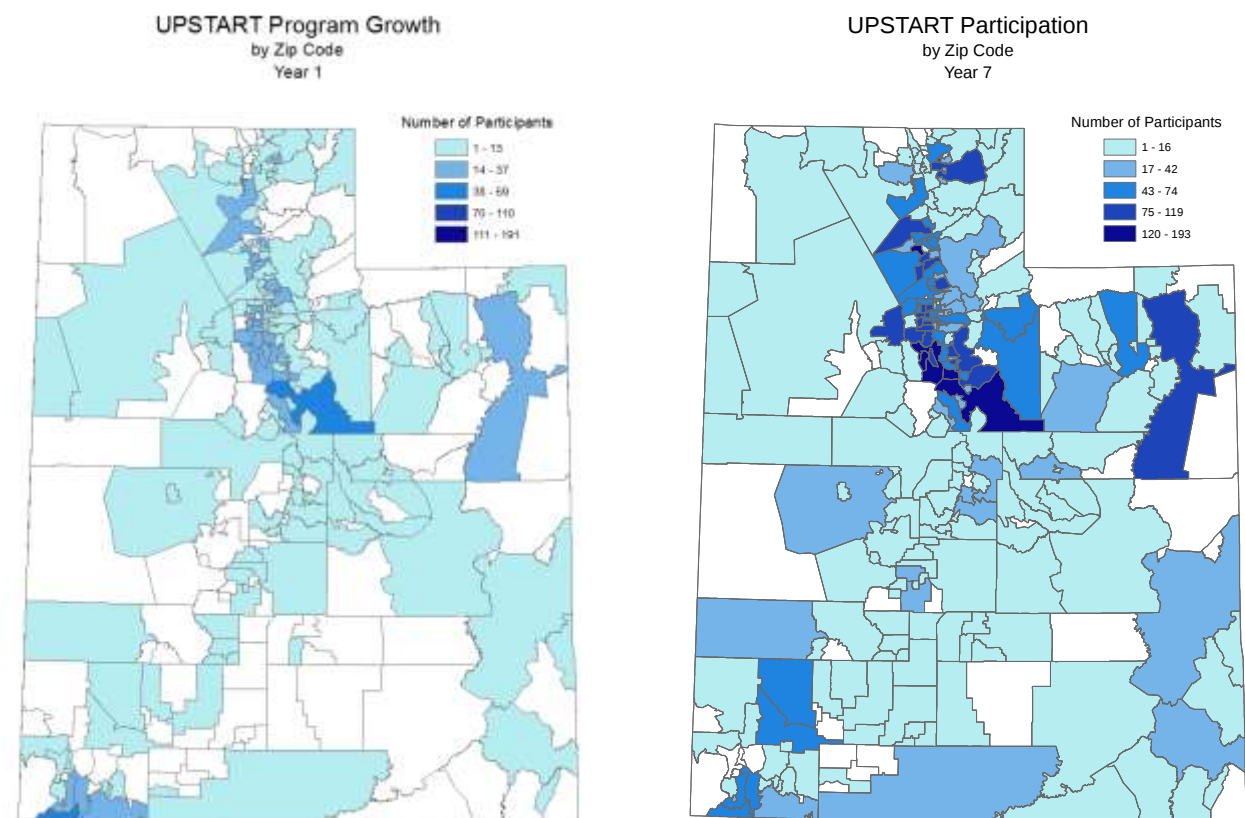
UPSTART Program Implementation

Findings reviewed in the UPSTART implementation section include seventh 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 2015-16 program year was a continued expansion of UPSTART enrollment, as the number of preschool students participating in the program in Year 7 (N=6,639) grew by 1,548 students from the previous year (Year 6, N=5,091), a 30 percent increase. Since the inception of the program, the number of students enrolled in the program rose from 1,577 children in Year 1 to 6,639 students in Year 7, an increase of over 320 percent. The maps depicted in **Figure 1** showcase UPSTART program participation by student zip code from the inception of the program (Year 1, N=1,248) to the most recent Year 7 (Year 7, N=6,639). As seen below in **Figure 1**, the UPSTART program has continued to further its reach over the past seven years and has increased enrollment in both urban and rural areas of the state.

Figure 1. Map of UPSTART program participation in Year 1 and Year 7



The Waterford Institute provided documentation for the seventh-year UPSTART enrollment of 6,639 children, including demographic information, provisioned educational technology, UPSTART program usage, and whether or not children completed program requirements. Some basic demographic characteristics of the C7 population are presented below in **Table 4**.

Table 4
Demographic Characteristics of C7 Population

Demographic Categories		All C7 UPSTART (N=6,639)
Child's Gender	Male	51%
	Female	49%
Child's Ethnicity	White	77%
	Hispanic	16%
	Asian/Pacific Islander	3%
	African American	1%
	Native American	<1%
	Other	2%
Child's Language	English	88%
	Spanish	11%
	Other	1%
Parent Educational Attainment	Some High School	5%
	High School Graduate	15%
	Some College	38%
	College Graduate	35%
	Advanced Degree	8%
Parent Marital Status	Married	91%
	Otherwise	9%
Household Poverty Level	Below 100%	20%
	Below 185%	59%
	Below 200%	64%

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

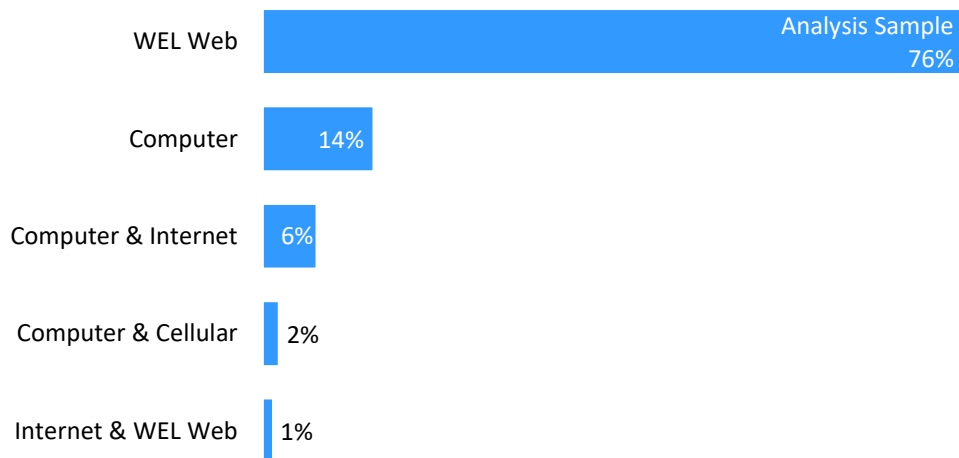
Slightly more C7 boys (51%) were enrolled than girls (49%) and in terms of ethnicity, the majority (77%) of the C7 enrollment was White, with 16% of the children being of Hispanic origin. Fifty-nine percent of the C7 UPSTART participants lived in families with incomes less than 185% of the federal poverty level.⁵

⁵ The federal poverty definition consists of a series of thresholds based on family size. In 2016, a 100% poverty threshold for a family of four was \$24,300, while a 185% threshold for a family of four was \$44,995.

Provided UPSTART Equipment

The type of education technology provided to UPSTART children in Year 7 of the program is shown in **Figure 2** for all 6,639 children enrolled in the program. The majority of UPSTART children (76%) 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 C7 Participants by Waterford



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

Second most frequently, UPSTART provided free personal computers to 14% of the C7 children while they participated in the program. Another 6% of the C7 program participants were provided with internet access and personal computers. The remaining 4 percent of the C7 enrollment received computers and wireless access (2%), internet and access to the Waterford website (1%), or participated in a lending library program (1%) to enable them to access the UPSTART curriculum (see **Figure 2** for details).

UPSTART Usage

We reviewed program usage (time spent using the software program) for three groups: all UPSTART participants, UPSTART program graduates, and the evaluation analysis sample. The hours of instruction observed for all children documented as enrolled in the seventh year of UPSTART are summarized in **Table 5**, and are compared to program “graduates”. The average level of usage for all students enrolled in the seventh year of UPSTART (N=6,639) was approximately 60 hours of instruction; this is less than the average level of usage as documented in the sixth year of the program (67 hours for C6; see Evaluation and Training Institute, 2016). The C7 academic year covered 44 weeks of instruction, beginning the week of August 31, 2015 and ending June 27, 2016.

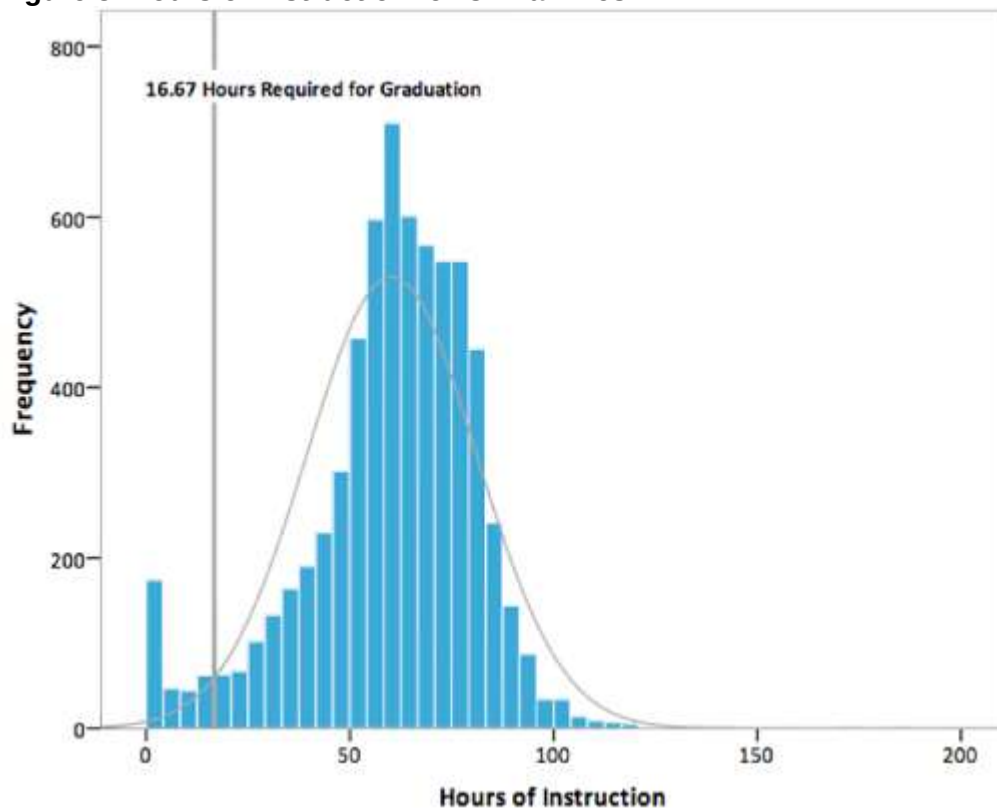
Table 5
C7 Hours of UPSTART Instruction

Group	N	Mean	SD	Range
All UPSTART	6,639	60.25	20.83	00.00 - 183.67
UPSTART Graduates	5,804	65.78	14.93	16.81 - 183.67
UPSTART Analysis Sample	208	65.48	19.00	4.40 - 127.97

Sixty-five of the enrolled families who were provided instructional equipment (e.g., computers, an Internet subscription, and a computer drive) did not log any instructional time in the UPSTART curriculum during Year 7 of the program. For enrolled families whose children did use the curriculum, the average duration in the program was approximately 41 weeks. This usage pattern is similar to that observed in the sixth year of the program. The children in the C7 evaluation analysis sample used the UPSTART curriculum for approximately 65 hours of instruction on the average (see **Table 6**).

The histogram in **Figure 3** shows the distribution of hours of instruction for the total C7 population (N=6,639). As noted previously, sixty-five of the enrolled children logged zero hours of instruction during their time in UPSTART. At the other end of the spectrum, seven children logged over 130 hours of instruction.

Figure 3. Hours of Instruction for C7 Families



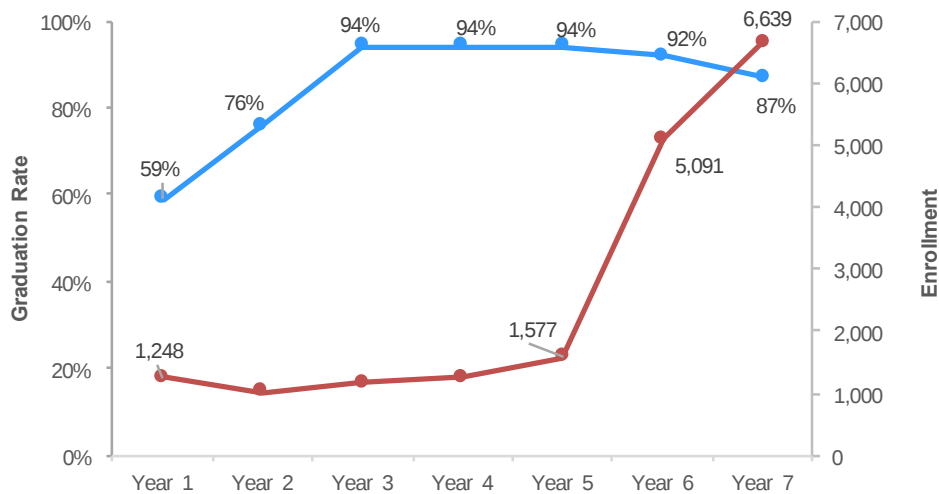
The bottom quartile of the C7 population completed 50.92 hours of instruction or less, the midpoint of the C7 distribution was 62.34 hours, and the top quartile completed in excess of 74.35 hours of instruction.

UPSTART Graduation Rate

Of the 6,639 children documented as enrolled in UPSTART in the seventh year of the program, the Waterford Institute classified 5,804 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 ($r = .70$) and with the number of weeks in the program ($r = .58$).

By these usage requirements, Cohort 7 achieved a graduation rate of 87% (i.e., $5,804/6,639 = 0.87$). As seen in **Figure 4**, this is a lower rate than the previous four years, which may be due to increased enrollment and the steady rise in participants across the last two years (the UPSTART population increased from 1,577 children in Year 5 to over 5,000 children in Year 6 and over 6,600 in Year 7).

Figure 4. UPSTART Graduation Rates over Time



In order to further examine the features of program graduates and non-graduates, **Table 6** displays the demographic characteristics of UPSTART graduates and non-graduates in Cohort 7. Children who did not meet the program usage requirement were significantly 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, and reside in families with higher levels of poverty. We further explore the impact of graduation status on literacy outcomes in the **Program Impacts** section of this report.

Table 6
Demographic Characteristics of C7 Population

Demographic Categories		UPSTART Graduates (N=5,804)	UPSTART Non-Graduates (N=835)
Child's Gender	Male	51%	52%
	Female	49%	48%
Child's Ethnicity	White	79%	68%
	Hispanic	15%	23%
	Asian/Pacific Islander	3%	4%
	African American	1%	1%
	Native American	1%	2%
	Other	2%	3%
Child's Language	English	88%	84%
	Spanish	11%	15%
	Other	1%	1%
Parent Educational Attainment	Some High School	4%	10%
	High School Graduate	13%	22%
	Some College	37%	42%
	College Graduate	37%	23%
	Advanced Degree	8%	3%
Parent Marital Status	Married	93%	80%
	Otherwise	7%	20%
Household Poverty Level	Below 100%	18%	35%
	Below 185%	57%	72%
	Below 200%	62%	75%

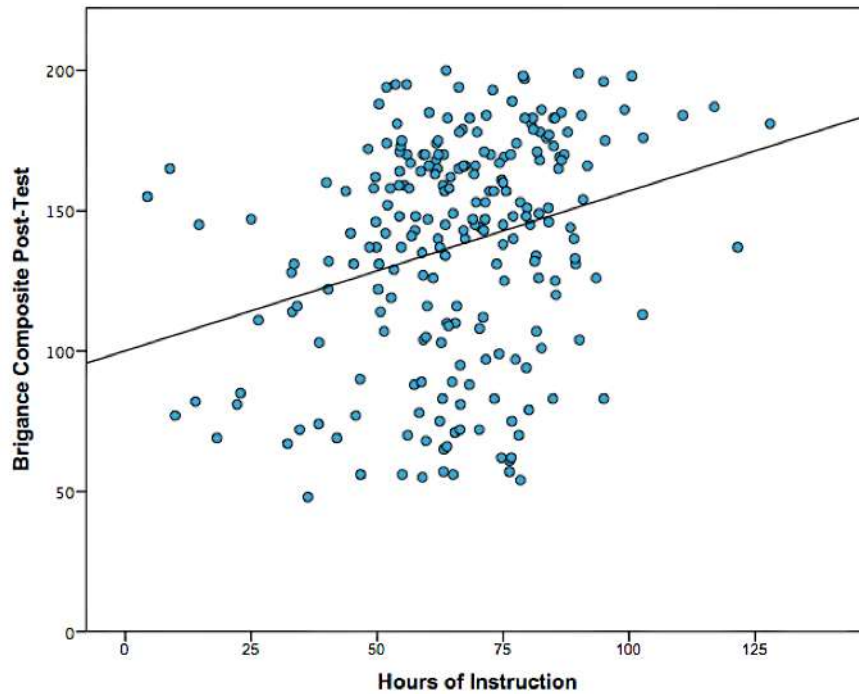
Note: Percentages may not add to 100% due to rounding. Data is from Waterford participant records.

UPSTART Usage and Literacy Outcomes

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

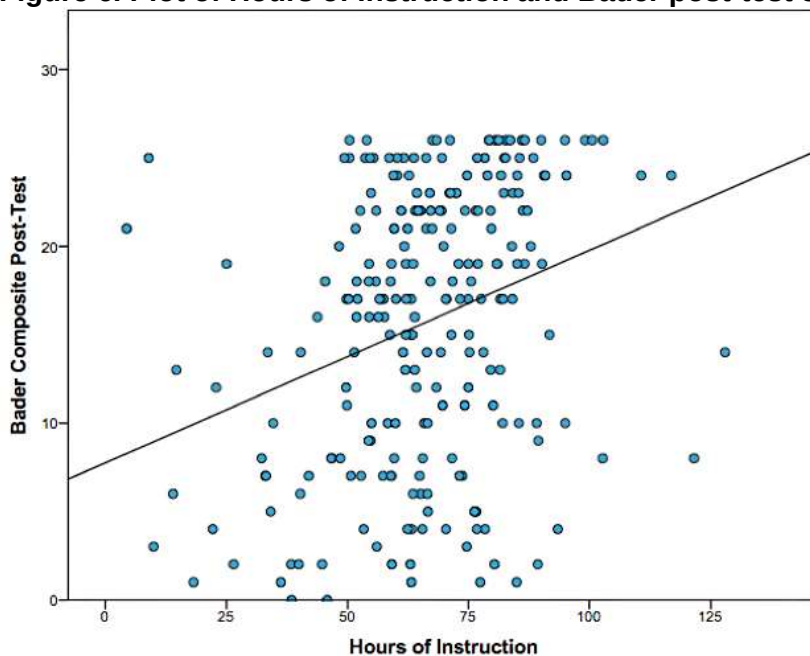
The plot in **Figure 5** on the following page shows a small positive relationship between UPSTART usage (measured in hours of instruction) and Brigance post-test scores ($r = .28$). That is, Brigance post-test scores tend to increase with increasing hours of UPSTART usage.

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



Similarly, the plot presented in **Figure 6** displays the relationship between hours of UPSTART instruction and the Bader composite post-test score indicates a small positive linear association between instruction time and scores on the Bader post-test ($r = .30$). This suggests that the acquisition of early phonological skills as measured by the Bader tend to improve with increasing levels of exposure to UPSTART curriculum.

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



UPSTART Program Impacts on Literacy

This section includes results based on statistical comparisons of literacy achievement (test scores) for matched treatment and control groups. The impact of the UPSTART program is shown through two lenses: effect sizes and growth scores. Both methods provide salient feedback about the impact of UPSTART. The first method helps stakeholders understand how large an impact UPSTART had on participants, while the second method shows how UPSTART students grew (compared to control students) based on two points of time.

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

Research Question 1: *Do UPSTART students have better literacy skills at Kindergarten than control students?*

Research Question 2: *Do UPSTART students show stronger literacy growth rates from preschool to Kindergarten than control students?*

The results of the matched sample are presented for each research question above, and the statistically significant ($p < .05$) findings are depicted visually⁶. We conducted a series of models that explored the impact of household income level on the outcomes of interest and the results were not meaningfully different from our initial analysis. 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.

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

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

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

However, it is important to note that Cohen's broad categories were designed for a range of effect sizes across a wide spectrum of social and behavioral research and are not specifically tailored for education interventions, studies, or samples. A more appropriate and meaningful benchmark for assessing the significance of an

⁶ To create a concise report that highlights the most important findings for stakeholders, we did not present findings that were non-significant in figures.

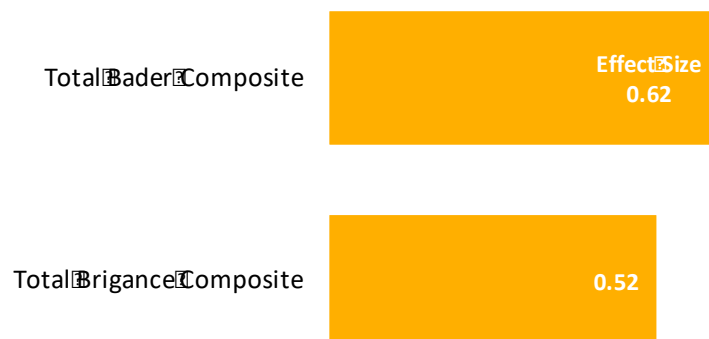
intervention's effect size is to compare it with the effects found for similar interventions with comparable research samples and outcome measures (Lipsey et. al, 2012). If an effect is larger than those of similar interventions, it has practical significance by virtue of being larger than previously reported effect sizes. Conversely, if an effect size is lower than comparable interventions and education research studies, then the impact may not as impressive or significant.

How then, do we determine appropriate benchmarks for interventions similar to UPSTART? Researchers at the Institute of Education Sciences reviewed 829 effect sizes from 124 education research studies and determined that the average effect size for an evaluation that used a standardized subject outcome measure (like the Brigance/Bader) to assess a comprehensive educational intervention program that targeted individual students like UPSTART was .26 (Lipsey et. al, 2012). We provide this benchmark to contextualize the effect sizes presented in this report and to aid the reader in determining the practical significance of the effect of UPSTART – any effect size above .26 is higher than the average effect size seen in similar education evaluations. **Appendix B** provides greater detail on how the benchmark was determined.

Effect sizes⁷ were calculated to show the magnitude of UPSTART's impact at post-test as measured by each of the 11 literacy subtests (8 Brigance subtests and 3 Bader subtests), and the Total Brigance and Bader Composites (composites include aggregated results of the subtests). **Effect sizes above the .26 benchmark are presented in gold to provide context showcase their practical significance.**

Combined post-test results showed that UPSTART participation had a medium impact on students' early literacy skill development. In the matched post-test sample⁸ (N=416), UPSTART produced medium effects (.62 and .52) as measured by the total Bader and Brigance composite scores that are well above the observed .26 effect size for similar interventions and evaluation studies (see **Figure 7**).

Figure 7. Brigance and Bader Posttest Analysis of Composite Scores



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

⁸ Treatment Group (N = 208); Control Group (N = 208)

UPSTART children scored significantly higher than control children on seven of the eight Brigance tests and all three Bader subtests on the post-test, showing strong empirical evidence that UPSTART was successful helping children develop key early literacy skills. The ES estimates for individual subtests ranged from .21 (Recites Alphabet) to .74 (Pre-primer Vocabulary) and would be considered medium to large effects. The Expressive Vocabulary subtest was the only subtest in which the treatment and control groups' post-test scores were not significantly different.

The effect size estimates for each statistically significant literacy subtest (10 out of 11), as measured by the Brigance and Bader instruments, are presented below in **Figure 8**. The results are organized according to the subtests' respective literacy constructs: decoding, phonological awareness, letter knowledge, and pre-literacy discrimination. Please refer to the **Outcome Measures** section beginning on page 14 for a discussion of the measurement constructs and **Table 3** for a list of all 11 subtests and their corresponding constructs.

Figure 8. Effect Size Estimates by Literacy Construct

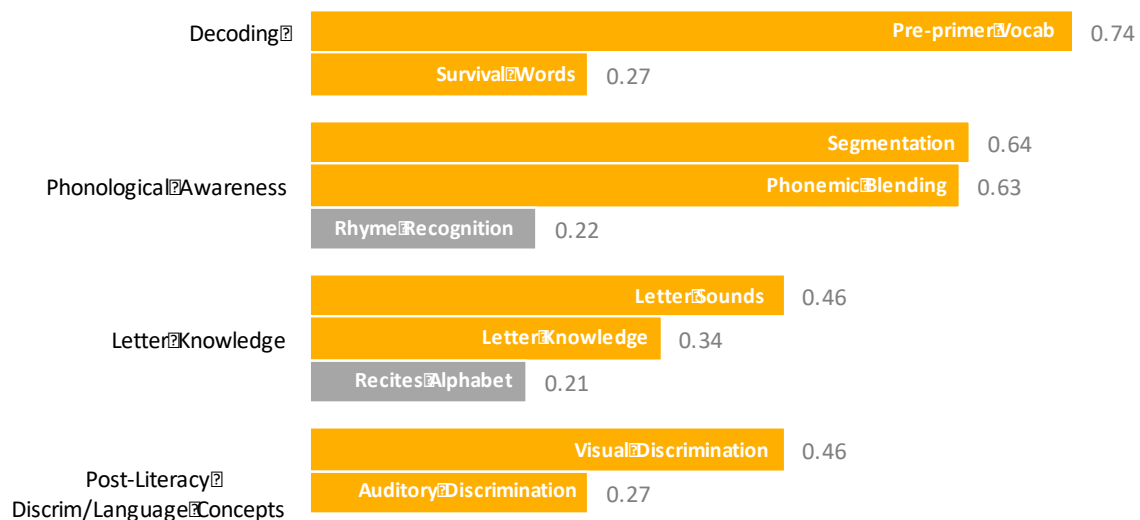
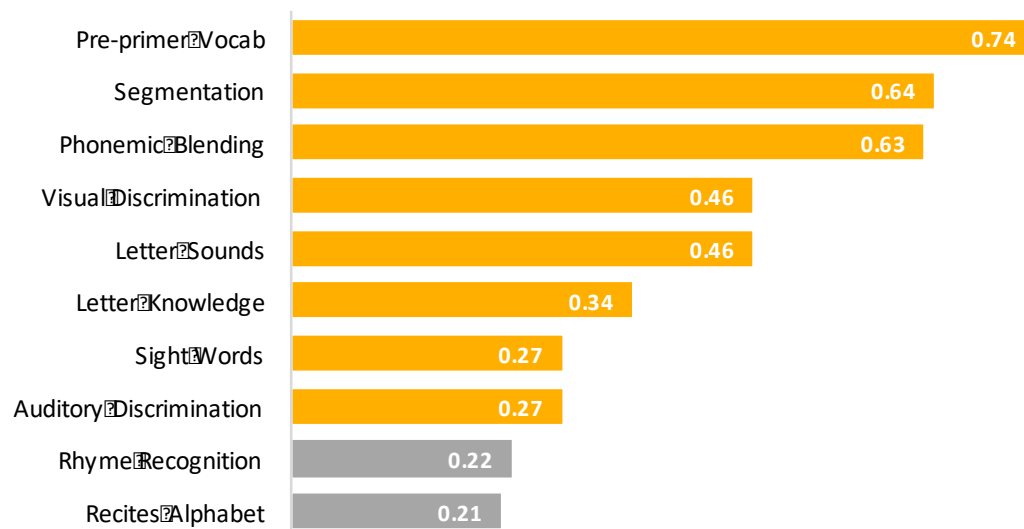


Figure 9 presents the ES of each literacy subtest by the size of their effects along with the .26 effect size benchmark from similar education intervention studies. UPSTART had the largest impact on pre-primer vocabulary (.74), phonemic segmentation (.64), and phonemic blending (.63). Effect sizes from five subtests measuring decoding, phonological awareness, phonics, and discrimination were above the average .26 effect size benchmark from other similar education interventions and should be considered practically significant and consequential.

Figure 9. Effect Size Estimates by Magnitude of Effect



Regression Results. In addition to computing effect sizes, we ran regression analyses to determine if pre-existing differences between the treatment and control groups on demographics and pre-test measures affected the results. The regression analyses did not essentially change the initial estimate of the mean overall impact on the Bader at post-test, however the linear regression analyses improved the estimate of UPSTART's overall impact on the Bader post-test from 4.69 to 5.76 points (see **Table 7**).

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

	T-Test	Regression
Bader	4.69	5.76
Brigance	20.86	21.29

Using Brigance pre-test scores as covariates did not improve the estimate of UPSTART's overall impact. However, using spouse working status (spouse working full time or spouse not working) in addition to the Brigance pre-test resulted in a slight improvement of the estimate of UPSTART's overall impact from 20.86 points to 21.29 points.

How did UPSTART Non-Graduates fare?

As noted in the **Program Implementation** section, the UPSTART graduation rate has declined over the past two years. To determine if graduation status had an impact on literacy outcomes, we conducted an exploratory analysis of our full evaluation treatment sample. Of the 230 treatment children tested, 17 did not meet UPSTART usage requirements for graduation while 213 children completed the program and graduated. The small number of non-graduates in our evaluation sample limits the generalizability of our analyses, and will require further evaluation to confirm our findings.

We found that the children who graduated from the program had higher post-test scores on the Brigance composite, as well as higher scores on the visual discrimination, auditory discrimination, survival sight words, and pre-primer vocabulary subscales than

non-graduates. Moreover, UPSTART graduates scored significantly higher than non-graduates on the Bader phonological awareness composite test and all Bader subscales (rhyming, segmenting, and blending).

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

We studied literacy growth rates while in the program as an additional way to evaluate program impacts beyond outcome score comparisons. Paired samples t-tests were performed to examine growth rates as measured by the Brigance and the Bader total test composite scores for the treatment and control group children and for eight key subtests (Rhyme Recognition, Phonemic Blending, Phonemic Segmenting, Visual Discrimination, 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 ($p < .05$) stronger mean literacy growth rates compared to the control group on the Total Bader and Brigance Composites, with the treatment group scoring an average of 6 points higher on the Bader and 21 points higher on the Brigance.
- The treatment group showed statistically stronger ($p < .05$) literacy growth rates compared to the control group on three out of five Brigance subtests (Letter Sounds, Survival Sight Words, and Basic Pre-Primer Vocabulary) and all three Bader subtests (Rhyme Recognition, Phonemic Blending, and Segmentation).
- There was no difference in growth rates between the treatment and control group on the following subtests: Visual Discrimination (measures discrimination between shapes, letters, and words) and Auditory Discrimination (measures auditory discrimination between similar word pairs).
- Of the four constructs in which the Brigance and Bader subtests measure, Pre-Discrimination was the only construct in which growth rates between the treatment and control students were not statistically significant ($p < .05$).

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

⁹ This section presents outcomes that were statistically significant, and therefore a figure for pre-literacy discrimination is not depicted here.

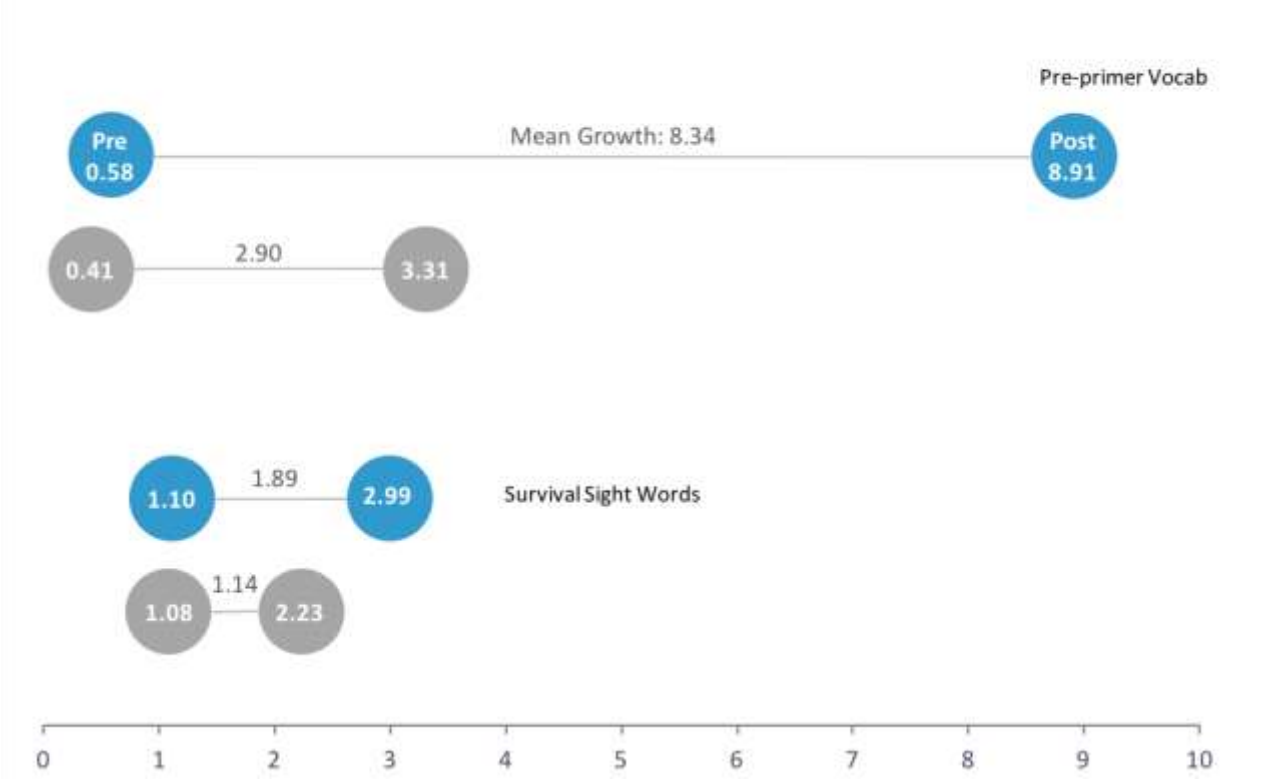
UPSTART children experienced significant, higher mean growth from pre-test to post-test compared to control children on all three subtests (rhyme recognition, phonemic blending and segmenting) that measure **Phonological Awareness**.

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



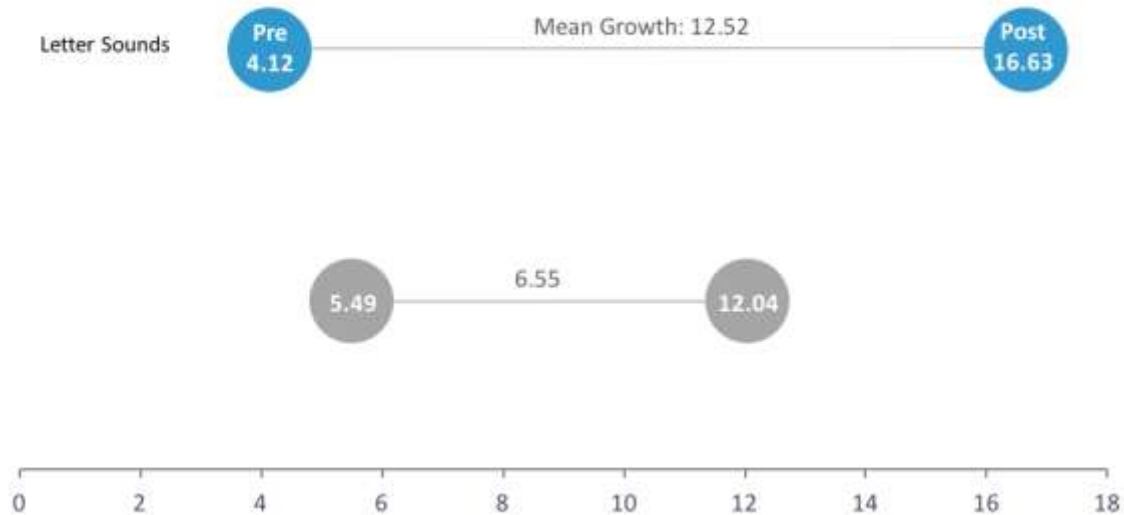
UPSTART students experienced significant, higher mean growth compared to the control group on both subtests used to measure children's **Decoding** ability, including pre-primer vocabulary and survival sight words.

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



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

**Figure 12. Letter Knowledge: *Treatment* and Control Group
Pre-and-Posttest Mean Scores**



Discussion and Recommendations

The final section of the Cohort 7 (C7) evaluation report summarizes findings and trends for UPSTART implementation and impacts on early literacy skills. Based on the results and additional discussion about the evaluation design, we also include summary recommendations for the program and future research efforts to help the state monitor its impacts.

Program Implementation

Based on the data provided by UPSTART program officers, the program was implemented with great success. UPSTART enrollment increased from 1,577 children in Year 5 to 6,639 children in Year 7, an increase of over 320 percent over the past two years. Enrollment has increased in areas across state of Utah and UPSTART has reached families in both rural and urban communities. Nearly two-thirds of the children enrolled in Year 7 lived in families with incomes less than 200% of the federal poverty level and the majority of children were White (77%) and English speaking (88%).

Most of the C7 children accessed the UPSTART curriculum through the Waterford website (76%). Approximately 14% of the sixth year participants received a computer loan and 8% were provided with a computer and Internet. Graduation rates were approximately 5% lower than in previous years, which could be due to the increased enrollment across the state. Families with children who did not graduate from UPSTART tended to have lower levels of parental education, higher levels of poverty, and be members of underrepresented racial, ethnic, and linguistic groups.

Program Impacts on Literacy Development

While program implementation findings are important for monitoring how resources were used to enroll and graduate students, findings about literacy testing outcomes is the most important indicator of program success. UPSTART participation had a strong impact on children's emerging literacy skills based on the results from effect size and growth score analyses. The program produced statistical effects (Bader ES = .62; Brigance ES = .52) on learning compared to non-program children that are stronger, on average, than other educational evaluation studies on similar interventions with comparable outcomes and participants. The effects were seen across different measures of literacy: decoding skills, letter knowledge, pre-literacy discrimination, and phonological awareness.

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

We reported findings for focused literacy tests, and a majority of the results from the Brigance and Bader scales were shown to have small to medium effects (effect sizes ranged from .21 to .74). Overall, the results of both analyses illustrate that UPSTART program participation had a strong impact on facilitating UPSTART students' literacy skill development in a variety of key areas. The largest impacts were found for pre-primer vocabulary (measures decoding skills), and phonemic segmentation and blending (measures phonological awareness). Students who did not graduate from the UPSTART program, however, did not attain the benefits of improved skills as they had lower levels of literacy achievement than students who completed the full UPSTART usage requirements.

UPSTART students also experienced greater growth from pre-test to post-test compared to control students in three out of five literacy constructs (phonological awareness, decoding, and letter knowledge), with the exception of the vocabulary and pre-literacy discrimination domains. Group differences in the vocabulary subtests and auditory/visual discrimination subtests were not statistically significant in the post-test analyses, indicating that these are skill areas in which UPSTART did not have a positive impact in Cohort 7.

Limitations

This evaluation report marks the seventh year of the UPSTART evaluation. Each year we like to discuss the implication of the evaluation results on future research efforts. We used an unmatched group (pre-/post-test) design in years past, but for Cohorts 6 and 7 we scaled-up our data collection to gather information from more students and used a matched group design. There are several benefits to balancing students using a one-to-one matching technique, but the method requires large groups of treatment and control students to find the matches, and treatment students can be removed from the analyses if they do not have an equivalent control student. Removing treatment students from our matched sample could reduce our statistical power to detect smaller treatment effects. Treatment students are randomly matched to control students with similar matching variables (see our method section for more information), but there is no way to

determine if the students who were not matched would have influenced the results since they were not included in the analyses.

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

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

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

Summary and Recommendations

The UPSTART program shows continued success at helping preschool age children develop literacy skills and prepare for entry into kindergarten. These outcomes would have specific benefits to at-risk children, whose families struggle with poverty and other issues, and often lack the resources to help their children develop the literacy skills needed to succeed in school. Given the success at improving literacy test scores, we recommend that the state continue providing the UPSTART program to children. The strong program effects support wide-scale implementation across at-risk preschool populations. This year, however, a notable trend in declining graduation rates was found and needs to be carefully monitored across future cohorts.

During the 2015-2016 program year, less C7 students were classified as graduates when compared to previous cohorts (87% graduation rate in C7 versus a 92% graduate rate in C6). In addition to the drop in graduation rates, average program usage dropped

approximately 4 hours when compared to the previous year (C7 average use was 60 hours versus 67 hours of average use for C6). The decreased program usage is of particular concern, as children who failed to meet the program requirements for graduation had, on average, significantly lower literacy outcome scores when compared to UPSTART graduates. Moreover, families that did not meet usage requirements were more likely to have other indicators of risk, such as lower levels of parental education and household incomes below the federal poverty level. In other words, there is a negative trend in enrollment and program usage for the families that the UPSTART program is designed to benefit most. This trend needs to be carefully monitored because a continued decline might begin to erode literacy outcomes for the most at-risk students.

Program Recommendations. We recommend that the program vendor work with the evaluator and USOE staff to monitor program implementation carefully and to be sure that increased enrollment does not erode graduation or usage rates, two key areas for ensuring strong student literacy achievement and future program success. 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:
 - The use of social networks and/or social media to reach parents who are difficult to communicate with.
 - Developing targeted incentives for families with the highest risk factors for not meeting program usage requirements, such as monthly awards (extrinsic), being highlighted in UPSTART communications to social networks as “Gold Star Families” (intrinsic).
- The program vendor could highlight the benefits of meeting usage guidelines by showing parents evaluation information that links graduation status with reading outcomes (e.g., through slides during training, brief informational guides, e-mails, etc.). The evaluation data can provide evidence about program effects and motivate parents to meet UPSTART usage recommendations.

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

In addition to a strong focus on finding additional sources for control student families, we recommend that USBE work with us to explore alternate evaluation designs that might reduce the need to compete with existing preschool programs for control families. One possible option to explore would be the regression discontinuity design (RDD), where children are assigned to the treatment or control condition based on a cutoff value on a variable, such as age. This design has an internal validity comparable to randomized experiments (Lipsey et al., 2015), however, it is also difficult to implement and requires careful about sampling (students) well in advance of data collection timelines.

References

- Bader, L. A., & Pearce, D. L. (2008). Bader Reading and Language Inventory (6th ed.). New York, NY: Pearson.
- Brigance, A. H. (2004). Brigance Inventory of Early Development II (IED-II) (2nd ed.). N. Billerica, MA: Curriculum Associates.
- Cohen, J. (1988) Statistical Power Analysis for the Behavioral Sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Davis-Kean, P. E. (2005). The influence of parent education and family income on child achievement: The indirect role of parental expectations and the home environment. *Journal of Family Psychology*, 19(2), 294–304.
- Evaluation and Training Institute. (2011, August). *Utah UPSTART education program evaluation kindergarten outcomes: Program impacts on reading proficiency (Cohort 1/Year 1 Results Technical Report)*. Culver City, CA: Author.
- Evaluation and Training Institute. (2012a, November). *Utah UPSTART program evaluation: Year 1/Cohort 1 First Grade (Cohort 1 First Grade Technical Report)*. Culver City, CA: Author.
- Evaluation and Training Institute. (2012b, March). *Utah UPSTART program evaluation kindergarten outcomes: Program impacts on reading proficiency (Cohort 2 Results Technical Report)*. Culver City, CA: Author.
- Evaluation and Training Institute. (2013, March). *Utah UPSTART program evaluation program impacts on early literacy: Third Year Results (Cohort 3 Technical Report)*. Culver City, CA: Author.
- Evaluation and Training Institute. (2014, February). *Utah UPSTART program evaluation program impacts on early literacy: Fourth Year Results (Cohort 4 Technical Report)*. Culver City, CA: Author.
- Evaluation and Training Institute. (2015, March). *Utah UPSTART program evaluation program impacts on early literacy: Year 5 Results (Cohort 5 Technical Report)*. Culver City, CA: Author.
- Evaluation and Training Institute. (2016, January). *Utah UPSTART program evaluation program impacts on early literacy: Year 6 Results (Cohort 6 Technical Report)*. Culver City, CA: Author.
- Guryan, J., Hurst, E., & Kearney, M. (2008). Parental education and parent time with children. *Journal of Economic Perspectives*, 22(3), 23-46.
- 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.
- 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>
- Shadish, Cook, and Campbell (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton Mifflin Company.
- Snow, C.E., Burns, M., S., & Griffin, P. (1998). *Preventing Reading Difficulties in Young Children*. Washington, DC: National Academy Press.

Appendix A: Comparison of Evaluation Samples

The matched and unmatched samples are compared with the C7 population on key demographic characteristic reported by the program vendor in **Table A.1**. Both of the unmatched and matched samples are more homogenous than the full population of preschoolers who were enrolled in Cohort 7, with 80% and 83% of unmatched and matched children, respectively, being Caucasian and 100% classified as English speakers.¹⁰

Table A.1
Sample Comparisons on Key Waterford Demographics

Demographic Categories		C7 Population (N = 6,639)	Unmatched Sample (N=230)	Matched Sample (N=208)
Gender	Male	51%	51%	51%
	Female	49%	49%	49%
Ethnicity	Caucasian	77%	80%	83%
	Hispanic	16%	11%	10%
Child Language	English	88%	100%	100%
Parent Education Level	Some College	38%	46%	47%
	Bachelor's Degree	35%	30%	30%
Parent Marital Status	Married	91%	90%	90%
Poverty Status	Under 185%	59%	100%	100%

The C7 population had parents with slightly higher college graduation levels and lower levels of poverty. Whereas 35% of the parents in the overall C7 population have a college degree, the modal level of parent education in the unmatched and matched sample was some college (46% and 47%, respectively). Additionally, 59% of families in the C7 sample were under the 185% federal poverty rate compared to 100% of families in the unmatched sample and 100% of families in the matched sample. As mentioned in the main body of the report, we focused on recruiting low-income families for our treatment sample to reflect the prioritization of these families by the state in the recent legislative extension of the UPSTART program.

The unmatched sample is slightly closer to representing the characteristics of the C6 population. However, the matched sample ensures that the treatment group's characteristics best mirror the control group to estimate program impact with the greatest accuracy. UPSTART outcome findings are reported in the main body of the report from the matched treatment-control sample.

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

Appendix B: Determining UPSTART Effect Size Benchmark

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

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

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