Indiana UPSTART Evaluation

Preschool Impact Study

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

The Evaluation and Training Institute (ETI) was contracted by the Waterford Research Institute to evaluate the UPSTART program, an at-home school readiness software program developed and provided by the Waterford Institute to prepare young children for school entry and future academic success. The evaluation consists of two complementary quasi-experimental evaluation designs with matched treatment and control groups that will be used to determine the program's impact on participants' literacy and numeracy skills as well as their social-emotional development: a **Preschool Study** and a **Third Grade Study**. This report is focused on presenting findings for the Preschool Study because data for the Third Grade Study will not be collected until the fifth year of the program (2023), at which point ETI will collaborate with the state and Local Education Agencies (LEAs) to have student state standardized test scores transferred to ETI for analysis.

The Preschool Study involved the testing of preschool students enrolled in the UPSTART program (treatment group) and students who were not enrolled in the program (control group) before and after the program was implemented. The treatment and control groups were pre-tested from October 2018 to January 2019 and post-tested from July to August 2019. Findings for the Preschool Study were presented as follows:

- The *Preschool Impact Analysis* presents information on program student outcomes using a pre-test/post-test design with a statistically matched control group in order to assess the program's impact on developing children's early literacy skills. Our research findings cover three areas:
    1) What types of impacts the program had on children's literacy skills,
    2) What types of impacts the program had on children's math skills, and
    3) What types of impacts the program had on children's social skills.
- The *Usage Analysis* describes how the program was implemented and used throughout the preschool year.

The executive summary presents key findings for each reporting area, along with selected recommendations for improving the program for future evaluation efforts.

Preschool Findings - Impacts on Participants' Literacy, Math, and Social Skills

We presented 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 presents it in standard deviation units. Effect sizes describe the magnitude of the difference between two groups, and essentially create a standardized scale to facilitate interpretation of results. Following recommendations from the What Works Clearinghouse (WWC) (What Works Clearinghouse, 2017) and a meta-analysis of similar educational interventions and studies (Lipsey et al., 2012), we set an effect size threshold of .26 to denote effects that have practical significance and are substantively important.

The UPSTART program had a significant effect on children's literacy skills based on results from the regression model and effects size estimates. Children enrolled in UPSTART produced statistically significant positive effects compared to control children on the *Brigance Literacy Composite scale*, which measures phonological awareness, letter knowledge, and pre-literacy discrimination (ES = .48). However, UPSTART participants did not perform significantly better on the Brigance Math composite (ES=0.09). Social skill development was not negatively or positively affected by the UPSTART program. In other words, the home-based, computer program did not hinder children's social skill development, with results showing similar age-appropriate social skill growth to control subjects from pre to post tests.

Students who participated in UPSTART experienced a large improvement in their *letter knowledge skills*. The letter is the most basic unit of reading and familiarity with the letters of the alphabet has been shown to be a strong predictor of reading achievement (Piasta & Wagner, 2010). Additionally, understanding the connection between written letters and the sounds of speech is a precursor to decoding.

- UPSTART children had medium to large effects in their word recognition (ES = .86), sign recognition (ES = .43), and recites alphabet (ES = .46).

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 visually discriminate between different shapes, letters, and words (ES = .29), and even stronger effects in their ability to audibly distinguish if two words sound the same (ES = .47).

Phonological skills have been identified as one of the most important predictors of reading success and involve 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.

- There were no statistically significant differences found between the treatment and control children for the two subscales measuring children's phonological skills: phonological awareness or phonological manipulation.

Summary and Recommendations

The UPSTART program findings highlighted its success in helping preschool-aged children develop their literacy skills as they prepared for entry into kindergarten, however the program had no effect on children's math. Children in both treatment and control groups had positive, age appropriate social skill development during the course of the study. It is possible the current study under-represented the treatment effects considering the low sample size of participants tested, which is a limitation of our evaluation. Smaller sample sizes tend to have decreased power and type II error or a false negative. In addition, UPSTART program usage, as identified by graduation rates, was considered low in our sample, and was approximately 56%. UPSTART usage was positively correlated with increased literacy and math scores, and stronger effects than what we saw in this report may have been possible with increased program usage and resultant graduation rates.

A second limitation, the research design was in conjunction with approval from the Indiana Family and Social Service Agency (FSSA), which claimed a design oversight mandate but failed to provide a field endorsement. The lack of a field endorsement and a basic linking of state agencies to the independent evaluator contributed to higher attrition rates in the control group than in cases where state agencies simply endorse and support the evaluation activities. These actions also conspire to under-represent possible treatment effects. Low sample size and high attrition rates negatively the study and ultimately the state's ability to review the program's full impact, as detailed in our section on study limitations.

Even with the stated limitations, due to the strong impact on early literacy development we recommend that the state continue to provide the UPSTART program to children. Given the importance of UPSTART program graduation on literacy achievement outcomes, we recommend that the program vendor continue to work with the evaluator and the Indiana Family and Social Services Administration to monitor program implementation carefully and encourage higher graduation rates. Specifically, we recommend that the program vendor consider the following:

- 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). Some potential strategies might include:
    - Creating peer support partnerships with similar community groups in high risk geographic locations to discuss strategies for increasing children's program use.
    - Increasing communication with high-risk families to evaluate potential barriers to program usage.
    - Developing targeted incentives for families with the highest risk factors for not meeting program usage requirements, such as monthly awards (extrinsic), being highlighted in UPSTART communications to social networks as "Gold Star Families" (intrinsic).

Introduction

The State of Indiana requested proposals from education technology providers to develop school readiness programs for preschool children (RFI-18-078). The Waterford Research Institute was awarded a contract to implement its UPSTART program, an at-home adaptive educational technology program, to Indiana preschoolers. As part of the project, the state required an independent firm to conduct an evaluation of the program's effectiveness for preparing children for success in school. The Family and Social Services Administration (FSSA) was tasked with oversight of the project and the evaluation to ensure that both comply with the legislation supporting the initiative. The Waterford Research Institute contracted with the Evaluation and Training Institute (ETI) to conduct an independent evaluation of the UPSTART program for Indiana children that compares the school readiness of children who enrolled in UPSTART during their pre-kindergarten year with a group of comparison children who did not participate in the program. The FSSA reviewed and approved the evaluation design and scope of work.

UPSTART Program Description

UPSTART is a home-based software program created by the Waterford Research Institute that was designed to promote the development of literacy skills[1] that will prepare young children for entry into school ("school readiness"), as well as to provide reading instruction to students up to third grade. The program provides an individualized learning experienced by adapting to each child's skill level as they move along the program completing assignments and tasks. The reading skills covered in Level 1 of the Waterford Early Learning Program include:

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

For preschool-aged children the recommended program usage was **15 minutes a day, 5 days a week**, with an expectation that the child would complete 1,000 minutes prior to entering kindergarten.

Methods

Study Design

The Indiana UPSTART evaluation used a quasi-experimental, repeated measures research design with naturally occurring groups of students in treatment (program) or control (no program) conditions to determine how the UPSTART program impacted their literacy and math skills and their social-emotional development. Program students were compared to a comparison group of peers who did not use the program. Data were collected at entrance to preschool (pre-test) and exit from preschool (post-test). We plan on testing students in the third grade (delayed posttest),

[1] The UPSTART program was designed to promote reading (literacy), but also had curriculum developed to support Math and Science learning.

but we will also work with our district partners and explore the potential for using state testing data alongside of the data we have already collected. This would allow us to present to Indiana a comprehensive research report that combines state testing data, such as the successor to the ISTEP (ILEARN; expected to be launched in 2019), the IREAD-3, and our own primary data from widely accepted measures of math and literacy.

The diagram below depicts the study design. The UPSTART/treatment children received the Waterford Early Learning Program in the year prior to kindergarten, whereas the control group children did not. The measurements taken the summer before preschool enrollment, Measurement 1 (pretest), summer after preschool and just before enrollment in kindergarten, Measurement 2 (posttest), and a delayed posttest taken during third grade, Measurement 3.

			Preschool (2018-2019)		**Kindergarden-2nd grade**	**3rd Grade (2023)**
Not Randomly Assigned	Treatment	Measurement 1	UPSTART	Measurement 2		Measurement 3
	Control					

Research Questions

We hypothesized that if the UPSTART home-based program had no effect on improving early literacy skills or early math skills, then the children who participated in UPSTART (treatment group) would be expected to perform at the same level as nonparticipants (control group) following the preschool year. Alternatively, if UPSTART did have an effect on these skills, then the treatment group should perform significantly better than the control group on the first post-test taken prior to enrollment in Kindergarten. In addition, we tested if the UPSTART program had an impact on children's socials skills, in that we hypothesize that children who participated in UPSTART (treatment group) would be expected to perform at the same level as nonparticipants (control group) following the preschool year.

Our preschool study research questions were as follows:

- **Did four-year-old children who participated in UPSTART during their preschool year have higher scores on measures of literacy and math before entry into kindergarten than four-year children who did not participate in the program?**
- **Did prekindergarten children who participated in UPSTART have social skills (e.g., communication, cooperation, empathy) before entry into kindergarten that were similar to children who did not enroll in the program?**

Forecasting our follow-up study research questions, if UPSTART has no effect on sustaining gains in early literacy development, then the third-grade children who participated in UPSTART – the treatment group – would be expected to perform at the same level as the control group when given a delayed post-test in the third grade. Alternatively, if UPSTART has an effect on sustaining literacy gains through third grade, then the treatment group should perform significantly better than the control group when re-tested during the third grade.

The longitudinal research questions were:

- **Did UPSTART sustain improvements in literacy skills through grade three?**
- **Did UPSTART sustain improvements in math skills through grade three?**
- **Did UPSTART sustain social skills through grade three?**

Data Collection

Data collection and assessments of preschool children occurred prior to school entry during the summers of 2018 and 2019. The outcomes of interest were measures of early literacy skills relevant to emerging readers (phonemic awareness, letter recognition, awareness of concepts of print, and oral language comprehension), early math skills (counting, sequencing, patterns, shapes, and others) and social emotional learning (communication, cooperation, Assertion, responsibility, empathy, and engagement). The first two skill set measures align with the skills taught by the Waterford Early Learning Program at Level 1 of the curriculum. To measure the literacy skills of pre-literate children, we used the Brigance IED III - Academic Skills/Cognitive Development: Literacy scale. We also used subscales from the Brigance IED III - Academic Skills/Cognitive Development: Mathematics scale to measure math skills. Preschoolers' social and emotional development were assessed with the parent survey of the Social Skills Improvement System (SSIS) rating scale. A detailed description of each of these outcome measures is provided in the **Instrumentation** section below.

Measures

The outcomes of interest in the Indiana UPSTART evaluation included:

- Early literacy skills relevant to emerging readers - early phonemic awareness, letter recognition, vocabulary, and decoding.
- Emerging math skills - number sequencing, comparing amounts, recognizing written numerals, counting, and basic arithmetic.
- Social and emotional development – communication, cooperation, assertion, responsibility, empathy, engagement, and self-control.

We collected data using two standardized instruments, the Brigance Inventory of Early Development (3rd edition) (IED-III) and the Social Skills Improvement System (parent rating scale).

The **Brigance Inventory of Early Development** (IED-III) literacy scale measured children's letter knowledge, phonological awareness, print conventions, and word recognition. The Brigance mathematics scale focused on the development and understanding of basic mathematical concepts such as rote counting, number-object correspondence, comparisons of groups, knowledge of numerals, problem solving, and basic computation.

The Brigance IED III was standardized on a nationally representative sample in terms of geographic, demographic, and socioeconomic characteristics and psychometric studies show evidence of strong reliability and validity. With regard to reliability, children's scores were consistent when examined repeatedly (test-retest reliability for preschool children = .93), consistent across multiple examiners (inter-rater reliability ranged from .73 to .94), and test items were correlated with one another (internal consistency reliability from .90 to .99 with a mean of .96). Inferences based on the Brigance were valid and supported by justifying evidence as confirmatory factor analysis showed a high degree of construct validity or internal structure between Brigance subtests, and the Brigance correlates with other achievement tests such as the Woodcock Johnson, Wechsler Intelligence Test, and the Battelle Development Inventory.

The **Social Skills Improvement System** (SSIS) rating scales evaluated social skills (e.g., communication, cooperation, assertion, responsibility, self-control) and problem behaviors (e.g., bullying, externalizing, internalizing, hyperactivity/inattention). The instrument has been standardized on a nationwide sample matched to the US population estimates for race, region, and socioeconomic status and allows for the determination of national norms for preschoolers and by gender. Psychometric properties indicate high median scale reliabilities (mid to upper .90s) for all scale subdomains (social skills, problem behaviors, academic competence) and for all age groups (3-5 year old children, 5-12 year old children, and 13-18 year old children). Test-retest reliability statistics ranged from .83 to .87 for the parent rating scale.

Data Collection Procedures

The evaluation used pre-kindergarten assessment data collected by field staff at the following time points: (1) before the program (pretest, Summer 2018) and (2) at exit from the program (posttest, Summer 2019; serves as entrance into Kindergarten data).

Assessments were administered individually to children by trained staff. Assessors had undergone criminal background checks and been thoroughly vetted before being hired. In addition, staff signed a confidentiality agreement to assure that no experimental data or results from the assessment were shared outside of the research team.

The study included repeated measures administered over time by staff trained in the research protocol. Assessment sessions took approximately 30 minutes and were held in testing centers central to each district (e.g., elementary schools, libraries, conference meeting rooms). During the assessment, field staff conducted the following activities:

1. Informed consent procedure and signature form
2. Parent Surveys
   a. Intake Form (demographic information)
   b. Social Skills Improvement System (SSIS) parent rating scale
3. Brigance IED III Literacy
4. Brigance IED III Math

Prior to beginning the assessment, the testing administrator provided parents with informed consent information. Assessors answered any questions posed by parents and cosigned the consent form to indicate that they followed proper informed consent procedures. Signed consent forms were required for participation and kept in secure storage along with completed testing materials for the duration of the study. Parents and children were advised that they could stop participating in the study at any time. Parents were given a copy of the informed consent, which included contact information for the ETI project director, Dr. Jon Hobbs.

Parents also completed an intake survey that asked them to provide some basic demographic information (annual income, employment status, marital status, etc.) along with some questions that asked about their child's social habits while their children were tested on measures of early literacy and math.

Evaluation Sample

The Waterford Research Institute transferred a list of UPSTART treatment families to the principal investigator at ETI. Using the provided data, ETI performed the following tasks: (1) generated a unique identification number for each family, and (2) randomly selected a subsample of eligible families to serve as potential treatment families for the evaluation.

Waterford also securely transferred a list of waitlisted families who would not be participating in UPSTART, due to size constraints, to ETI to serve as potential control families. We also recruited control families through local preschools and announcements in community organizations (e.g., libraries, local newspapers, parent groups).

Our sample was drawn from four Indiana school districts who agreed to participate in the UPSTART program. Our total sample consisted of 170 students, 66 program students and 104 comparison students. Some basic demographic characteristics of the Indiana population are presented below in **Table 1**, which shows that a majority of evaluation participants identified as White and had parents who were married, had an annual household income between $10,000-74,999, and had completed some college or less.

Table 1. Pretest Treatment-Control Comparisons on Key Demographics

Demographic Categories		Treatment (N=66)	Control (N=104)
Child Gender	Male	45%	52%
	Female	55%	48%
Child Ethnicity	Hispanic	9%	0%
	African American	2%	11%
	White	85%	87%
Child Language	English	98%	100%
Parent Education Level	HS Diploma	44%	40%
	Some College	36%	33%
	Bachelor's Degree	8%	16%
Parent Marital Status	Married	58%	57%
Household Income	$0-$10,000	6%	4%
	$10,000-$24,999	35%	14%
	$25,000-$49,999	40%	44%
	$50,000-$74,999	11%	20%
	$75,000-$99,999	2%	6%

$100k or more	2%	6%

Table 2 displays the number of participants that were tested, and the number of participants used in the analysis. The majority of participants that were excluded from the final sample did not complete post -testing and therefore were removed from the final analysis due to not having outcome test scores. Control families were more likely to drop-out of the study at post-test than treatment (23% vs. 30% attrition rates).

Table 2. Participant Counts and Attrition

Experimental group	**Number Tested**	**Number in Analysis**	**Attrition**	**Differential Attrition**
UPSTART	66	51	23%	7%
Control	104	73	30%	
Total	170	124	27%	

**Note.* Some children moved from one experimental group to another from pre-to-posttest.

Our final sample of students used in the evaluation analysis (**Table 3**) includes 51 treatment and 73 control students. Our demographics changed due to post-test attrition, and the final sample control group had higher household incomes and pretest scores. These two qualities put the control students at an advantage over the lower-income and lower scoring (pre-test) treatment students, however, these covariates were controlled for in the final statistical models (along with others, such as ethnicity and language status).

Table 3. Final Sample Treatment-Control Comparisons on Key Demographics

Demographic Categories		**Treatment (N=51)**	**Control (N=73)**
Gender	Male	49%	45%
	Female	51%	55%
Ethnicity	Hispanic	12%	0%
	African American	2%	11%
	White	80%	86%
Child Language	English	98%	100%
Parent Education Level	HS Diploma	42%	36%
	Some College	34%	37%
	Bachelor's Degree	11%	36%
Parent Marital Status	Married	61%	59%
Household Income	$0-$10,000	6%	4%
	$10,000-$24,999	31%	10%
	$25,000-$49,999	41%	42%
	$50,000-$74,999	14%	21%
	$75,000-$99,999	2%	7%
	$100k or more	2%	8%
Skills	Brigance Literacy Composite (Pre-Test)	52.98	73.78
	Brigance Math composite (Pre-Test)	13.82	19.82

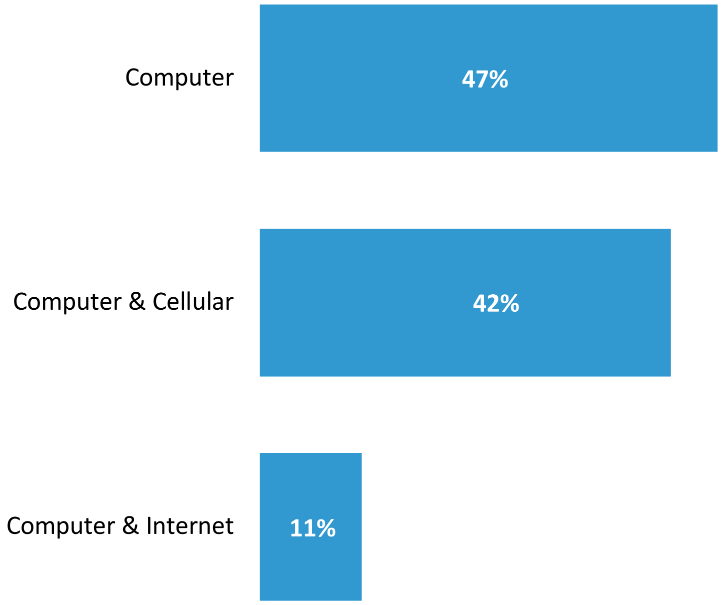
Program Implementation Findings

Findings reviewed in the UPSTART implementation section included 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.

Provided UPSTART Equipment

The type of education technology provided to UPSTART children is shown in **Figure 2** for all 66 children. The majority of UPSTART children (47%) used the UPSTART provided free personal computers, which allowed families to access the UPSTART curriculum through these provided computers. The remaining families were provided with either the computer and cellular equipment (42%) or the computer and internet installation (11%)

Figure 1. Equipment Provided to UPSTART Participants by Waterford

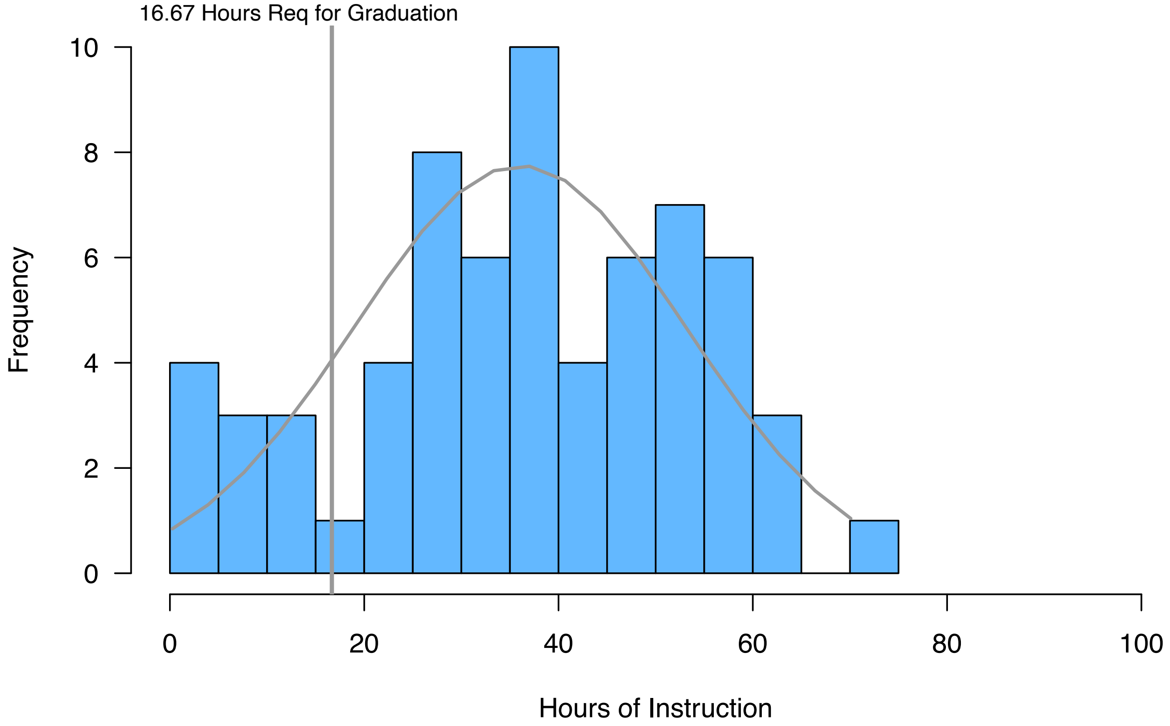


UPSTART Usage

All of the 66 enrolled families who were provided instructional equipment (e.g., computers, an Internet subscription, and cellular data) logged at least 1 hour of instructional time in the UPSTART curriculum. For these enrolled families who used the curriculum, the average duration in the program was approximately 29 weeks with approximately 36 hours of instruction on average (see **Figure 3**).

The bottom quartile of the UPSTART population completed 26.76 hours of instruction or less, the midpoint of the UPSTART distribution was 36.29 hours, and the top quartile completed an excess of 50.55 hours of instruction.

Figure 2. Histogram of Hours of Instruction for UPSTART Participants (N=66)



UPSTART Graduation Rate

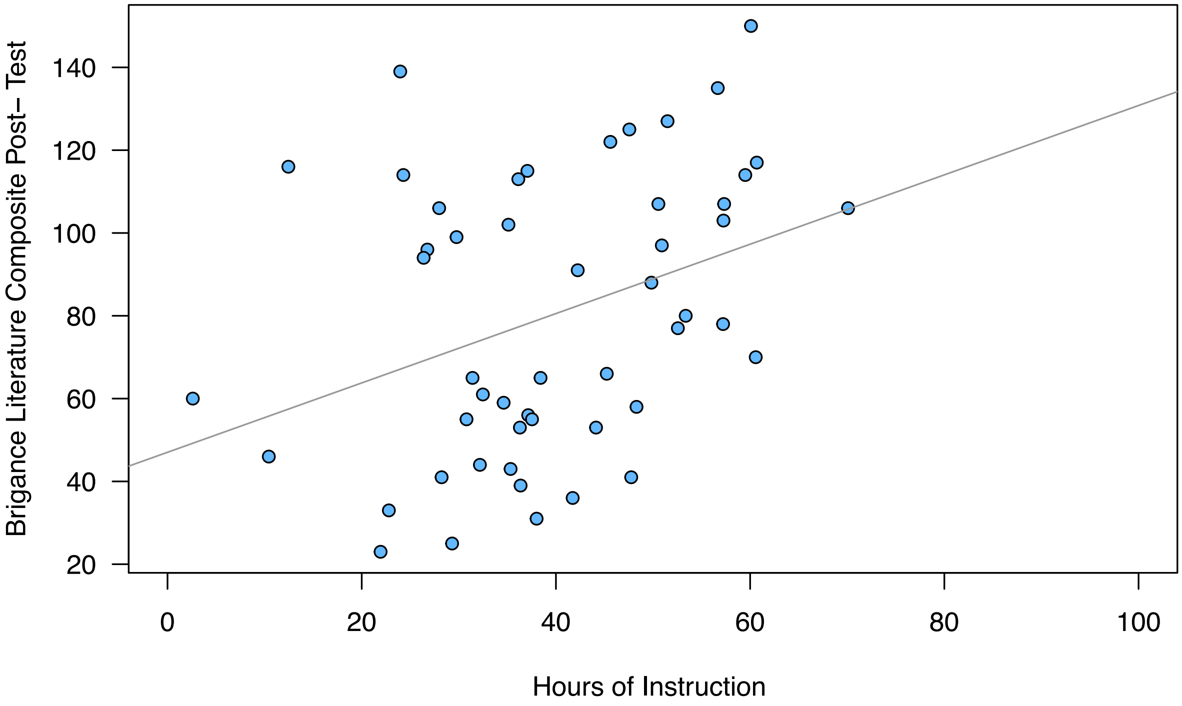
Of the 66 children documented as enrolled in the UPSTART program, the Waterford Institute classified 37 as children who had met the program's usage criteria and were thus considered to be graduates of the program (56%). The usage criteria involved (a) logging more than 1,000 minutes (16.67 hours of instruction) with the UPSTART curriculum and (b) averaging at least one hour of instruction per week while participating in the program. UPSTART graduates tended to have significantly more hours of instruction (Graduate mean: 47.73, Nongraduate mean: 21.09) and number of weeks in the program (Graduate mean: 34.22, Nongraduate mean: 26.66).

UPSTART Usage and Outcomes

UPSTART curriculum usage was found to be significantly and positively related to literacy and math outcomes as measured by composite scores on the Brigance literacy and math measurements.

*The plot in **Figure 4** on the following page shows a small positive relationship between UPSTART usage (measured in hours of instruction) and Brigance literacy post-test scores (r=.38). That is, Brigance literacy post-test scores tended to increase with increasing hours of UPSTART usage.*

Figure 4. Plot of Hours of Instruction and Brigance Literature Post-test scores



Similarly, a correlation analysis of the relationship between hours of UPSTART instruction and Brigance math composite post-test scores indicated a positive linear association between instruction time and scores on the Brigance Math post-test (r = .31). This suggests that the acquisition of early mathematical skills as measured by the Brigance also tended to improve with increasing levels of exposure to UPSTART curriculum.

Program Impact Findings: Literacy and Math

Description of Impact Analysis

This section included results based on statistical comparisons of literacy and math achievement (test scores) for treatment and control groups during the first year of UPSTART implementation. The impact of the UPSTART program is shown through two lenses: effect sizes and regression scores. Both methods provided salient feedback about the impact of UPSTART. The first method helps stakeholders understand how large an impact UPSTART had on participants, while the regression will take into account whether any preexisting factors had an effect on testing outcomes.

Results in this section are based on series of ordinary least squares regression models (OLS regression) that explored the impact of the treatment (UPSTART) on the outcomes while controlling for other meaningful predictor variables that could affect students' posttest scores, such as household income level, ethnicity, and parent's education. 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. The simple regression model included a full factorial combination of Brigance Literacy composite, testing period, and treatment group.

In addition to OLS regression, we have presented effect sizes provide additional context for our findings. An effect size (ES) takes the difference between two group means on an outcome variable and represents it in standard deviation units. For example, an effect size of .30 would indicate that the difference between a treatment and control group is .30 standard deviation units. Effect sizes describe the magnitude of the difference between two groups, and essentially create a standardized scale so the results are easy to interpret and have meaning. According to Cohen's (1988) general categorization of effect sizes as small (0.2), medium (0.5), and large (0.8) as a general rule of thumb. However, it is important to note that Cohen's broad categories were designed for a range of effect sizes across a wide spectrum of social and behavioral research and are not specifically tailored for education interventions, studies, or samples. A more appropriate and meaningful benchmark for assessing the significance of an intervention's effect size is to compare it with the effects found for similar education interventions with comparable research samples and outcome measures (Lipsey et al., 2012). If an effect is larger than those of similar interventions, it has practical significance by virtue of being larger than previously reported effect sizes. Conversely, if an effect size is lower than comparable interventions and education research studies, then the impact may not be as impressive or significant.

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

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

Effect sizes[2] were calculated to show the magnitude of UPSTART's impact at post-test as measured by each of the eight literacy subtests and the Total Brigance Literacy and Math Composites (composites include aggregated results of the subtests, (8 Brigance Literacy subtests and 3 Brigance Math subtests). **Graphs of effect sizes in this report provide a line marking**

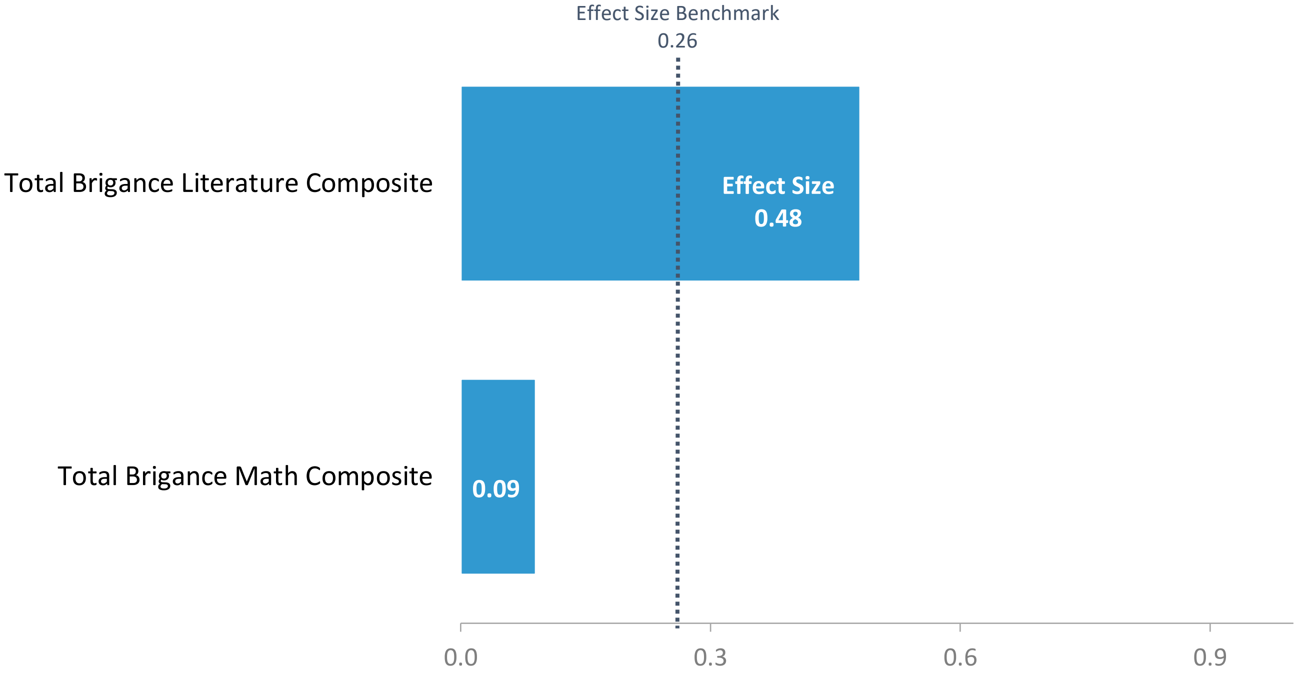
[2] Effect size (Cohen's *d*) was calculated for each test as the treatment group mean minus the control group mean divided by the pooled standard deviation.

the .26 benchmark to provide context and to showcase findings that have practical significance.

Program Effects on Literacy
Combined post-test results showed that UPSTART participation had a substantive impact on students' early literacy skill development, but not mathematical skill development. In the matched post-test sample for the Brigance Composite Scores UPSTART produced medium effects for literature skills (.48), while there was a negligible effect for mathematical skills (0.09) as measured by the total Brigance literacy and math composite scores that are well above the observed .26 effect size for similar interventions and evaluation studies (see **Figure 5**).

Figure 5. Brigance Post-test Analysis of Composite Scores

On average, children participating in UPSTART scored 52.98 points on the Brigance Literacy Composite before beginning the program and 80.31 points on the Brigance after the program was completed. Conversely, control children who were not enrolled in UPSTART scored 73.78 points on the Brigance pre-test and 87.03 points on the Brigance post-test.



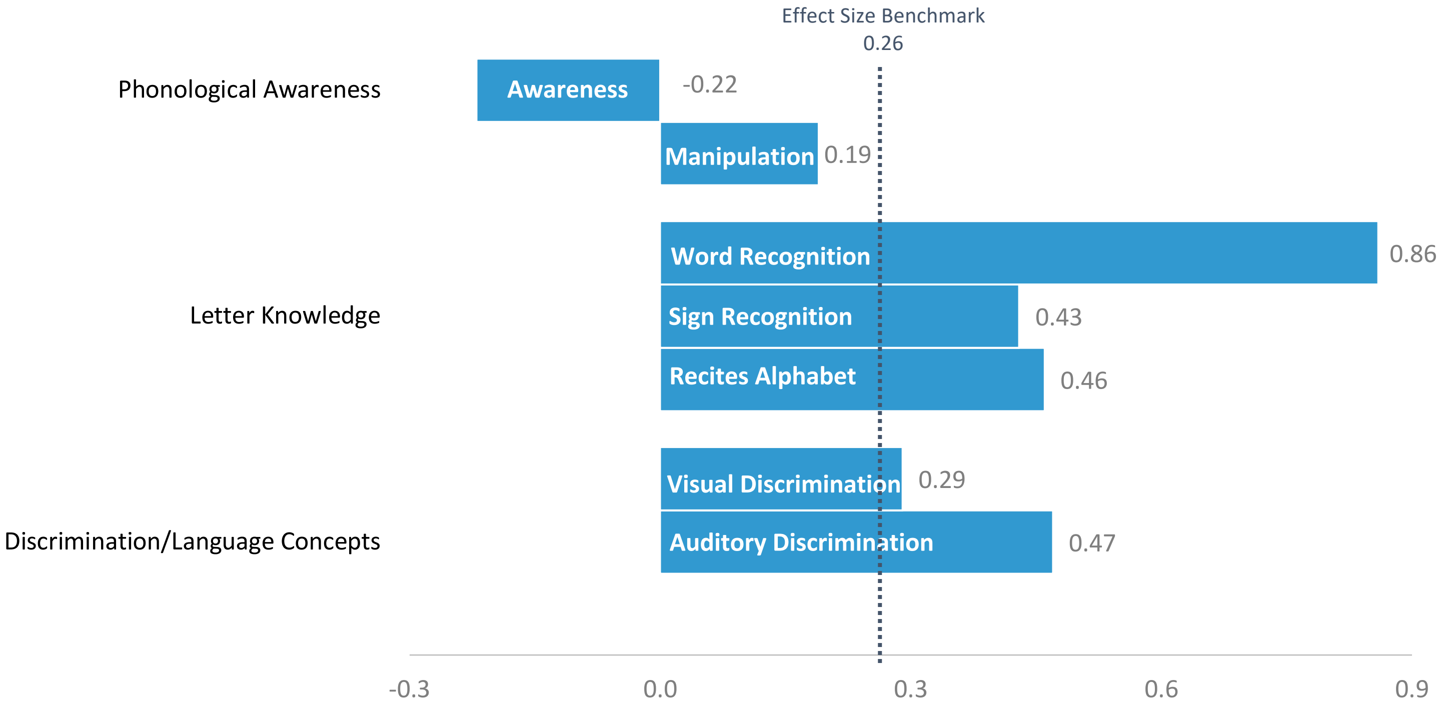
With regard to the Brigance Math Composite, UPSTART children scored 13.82 points on the instrument at pre-test and 24.31 points at post-test, while their control counterparts scored 19.82 points on the Brigance Math Composite pre-test and 28.16 points on the Brigance Math post-test.

UPSTART children scored significantly higher than control children on all seven out of the eight Brigance Literacy post-tests, showing strong empirical evidence that UPSTART was successful in helping children develop key early literacy skills. The results are organized according to the subtests' respective literacy constructs: phonological awareness, letter knowledge, and pre-literacy discrimination. The ES estimates for individual subtests on the Brigance Literacy ranged from .86 (word recognition) to -.22 (Phonological Awareness) with some effects considered medium to large effects, while other small to negligible.

Figure 6. Effect Size Estimates by Literacy Construct

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 influenced the results.

Simple linear regression analysis using only the treatment as a predictor produced an estimated effect of 15.09 on the Brigance Literacy post-test, indicating that participant in the UPSTART results in higher literacy score. Using pretest as a covariate slightly improved the estimate of UPSTART's impact and the linear combination of the treatment and the Brigance composite pretest was significantly related to the Brigance posttest, $R^2 = .66$, adjusted $R^2 = .71$, $F(18, 97) = 13.51$, $p < .0001$. and accounted for 66% of the explained variability in posttest outcomes. The full OLS regression model can be seen in Appendix A.



Social Skill Results

For the purposes of this study, ETI utilized the SSIS Rating Scales to measure social skills recorded in the parent form and the benchmark scores can be referenced in **Appendix C**. The central question that we asked was, "Do treatment and control students have equal levels of social skills development across their preschool year?" The answer would aid in understanding the impact of an at-home computer-based school readiness program on the social and emotional development of students using the program, with one concern being the students may not interact with other children as frequently as students enrolled in a traditional preschool setting. ETI hypothesized that if the program had no negative effects, differences between the treatment and control group would be minimal and that both groups would, generally, experience similar rates of social skills development.

In general, both treatment and control students developed social skills at about the same rates. There were no statistically significant differences between treatment and control group social skill outcomes while controlling for baseline social skill scale scores.

To help stakeholders review descriptive findings about social skill development, the following section is organized by the seven different social skill subscales and ends with an analysis of the composite scores across all subscale categories:

- Communication
- Cooperation
- Assertion
- Responsibility
- Empathy
- Engagement
- Self-Control
- Composite

Communication

This subscale (score range 0-21) measures a child's ability to take turns and make eye contact during a conversation, using appropriate tone of voice and gestures, and being polite by saying "thank you" and "please".

As seen in **Table 4**, the treatment and control groups tended to perform about the same across most behavior levels from pre- to post-test, with a few exceptions. The most notable difference between the treatment and control groups is evident in the nine percent increase of treatment students considered above average at post-test. In comparison, the percentage of control students in that same behavior level remained at 15% for both pre and post-test. Although treatment students experienced the largest growth in students considered above average at post-test, many control students were already considered above average at pre-test and remained so at post-test. The percent of average treatment students decreased by ten percentage points from 88% at pre-test to only 78% at post-test and the percent of below average treatment students experienced no change from pre to post-test, remaining at eight percent. The pattern seen among treatment students between pre and post-test indicates that slightly more treatment students are improving their level of communication at post-test than control students who tend to experience little to no growth from pre- to post-test.

Table 4. Level of Communication among Treatment and Control Students at Pre- and Post-Test

	Below Average		Average		Above Average	
	TX	CTRL	TX	CTRL	TX	CTRL
PRE	5	4	58	84	3	16
	8%	4%	88%	81%	5%	15%
POST	4	2	40	60	7	11
	8%	3%	78%	82%	14%	15%

Cooperation (Score Range: 0 – 18)
This subscale measures a child's propensity to help others, share materials, and comply with rules and directions.

The percentage of treatment students remained consistent across all three levels of cooperation, varying by only one or two percentage points from pre-test to post-test (**Table 5**). Most treatment students had an average level of cooperation at pre-test (80%) and post-test (82%). Control students exhibited more variability than treatment students within the average level of cooperation and above average level of cooperation categories. Control students with an average level of cooperation decreased 10 percentage at post-test and those classified as above average experienced an increase of nine percentage points at post-test. The percent of control students with a below average level of cooperation increased by one percentage point at post-test. Outcomes between the treatment and control groups differed the most within the above average level of cooperation category. Twenty-two percent of control students were categorized as above average at post-test, but only 10% of treatment students were placed in that same category. It seems that more students in the control group improved their ability to cooperate than treatment students.

Table 5. Level of Cooperation among Treatment and Control Students at Pre- and Post-Test

	Below Average		Average		Above Average	
	TX	CTRL	TX	CTRL	TX	CTRL
PRE	6	7	53	83	7	14
	9%	7%	80%	80%	11%	13%
POST	4	6	42	51	5	16
	8%	8%	82%	70%	10%	22%

Assertion (Score Range: 0 – 21)
This subscale measures a child's attempt to initiate behaviors, such as asking others for information, introducing oneself, and responding to the actions of others.

As illustrated in **Table 6**, there was no significant difference between the treatment and control group at any level of assertion. For both treatment and control groups, the percent of students with a below average level of assertion decreased at post-test (0-4%). The percent of treatment students who displayed an average level of assertion increased from 82% at pre-test to 88% at post-test and the percent of control students in this category remained at 88% from pre-test to post-test, showing that treatment students were the only group to experience growth in this category. More students in the control group were considered above average at post-test (12%) compared to those in the treatment group (8%) which experienced a decrease of one percentage point from pre-test. In this case, control students at post-test displayed a slightly greater level of assertion than treatment students.

Table 6. Level of Assertion among Treatment and Control Students at Pre- and Post-Test

	Below Average		**Average**		**Above Average**	
	TX	**CTRL**	**TX**	**CTRL**	**TX**	**CTRL**
PRE	6	3	54	92	6	9
	9%	3%	82%	88%	9%	9%
POST	2	0	45	64	4	9
	4%	0%	88%	88%	8%	12%

Responsibility (Score Range: 0 – 18)

This subscale measures a child's ability to show regard for property or work and demonstrate the ability to communicate with adults.

Treatment and control students displayed similar results at post-test regardless of level of responsibility, with only slight differences observed between the two groups (**Table 7**). One difference was that the percentage of students in the control group considered average or above average increased by two to four percentage points at post-test, while experiencing a six percent decrease in the below average category from pre-test (9%) to post-test (3%). In contrast the percentage of treatment students in the below average and average categories decreased by exactly one percent at post-test but increased by two points in the above average category. For this subscale, control students developed a marginally greater level of responsibility at post-test when compared to treatment students.

Table 7. Level of Responsibility among Treatment and Control Students at Pre- and Post-Test

	Below Average		Average		Above Average	
	TX	CTRL	TX	CTRL	TX	CTRL
PRE	7	9	50	80	9	15
	11%	9%	76%	77%	14%	14%
POST	5	2	38	58	8	13
	10%	3%	75%	79%	16%	18%

Empathy (Score Range: 0 – 18)
This subscale measures a child's ability to show concern and respect for others' feelings and viewpoints.

In **Table 8** below, students in the treatment and control groups displaying an above average level of empathy experienced a growth of seven to twelve percentage points from pre-test (17-26%) to post-test (29-33%). These results suggest that across all subscales, students regardless of condition (treatment vs. control) performed the best in this subscale, but control students did seem to achieve greater gains in their level of empathy.

Table 8. Level of Empathy among Treatment and Control Students at Pre- and Post-Test

	Below Average		Average		Above Average	
	TX	CTRL	TX	CTRL	TX	CTRL
PRE	4	1	51	76	11	27
	6%	1%	77%	73%	17%	26%
POST	4	0	32	49	15	24
	8%	0%	63%	67%	29%	33%

Engagement (Score Range: 0 – 21)
This subscale measures a child's ability to join activities that are in progress and inviting others to join, initiating conversations, making friends, and interacting well with others.

According to **Table 9**, students in the control group developed their level of engagement at a greater rate than treatment students. For example, six percent more control students were classified as having an average level of engagement at post-test (71%) than treatment students at that same point in time (65%). Three percent more control students were classified above average in their level of engagement at post-test (25%) than treatment students (22%). Treatment students also differed in that they experienced an increase in the percentage of students classified as below average at post-test. Overall, control students experienced somewhat greater gains than treatment students when measuring their level of engagement.

Table 9. Level of Engagement among Treatment and Control Students at Pre- and Post-Test

	Below Average		Average		Above Average	
	TX	**CTRL**	**TX**	**CTRL**	**TX**	**CTRL**
PRE	6	5	48	78	12	21
	9%	5%	73%	75%	18%	20%
POST	7	3	33	52	11	18
	14%	4%	65%	71%	22%	25%

Self-Control (Score Range: 0 – 21)
This subscale measures a child's ability to respond appropriately in conflict (e.g. disagreeing, teasing) and non-conflict situations (taking turns and compromising).

Self-control was the most difficult social skill for students to develop. This subscale had the highest rate of treatment and control students considered below average out of all seven subscales, starting with 13-17% of students at pre-test and showing little improvement with 10-14% of students in the same category at post-test. Five percent more treatment and control students were characterized as having an average level of self-control at post-test than at pre-test. There was no positive change among students considered above average at post-test because the results for control students were equal to the results at pre-test and the percent of treatment students decreased from 14% to 12%. Although control students may have developed their level of self-control a bit more than treatment students, both groups had difficulty with this social skill when measured.

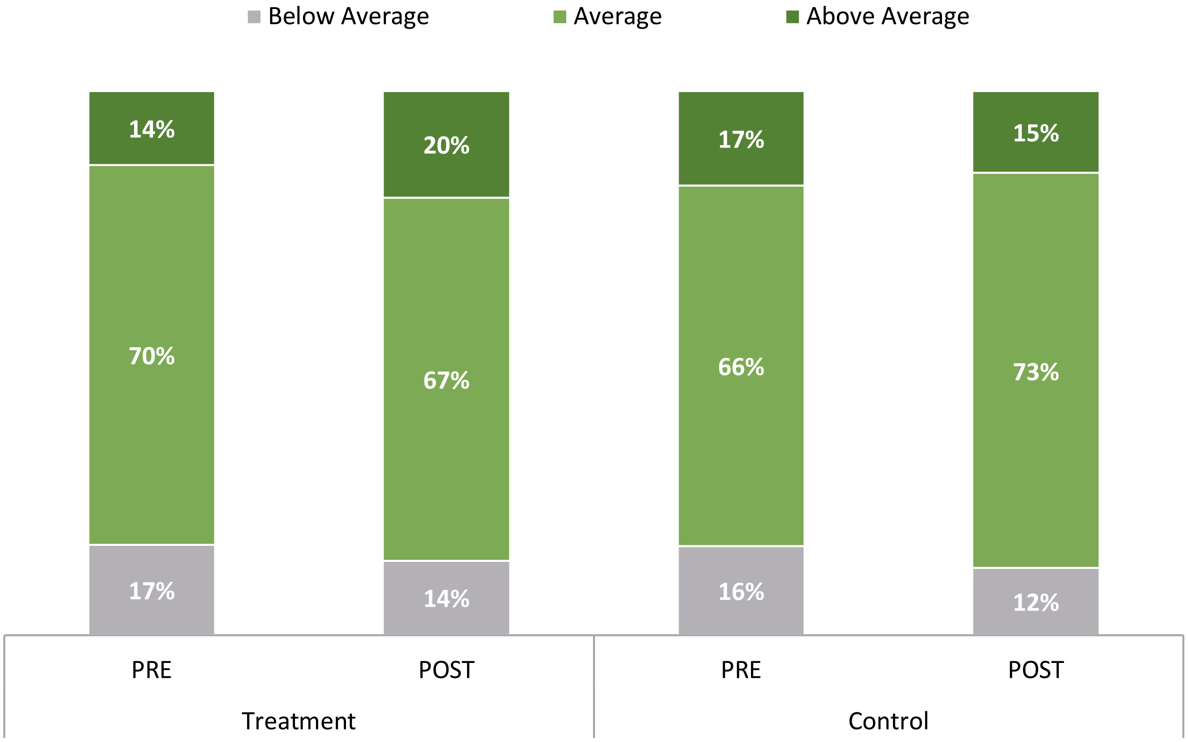
Table 10. Level of Self-Control among Treatment and Control Students at Pre- and Post-Test

	Below Average		**Average**		**Above Average**	
	TX	**CTRL**	**TX**	**CTRL**	**TX**	**CTRL**
PRE	11	14	46	68	9	22
	17%	13%	70%	65%	14%	21%
POST	7	7	38	51	6	15
	14%	10%	75%	70%	12%	21%

Composite (Score Range: 0 - 138)

In general, students in the treatment and control groups performed about the same with only a few marginal differences characterizing the experience of students in either group. For both groups the percent of students whose overall social skills were considered below average decreased from pre-test (16-17%) to post-test (12-14%). The majority of students in each group also experienced growth, although that growth was concentrated in a different social skills category for each group of students. For instance, treatment students experienced growth in the percent of students with an above average level of social skills, from 14% at pre-test to 20% at post-test. In contrast, there was an increase among control students in the percentage of those considered having an average level of social skills, from 66% at pre-test to 73% at post-test.

Figure 7. Behavior Level Corresponding to Composite SSIS Score for Treatment and Control Participants at Pre- and Post-Test



Overall, the differences between the treatment and control groups across all subscales and composite score were negligible with progress or regression of social skills development being represented by the difference of only a few percentage points. Although the results from analyzing student performance in the SSIS did not conclusively state that one group performed better than the other, control students did tend to experience marginally greater gains than treatment students in six out of the seven subscales – Cooperation, Assertion, Responsibility, Empathy, Engagement, and Self-Control.

Study Limitations

Study limitations are important to consider when weighing results. By necessity this study used a quasi-experimental research design, not a randomized control trial, and it is possible that due to a lack of random assignment to treatment there were pre-existing differences in treatment students that resulted in higher outcome scores on measures of literacy. While possible, this is highly unlikely, and in fact the opposite is more likely true: due to a lack of random assignment to treatment or control conditions, children in the treatment condition (the UPSTART program) were selected from families with higher levels of poverty than the control group families. This means that in effect control families were more advantaged and did not face the same hardships as many of the control families with higher levels of poverty.

The evaluator employed multiple strategies to recruit control families into the evaluation with equal levels of poverty. Normally this process is guided with state assistance. The FSSA did not assist in identifying low-income populations outside the program areas (i.e., equivalent on demographic and SES measures), and its lack of assistance pitted Waterford's program enrollment process against evaluation control group enrollment. This prevented the evaluator from recruiting higher numbers of control families at the same levels of poverty, which would have effectively reduced the number of families the program could serve. It would have been unethical to prevent a high poverty family from enrolling in UPSTART- a program that could result in great benefits for their child. State assistance in finding similar treatment and control student populations is common practice when states seek to optimize a research design, a benefit for their policy decision making. Overall this lack of state assistance conspired to make it harder to find and retain research subjects, and ultimately decreased our power to determine program effects.

Summary and Recommendations

The UPSTART program shows continued success at helping preschool aged children develop literacy skills and prepare for entry into kindergarten. Given its success at improving literacy test scores, while not hindering social skills, we recommend that the state continue providing the UPSTART program to children.

It is important to continually monitor and encourage appropriate program usage as this report's findings illustrate a positive relationship between increased program usage (as measured by hours of instruction) and significantly higher literacy outcome scores. Graduation rates need to be carefully monitored because a significant decline might erode literacy outcomes for the most at-risk students.

Program Recommendations. Although the graduation rates of UPSTART students were relatively low, 56%, UPSTART must 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). 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).

Evaluation Method Recommendations & Future Research. If a randomized control trial research design cannot be used in future research, we recommend that the state, FSSA, etc, review their goals and participate, even minimally, in assisting the evaluators in reaching targeted at-risk/low-income populations. Strong quasi-experimental research depends on collecting sufficient data from control students that are well balanced with treatment students across important covariates, such as income levels. To accomplish a balanced group design, we also recommend that the state work with the evaluators to strengthen relationships with other preschool providers that serve low-income families, specifically Head Start organizations, WIC and public preschool programs to widen our ability to collect data from non-program control families. This strategy is a win-win for all involved: low-income families can help move the bar on research into early literacy (and receive financial incentives while doing it) and the state can review results across more students and have more data for evidence-based decision making about their pre-Kindergarten school readiness programs.

References

Brigance, A. H. (2004). Brigance Inventory of Early Development III (IED-III) (3rd ed.). N. Billerica, MA: Curriculum Associates.

Cohen, J. (1988) Statistical Power Analysis for the Behavioral Sciences (2^{nd} 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.

Gresham, F. M., & Elliott, S. N. (2008). SSIS Rating Scales Manual. Minneapolis, MN: Pearson.

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.

Piasta, S. B., & Wagner, R. K. (2010). Developing Early Literacy Skills: A Meta-Analysis of Alphabet Learning and Instruction. Reading research quarterly, 45(1), 8–38. doi:10.1598/RRQ.45.1.2

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

Appendix A: Regression Model

To examine our main question - Does participation in the UPSTART program lead to better outcomes in literacy skills - we included a full factorial combination of Brigance Literacy composite, age, gender, parental education, parental marital status, family income, and poverty classification.

Table 11. Impact of Variables on Outcomes

Groups	**Variables**	**T-statistic**
Child groups	Treatment-control	t(115)=3.68***
	Brigance Literacy Post - Pre	t(115)=13.03***
	Age	t(115)=1.52
	Female-Male	t(115)=0.82
Parental Education	Some High School - High School	t(115)=-0.94
	High School - Some College	t(115)=-0.28
	Some College - Bachelor's	t(115)=-0.66
	Bachelor's - Master's	t(115)=-0.67
Parental Marital Status	Married- Separated	t(115)=-0.58
	Separated-Divorced	t(115)=-0.08
	Divorced-Unmarried	t(115)=0.41
Family income	<10,000 – 10,000 to 24,999	t(115)=1.73
	10,000 to 24,999 – 25,000 to 49,999	t(115)=1.63
	25,000 to 49,999 – 50,000 to 74,999	t(115)=0.78
	50,000 to 74,999 – 75,000 to 99,999	t(115)=-0.20
	75,000 to 99,999 - >100,000	t(115)=0.30
Poverty Classification	Poverty Level 200%	t(115)=-1.41

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.

Appendix C: SSIS Rating Scales

Behavior Levels Corresponding to Subscale Raw Scores for the Teacher and Parent Forms, Ages 3-5, by Norm Group *(pg.199 of SSIS Manual)*

Social Skills	**Below Average**	**Average**	**Above Average**
Communication	0 – 11	12 – 19	20 – 21
Cooperation	0 – 8	9 – 15	16 – 18
Assertion	0 – 10	11 – 19	20 – 21
Responsibility	0 – 7	8 – 15	16 – 18
Empathy	0 – 8	9 – 16	17 – 18
Engagement	0 – 10	11 – 19	20 – 21
Self-Control	0 – 6	7 – 14	15 – 21