An Evaluation of Istation Curriculum on Student Reading Growth: A Quasi-Experimental Study Using Propensity Score Analysis

Chalie Patarapichayatham[1], Ph.D.

Victoria N. Locke[2], Ph.D.

[1] Research Assistant Professor at Southern Methodist University, cpatarapichy@smu.edu

[2] Research Director at Istation, vlocke@istation.com

Introduction

Istation is an integrated learning system (ILS) that provides a formative assessment, intervention curriculum, and teacher resources to help educators identify students who need intensive intervention to succeed in the classroom. When a school uses both the assessment and the intervention curriculum, the program will route students to the appropriate parts of the curriculum where they need additional intervention based on their performance on the assessment. The Istation ILS also provides teachers with detailed reports on the students in the classroom, teacher-directed lessons to administer one on one or in groups, and a list of the students who would benefit from the additional instruction. These resources help teachers provide differentiated instruction in a timely manner that may help prevent students from falling behind. Istation also provides home access for students and professional development for educators on how to use the system and interpret its data and reports.

The formative assessment, known as Istation's Indicators of Progress Early Reading (ISIP™ ER), is an Internet- and Web-delivered computer-adaptive testing (CAT) system that provides continuous progress-monitoring assessments in the critical domains of reading in prekindergarten through eighth grade. It is built using two-parameter item response theory and driven by a fully CAT algorithm. Students take Istation assessments and/or Istation curriculum monthly from the beginning of the school year. Some students take Istation assessments and/or Istation curriculum three times a year under benchmarking assessment months (beginning of the year [BOY], middle of the year [MOY], and end of the year [EOY]). ISIP gathers and reports frequent information about student progress in the critical domains throughout and across academic years (Patarapichayatham, Fahle, & Roden, 2013). Reports are available for administrators, teachers, parents, and students to view on the website. The system alerts teachers

when students are not making adequate progress so that teachers can modify the instructional program before a pattern of failure becomes established (Mathes, Torgesen, & Herron, 2011).

Each year, more than 4 million students take the ISIP ER assessments. Several studies have shown a strong relationship between ISIP ER scores and state test scores. If a student does well on ISIP ER, it is very likely that they will do well on the state assessment test (LaPlante, 2018; Lester, 2017; Patarapichayatham, 2018; Patarapichayatham, Fahle, & Roden, 2013).

Istation Curriculum

Istation began in the 1990s, offering supplementary curriculum programs in early reading, designed for students who needed intervention. The interactive curriculum provides lessons based on theory and research for sound pedagogical practices for teaching particular reading skills. The curriculum uses engaging animation including characters and a game-like presentation to keep students interested in what they are learning. The instruction is divided into cycles, which start at foundational skills for the alphabet, alphabetic principle, print awareness, and other basic skills. As students grow in their reading skills, the system routes them into more difficult lessons that teach increasingly challenging material, including vocabulary.

After these cycles are completed, the system routes students into material that helps them develop reading comprehension skills. Teachers receive reports regarding students' performance within the curriculum as well as on the formative assessment, and these reports help teachers differentiate instruction for students. Istation's Priority Report indicates which students need additional instruction and suggests teacher-directed lessons, which are complete with a script and other resources needed for providing targeted instruction for individuals or groups of students.

While there is research on the formative assessment, there is less research available demonstrating the efficacy of the Istation curriculum. Putman (2016) evaluated the efficacy of

Istation curriculum using a quasi-experimental design with propensity score matching, using scores on the Dynamic Reading Assessment-2 (DRA-2) as an outcome measure, and found that Istation helped kindergarten students grow. The use of Istation explained 17.7% of the variance compared to the control group. When teacher literacy support was added to the model, 24.5% of the group differences were explained, and there was no interaction effect between Istation and teacher literacy support. Limitations of the study included its small sample size and unequal groups. Lundin (2017) looked into the effects of Tier 1 Response to Intervention (RtI) on elementary school reading fluency using Istation assessment and intervention. While this study did not evaluate the overall ISIP scores, the results showed significant growth between the pre- and post-test screenings on the Istation reading fluency assessment. Other studies have also demonstrated student growth on the Renaissance STAR following usage of the Istation curriculum in third grade (Luo, Lee, & Molina, 2017), but it was difficult to tease out the impact of Istation alone versus other classroom and home activities designed to increase student growth.

Other research with at-risk students (in the bottom 20th percentile) found that when compared to students above the 20th percentile, those who were at-risk showed greater gains in their ISIP scores when using the Istation curriculum. Students who used both the school and home components for more than 30 minutes a week showed the greatest gains (Sutter, Campbell, & Lambie, 2019).

The purpose of this study is to fill the gap in the research and evaluate the impact of the Istation curriculum on students' reading achievement in kindergarten, first grade, and second grade with a larger sample size. Most of the school systems that use the Istation ILS use the entire system, including the formative assessment and the curriculum; however, some schools only use the formative assessment, professional development, and teacher-directed lessons.

Two research questions guided our inquiry.

RQ1: Does use of the Istation curriculum make a measurable difference in students' growth in reading?

RQ2: If there is a measurable difference, does it vary by grade?

The design of this study is intended to meet the What Works Clearinghouse (WWC) 4.0 standards for quasi-experimental studies (WWC, 2017a) and the guidelines for a Level 2 rating for the Every Student Succeeds Act (ESSA) guidance for evidence-based research (U.S. Department of Education, 2016).

Data and Methods

The ISIP ER was used in this study to measure reading growth for students in kindergarten through second grade. The overall reading ability scores from the beginning of the year (BOY, September 2018) and the end of the year (EOY, May 2019) were used to compute a gain or growth score for the academic year. The students' scores for the EOY assessment month for the previous academic year (May 2018) served as a pretest and were used for propensity score weighting. Students with all three of these data points were included in the study, and students with missing data points were dropped from this analysis. In ISIP ER, as students master lower-level skills, they progress to more difficult skills, and therefore in each grade, different subtests comprise the overall ISIP ER score. In kindergarten, the overall scores are computed based on Listening Comprehension, Phonemic Awareness, Letter Knowledge, and Vocabulary subtests. In first grade, the overall scores are computed from the Phonemic Awareness, Letter Knowledge, Alphabetic Decoding, Reading Comprehension, Spelling, and Vocabulary subtests. In second grade, the overall scores are computed from the Reading Comprehension, Spelling, and Vocabulary subtests.

Sample Construction

Data were collected from the extensive Istation database of more than 1,000 schools across the state of Texas in the 2018-2019 academic year. We wanted to compare students who do not use the curriculum with those who do use the curriculum in schools where Istation is well-implemented. Since the quality of Istation implementation will vary across schools and may impact student growth, we set criteria defining good implementation of the curriculum. Istation has usage recommendations of 30 minutes per week for students above the 40th percentile and 40 minutes per week for students below the 40th percentile. Preliminary analysis showed that few schools met this threshold. Therefore, we set the good implementation cut point as an average of 400 minutes of curriculum usage per student during the school year, which equates to approximately 40-45 minutes per month. While this threshold is less than the Istation recommendations, previous research indicates it is sufficient to determine if use of the Istation curriculum is associated with positive gains in students' reading achievement (Patarapichayatham, 2014).

Our final sample consisted of students who used only the ISIP ER assessment as the comparison group, and students who were in schools that used the Istation curriculum an average of 400 minutes or more for all students in the school year comprised the treatment group. Students in schools that used the curriculum for less than 400 minutes on average were not included in this analysis.

Research Design

We used a quasi-experimental design with propensity score matching to assess whether there are differences in growth. Because of usage patterns, approximately 70% of the sample consisted of students who used both the ISIP ER assessment and Istation curriculum, whereas

about 30% of the sample used the ISIP ER assessment only. This creates an unbalanced sample between the treatment group and comparison group, which may introduce bias, and the results from the analysis may not be as reliable if these two groups are directly compared. To control for this discrepancy, we used propensity score weighting, which balances the treatment effect in the treatment and comparison groups (Austin, 2014).

Model

Propensity score matching is a statistical matching technique, popular in educational research and program evaluation, to estimate the effect of a treatment or intervention by accounting for the covariates that predict receiving the treatment. Propensity score matching attempts to reduce the bias in an estimate of the treatment effect obtained from simply comparing outcomes among units that received the treatment versus those that did not (Baker et al., 2018). Propensity score analysis is used to create balance and to compare treatment groups and conditions for the purposes of estimating unbiased treatment effects. This design is appropriate when a randomized comparison design study is not possible or is considered unethical or impractical (Guo & Fraser, 2015; Kamata et al., 2019; Rosenbaum & Rubin, 1983).

A potential drawback of propensity score matching is that a very large number of unmatched subjects are not used. Substantial amounts of data can be discarded from the analysis, which can bias the results (Austin, 2014). To control for this limitation and maintain a robust sample size, we used propensity score weighting so that all the subjects in the treatment and comparison groups can be used in the analysis, which helps reduce bias (Austin, 2014). Under propensity score weighting, propensity scores are used as a weight in a linear model such as a regression, ANOVA, or multivariate analysis.

Several methods are suitable for propensity score weighting analysis. However, Kamata, Gallegos, Patarapichayatham, and Kara (2019) investigated the effect of the number of covariates, types of covariates, and degrees of association between covariates and treatment and/or outcome in the propensity score model. They mainly focused on propensity score weighting (Guo & Fraser, 2015; Rosenbaum & Rubin, 1983) as the treatment effect estimation method through a series of simulation conditions. They found that overall the covariate balancing propensity score (CBPS) method performed better than a generalized boosted regression (GBR) method and traditional logistic regression. The CBPS method is used in this current study.

Propensity score weighting involves three main steps. First, a propensity score is derived for each student, and this score represents the likelihood of being in the treatment group. Logistic regression is then used with the outcome being whether or not the student is actually in the treatment group. Second, the derived propensity scores are then used to weight students in the treatment group and students in the comparison group. Third, a statistical analysis model is fit to the weighted sample.

Analysis

Before conducting the propensity score weighting, we evaluated the correlation between the BOY and EOY scores. The Pearson product-moment correlations were strong at 0.62 in kindergarten, 0.81 in first grade, and 0.84 in second grade. Next, we used propensity score weighting to minimize the bias in estimating the treatment effect in treatment and comparison groups. Third, the model, which evaluates the differences in the gain score between students in treatment and comparison groups, was fit.

To derive propensity scores, all demographic variables were assumed to be related to the outcome measure (e.g., Brookhart et al., 2006; Rubin, 1997). Five variables in total were used to compute propensity scores using logistic regression (see Table 1).

Table 1 about here

We weighted students on one school-level variable and four student-level variables. The school-level variable was the percentage of students receiving free or reduced-price lunch (FRPL). High-poverty schools consist of public schools where more than 75.0 percent of the students are eligible for FRPL, and mid-high poverty schools are those where 50.1 to 75.0 percent of the students are eligible for FRPL. Low-poverty schools consist of public schools where 25.0 percent or less of the students are eligible for FRPL, and mid-low poverty schools are those where 25.1 to 50.0 percent of the students are eligible for FRPL (https://nces.ed.gov). We combined the low and mid-low poverty schools into one group and the mid-high and high poverty schools into another group.

The four student-level variables were the pretest, gender (male or female), special education status (yes or no), and race/ethnicity (non-Hispanic white versus others). Complete demographics for the analytic sample are available in Table 2, and pretest scores are shown in Table 4. Because propensity score weighting requires complete data (pretest, BOY, EOY, and all five covariates), students with incomplete data were removed. The final samples consisted of 47,886 students across all three grades: 3,937 students in kindergarten, 21,567 students in first grade, and 22,382 students in second grade. The kindergarten sample is smaller because the matching criteria used the score from the previous year, which in this instance was from prekindergarten. Istation has fewer students enrolled in prekindergarten, and most of these students are in higher poverty schools. Table 2 shows that students' demographics were quite

different between the treatment and comparison groups. The propensity score weighting will play an important role to minimize bias from these two groups.

Table 2 about here

We used the CBPS method with the CBPS package in *R* statistical software, and all analyses were completed using *R*. The CBPS method simultaneously derives the propensity scores and weights for observations to optimize covariate balance between the treatment and comparison groups (Fong, Ratkovic, & Imai, 2019). We used logistic regression to determine the probability of membership in the treatment group or comparison group, given the specific set of selection covariates included. Propensity score weights were computed using the average treatment effect, which is the difference in the outcome variable between the average score for the treatment group and the comparison group. Results for the analytic sample after propensity score matching and weighting are available in Table 3.

After the propensity score weighting was implemented, the mean differences in the percent of students in mid-high to high-poverty schools were between 0% and 13%. Mean differences in the gender composition of the sample were reduced to 3% or less for each grade, and mean differences in students receiving special education were also reduced to 3% or less. Mean differences for the percentage of students who are non-Hispanic white were eliminated in kindergarten and were reduced by more than half for second and third grade.

Table 3 about here

We estimated the differences in pretest scores using propensity score weights in a linear regression model. Results are available in Table 4. Before propensity score weighting, mean differences between groups were 3.82 points in kindergarten, 3.55 in first grade, and 7.00 in

second grade. After weighting, mean differences in gain scores were less than 1.00 point in each grade, and less than .09 of a standard deviation. All five covariates were balanced.

Table 4 about here

Students' growth was measured by evaluating a gain score. BOY mean scores were subtracted from EOY mean scores for students in the treatment group and comparison group. We analyzed differences in growth at the student level. Prior research indicates that differences in usage are good indicators of implementation and thus score differences between schools (Patarapichayatham, 2014). We controlled for school-level effects by only selecting schools with good implementation and using the school-level FRPL variable in the propensity score weighting described above. Next, we estimated the differences in gain scores using propensity scores weights in a linear regression model to estimate an impact of the treatment effect. Finally, an independent-samples *t*-test was further conducted to test the gain scores' differences between the treatment and comparison groups. Cohen's D effect size was also conducted to test the size of the effects between these treatment and comparison groups.

Results

Results confirmed a significant and meaningful effect of Istation curriculum on reading growth scores after the groups were balanced using propensity scores across grades. For kindergarten, the estimated gain score from BOY to EOY of the comparison group was 18.32 (mean = 18.32, SD = 0.31), whereas the estimated gain score of the treatment group was 22.15 (mean = 22.15, SD = 0.41). Kindergarteners who used the curriculum (treatment) grew 20.91% more than kindergarteners who did not use the curriculum. For first grade students, the estimated gain score of the comparison group was 17.76 (mean = 17.76, SD = 0.11), whereas the estimated gain score of the treatment group was 20.23 (mean = 20.23, SD = 0.14). First grade students who

used the Istation curriculum grew 13.91% more than the comparison group. For second grade students, the estimated gain score of the comparison group was 11.24 (mean = 11.24, SD = 0.10), whereas the estimated gain score of the treatment group was 14.43 (mean = 14.43, SD = 0.14). Second grade students who used the Istation curriculum grew 28.38% more than the comparison group. These comparisons are available in Figures 1 and 2.

Figure 1 about here

Figure 2 about here

Results of the independent-samples *t*-test and Cohen's D confirm that these results are statistically significant, and these results are shown in Table 5. The independent-samples *t*-test showed significant differences in gain scores between the treatment group and comparison group for all grades (kindergarten: $t = -3.69$, $p < 0.001$; first grade: $t = -18.69$, $p < 0.001$; and second grade: $t = -13.70$, $p < 0.001$). The Cohen's D effect size also showed small effect sizes between the treatment group and comparison group across grades. They were 0.14, 0.14, and 0.16 for kindergarten, first grade, and second grade, respectively.

Table 5 about here

Discussion

This study confirms the use of the Istation curriculum makes a measurable difference in students' growth in reading. This study also confirms that students' growth in reading varies by grade level. Significant differences were shown in kindergarten, first grade, and second grade, confirming that the Istation curriculum helps students achieve basic and more complex concepts as they progress through the early elementary grades. While the technology does not replace the teacher, it appears that the Istation curriculum can serve as an effective supplementary tool for learning, and it helps students increase their growth in reading. These results confirm the

previous literature that found student growth when using the Istation curriculum (Putman, 2017; Luo, Lee, & Molina, 2017; Sutter, Campbell, & Lambie, 2019).

School administrators, teachers, and parents can use the results from this study to inform their estimates of students' scores when they use or do not use the Istation curriculum. For example, if a kindergartener scores 180 at the beginning of the year, their estimated EOY score would be approximately 198 if they do not use Istation curriculum, and the estimated EOY score would be approximately 202 if they do use the Istation curriculum. For a second grader, the gains are similar. If the student begins with a score of 220, they will have an EOY score of approximately 231 if they do not use the Istation curriculum and 234 if they do. This can make the difference between being in Tier 2 (40th percentile and below) versus being in Tier 1 (above the 40th percentile).

Limitations

There are a number of limitations to consider when interpreting these results. First, even though the propensity score analysis is popular in education program evaluation, it is important to point out that results may depend on the number and set of covariates in the model. More covariates at both the school and student levels (such as urbanicity and school size, school achievement, and student ELL status, enrollment in bilingual or dual-language programs, after-school enrichment, and parent education level) may impact student achievement. Second, other than setting criteria of 400 minutes of curriculum usage on average across all students, the fidelity of Istation implementation and usage at the teacher level were not controlled for in this study, and these variables may result in differences in student growth. We set a limit of 400 minutes of curriculum usage, which is lower than Istation recommendations. Adherence to the time recommendations may also affect the results. Finally, we restricted the sample to one state.

While Texas is a large and diverse state and has approximately 60% of students participating in the Istation reading program (as of May 2019), the generalizability to other states may be limited. A sample of students in different states with more variability regarding curriculum usage is recommended for future study.

Conclusion

The findings demonstrate that the Istation curriculum helped students with their reading achievement. Students who used Istation curriculum have higher gain scores from the beginning of the school year to the end of the school year. Istation usage recommendations are based on research indicating that time spent on Istation curriculum helps students make gains in reading (see usage recommendation at www.istation.com). Since several studies revealed that ISIP scores are predictive of passing state tests (LaPlante, 2018; Patarapichayatham, 2018; Patarapichayatham 2019; Campbell, Lambie, Sutter, Bicham, & Pulse, 2019), the findings from this study indicate that at-risk students who use the Istation curriculum have a higher chance of passing state tests than students who do not use the Istation curriculum.

References

Austin, P. C. (2017). Double propensity-score adjustment: A solution to design bias or bias due to incomplete matching. *Statistical Methods in Medical Research*, *26*(1), 201–222. https://doi.org/10.1177/0962280214543508.

Baker, S. K., Kamata, A., Wright, A., Farmer, D., & Nippert, R. (2018). Using propensity score matching to estimate treatment effects of afterschool programs on third-grade reading outcomes. J Community Psychology. 2019; 47: 117–134. https://doi.org/10.1002/jcop.22104.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Stürmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology*, *163*, 1149–1156.

Campbell, L. O., Lambie, G. W., Sutter, C. C., Bickham, A. R., & Pulse, L. P. (2018). Measuring the predictability of Istation Indicators of Progress Early Reading (ISIP ER) scores on Florida Standards Assessment (FSA) scores. University of Central Florida. www.ucf.edu/mirc.

Fong, C., Ratkovic, M., & Imai, K. (2019). CBPS: Covariate balancing propensity Score. R package version 0.20. https://CRAN.R-project.org/package=CBPS.

Guo, S., & Fraser, M. W. (2015). *Propensity score analysis: Statistical methods and applications* (Ed. 2). Thousand Oaks, CA: SAGE Publications, Inc.

Kamata, A., Gallegos, E., Patarapichayatham, C., & Kara, Y. (2019*). Evaluation of covariate balance as a quality indicator for propensity score matching analysis*. Paper presented to the annual meeting of American Educational Research Association, Toronto, ON, Canada, April, 2019.

LaPlante, J. (2018). Predictability Study of Istation ISIP (Math and Reading and Ohio AIR (Math and English Language Arts) Tests for 3rd – 8th Grade Students in the Youngstown City School District. http:/www.istation.com/studies.

Lester, A. A. (2017). The effectiveness of Istation when used for computer-assisted instruction (CAI) on the overall reading ability score of primary level students. Doctoral dissertation of Texas A & M University-Commerce.

Lundin, M. M. (2017). The effects of tier one RTI at the elementary school level using Istation on reading text fluency. Doctoral dissertation of University of Houston-Clear Lake.

Luo, T., Lee, G-L.., & Molina, C. (2017). Incorporating Istation into early childhood classrooms to improve reading comprehension. *Journal of Information Technology Education Research*, 16, 247-266. Retrieved from http://www.informingscience.org/Publications/3788 (CC BY-NC 4.0).

Mathes, P., Torgesen, J., & Herron, J. (2011). Technical manual: Istation's Indicators of Progress Early Reading: Computer Adaptive Testing System for Continuous Progress Monitoring of Reading Growth for Students Pre-K to Grade 3.

Olmos, A., & Govindasamy, P. (2015). A practical guide for using propensity score weighting in R. *Journal of Practical Assessment, Research & Evaluation, 20(13), 1-8.*

Patarapichayatham, C. (2014). Istation Reading Growth Study Grades 1 - 8. Dallas, TX: Istation.

Patarapichayatham, C. (2018). Predictability Study of ISIP Reading and Virginia Standards of Learning (SOL) for English Reading: 3rd – 5th Grade Students. Dallas, TX: Istation.

Patarapichayatham, C. (2019). Predictability Study of ISIP Reading and the Colorado Measurement of Academic Progress for English Reading: 3rd – 5th Grade Students. Dallas, TX: Istation.

Patarapichayatham, C., Fahle, W., & Roden, T. R. (2013). ISIP Reading versus STAAR Reading: The Predictability Study. Dallas, TX: Istation.

Putman, R. (2017). Technology versus teachers in the early literacy classroom: an investigation of the effectiveness of the Istation integrated learning system. *Education Technology Research and Development* 65: 1153. https://doi.org/10.1007/s11423-016-9499-5.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rosenbaum, P. R., & Rubin, D. B., (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41-55. http://dx.doi.org/10.1093/biomet/70.1.41.

Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine*, *127*(8_Part_2), 757–763.

Sutter, C. C., Campbell, L. O. & Lambie, G. W. (2019). Computer-Adaptive Reading to Improve Reading Achievement Among Third-Grade Students At Risk for Reading Failure. *Journal of At Risk Issues, 22(2), 31-38.*

Stone, C. A., & Tang, Y. (2013). Comparing propensity score methods in balancing covariates and recovering impact in small sample educational program evaluations. *Practical Assessment, Research & Evaluation, 18(13), 1-12.*

U.S. Department of Education, Office of Elementary and Secondary Education. (2016, September). Non-Regulatory Guidance: *Using Evidence to Strengthen Education Investments*. Retrieved from https://www2.ed.gov/policy/elsec/leg/essa/ guidanceuseseinvestment.pdf.

What Works Clearinghouse, Institute of Education Sciences, U.S. Department of Education. (2017a, October). *What Works Clearinghouse: Procedures Handbook (Version 4.0)*. Retrieved from http://whatworks.ed.gov.

What Works Clearinghouse, Institute of Education Sciences, U.S. Department of Education. (2017b, October). *What Works Clearinghouse: Standards Handbook (Version 4.0)*. Retrieved from http://whatworks.ed.gov.

Table 1: Demographics Variables Used in Propensity Score Weighting

Variable Level	Variable	Variable Type
School	FRPL (low and mid-low vs. mid-high and high)	Dichotomous
Student	Pretest	Continuous
	Gender (male vs. female)	Dichotomous
	Enrolled in Special Education (yes vs. no)	Dichotomous
	Race (non-Hispanic white vs. others)	Dichotomous

Table 2: Complete Demographics for the Analytic Sample Before Propensity Score Matching

Variable		Kindergarten		First Grade		Second Grade	
		Treatment	Comparison	Treatment	Comparison	Treatment	Comparison
		n = 3,838	n = 99	n = 17,911	n = 3,656	n = 18,362	n = 4,020
Gender	Male	1951(50.8%)	56 (56.6%)	9044 (50.5%)	1826 (49.9%)	9486 (51.7%)	2091 (52.0%)
	Female	1887 (49.2%)	43 (43.4%)	8867 (49.5%)	1830 (50.1%)	8876 (48.3%)	1929 (48.0%)
Race	Non-Hispanic White	278 (7.2%)	35 (35.4%)	3125 (17.4%)	1559 (42.6%)	2991 (16.3%)	1728 (43.0%)
	Others	3560 (92.8%)	64 (64.6%)	14786 (82.6%)	2097 (57.4%)	15371 (83.7%)	2292 (57.0%)
SES	High and Mid-High	3564 (92.9%)	71 (71.7%)	13325 (74.4%)	2663 (72.8%)	14451 (78.7%)	3122 (77.7%)
	Low and Mid-Low	274 (7.1%)	28 (28.3%)	4586 (25.6%)	993 (27.2%)	3911 (21.3%)	898 (22.3%)
Special	Yes	224 (5.8%)	11 (11.1%)	1420 (7.9%)	700 (19.1%)	1690 (9.2%)	804 (20.0%)
Education	No	3614 (94.2%)	88 (88.9%)	16491 (92.1%)	2956 (80.9%)	16672 (90.8%)	3216 (80.0%)

Table 3: Demographic Covariates Balanced Before and After Propensity Score Weighting: FRPL, Gender, Special Education, and Race/Ethnicity

Variable	Kindergarten		First Grade		Second Grade	
	Before	After	Before	After	Before	After
FRPL						
Treatment Group	0.92	0.91	0.68	0.61	0.69	0.61
Comparison Group	0.71	0.91	0.27	0.50	0.22	0.48
Gender						
Treatment Group	0.50	0.51	0.50	0.50	0.51	0.52
Comparison Group	0.56	0.54	0.49	0.50	0.52	0.54
Special Education						
Treatment Group	0.05	0.05	0.07	0.11	0.09	0.12
Comparison Group	0.11	0.05	0.19	0.11	0.20	0.15
Race/ Ethnicity						
Treatment Group	0.07	0.07	0.21	0.25	0.20	0.25
Comparison Group	0.12	0.07	0.42	0.35	0.42	0.35

Table 4: Pretest Means – Covariate Balanced Before and After Propensity Score Weighting

Grade	Treatment Group		Comparison Group		Difference in SD
	Mean	SD	Mean	SD	
Kindergarten					
Before propensity score weighting	180.81	11.87	177.99	11.69	.24
After propensity score weighting	**180.74**	**11.87**	**179.81**	**11.81**	**.08**
First Grade					
Before propensity score weighting	202.21	13.40	205.76	13.64	.26
After propensity score weighting	**202.73**	**13.44**	**202.12**	**13.40**	**.05**
Second Grade					
Before propensity score weighting	221.94	12.51	228.94	12.90	.56
After propensity score weighting	**222.17**	**12.58**	**221.93**	**12.51**	**.06**

Table 6: Estimated Students' Growth from BOY to EOY

Grade	Treatment		Comparison		Difference			
	Mean	SD	Mean	SD		*t*	*p*	Cohen's D
Kindergarten	22.15	0.41	18.32	0.31	3.83	−3.69	<.001	0.14
First Grade	20.23	0.14	17.76	0.11	2.47	−18.69	<.001	0.14
Second Grade	14.43	0.14	11.24	0.10	3.19	−13.70	<.001	0.14

Figure 1. Mean Differences in Score Gains between Istation Curriculum Users and Curriculum Nonusers

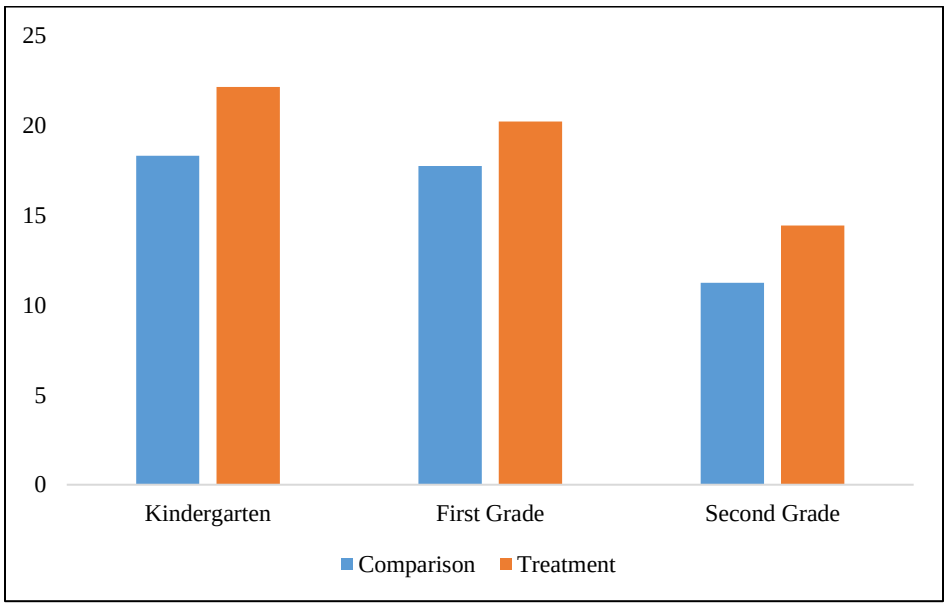

Figure 2. Percentage of Increased Gain Scores for Treatment Group over Comparison Group