RESEARCH RESULTS

HMH Into Reading: 2020–2022 QED Study

STUDY PROFILE

ESSA EVIDENCE LEVEL:
Promising

STUDY CONDUCTED BY:
Cobblestone Applied Research & Evaluation, Inc.

EVALUATION PERIOD:
2020–2022 School Years

STUDY DESIGN:
Quasi-Experimental Design

SAMPLE:
316 Elementary Schools
140 Treatment Schools
176 Control Schools

OUTCOME MEASURE:
State of Texas Assessments of Academic Readiness

THE CHALLENGE

According to the National Assessment of Educational Progress (NAEP), fourth grade reading scores have remained stable since 2005 with only about a third of students showing proficiency in reading ability. However, there has been a recent decline in fourth grade reading test scores, likely due to the COVID-19 pandemic and subsequent impacts on schooling. Scores dipped significantly by three points in 2022 when compared to 2019—lower than all previous assessment years since 2005 (National Center for Education Statistics [NCES], 2022). Decreases in scores were observed across demographic characteristics (e.g., sex, SES, and most ethnicities) and for students attending public schools (NCES, 2022). The Progress in International Reading Literacy Study (PIRLS) compared the reading ability of several countries, and the most recent assessment conducted in 2016 found that the United States is trailing many countries (Warner-Griffin et al., 2017). Due to the foundational nature of reading to general educational outcomes, it is imperative that reading education improve in the United States.

Given this, the 2015 federal Every Student Succeeds Act (ESSA) aims to identify and promote education programs that show evidence of success across four levels of evidence (Tier 1, Strong Evidence; Tier 2, Moderate Evidence; Tier 3, Promising Evidence; and Tier 4, Demonstrates a Rationale).

Several elementary literacy interventions, often multiple years into implementation, have demonstrated success under the Tier 1 level of ESSA standards. Research on reading programs meeting ESSA Tier 1 has demonstrated that students served by these programs score higher on norm-referenced reading assessments than control groups, with effect sizes ranging from 0.14 to 0.17 (Evidence for ESSA, 2019; Wilkerson & Savoy, 2013). Given these research findings, it is important to investigate how students acquire and retain reading knowledge through *Into Reading* in comparison to other competitor programs. Before HMH can assess the impact of *Into Reading* through a Tier 1 or 2 study, however, it is critical to understand if students who use the program outperform similar students who do not use *Into Reading* (Eddy et al., 2023).

THE SOLUTION

INTO READING

Into Reading is a K–5 literacy curriculum designed to facilitate reading and writing outcomes through student self-actualized learning. The curriculum is differentiated by design and aims to foster a culture of learning and growth in the classroom. The *Into Reading* curriculum materials include high-quality, engaging text sets, small-group lessons, assessments, easy-to-interpret reports, and instructional resources, as well as online professional support for teachers. In addition, for the purposes of this study, it is important to note that *Into Reading Texas* also includes assessments aligned to the Texas Essential Knowledge and Skills (TEKS) curriculum standards, and materials to prepare students for the State of Texas Assessments of Academic Readiness (STAAR®).

IMPLEMENTATION MODEL

Into Reading is a core curriculum designed to be used daily to support reading, writing, spelling, comprehension, and language development.

The program materials support both the student and teacher experience in the classroom:

STUDENT EXPERIENCE

- **Reading:** Students have access to a large library of culturally relevant and ethnically diverse texts to build cross-disciplinary knowledge, including Rigby® Readers and trade books. Notice and Note guides direct students to the identification of key signposts to interpret texts for meaning.
- **Writing:** The Writer's Notebook teaches students writing conventions through analysis, genre exploration, and a visual format for organizing their thoughts.
- **Integrated Instruction:** Reading and writing lessons are implemented in tandem. Books are used as a springboard for writing and discussion.
- **Scaffolding:** Students receive initial support and practice that prepares them to become independent learners.
- **Differentiation:** Lesson plans are customizable to provide appropriate support for each student.

TEACHER EXPERIENCE

- **Instructional Support:** A fully integrated online platform allows teachers to plan, teach, assess, and differentiate instruction. Small-group routines, scaffolds, and strategies for English learners are provided.
- **Professional Support:** Getting Started courses and online coaching videos provide information on teaching strategies and techniques.

THE STUDY

STUDY PURPOSE

Cobblestone Applied Research & Evaluation, Inc. (Cobblestone) initially launched a quasi-experimental design (QED) study to determine the potential impact of the *Into Reading* program on student reading outcomes in 2019, shortly after release of the program. The design of the 2019 study was sufficient to meet ESSA Tier 2 standards, pending a significant difference in the posttest variable of interest (2019–2020 STAAR reading scores) for treatment and control sites. However, due to the COVID-19 pandemic and the cancellation of the 2019–2020 STAAR student testing, the initial study was suspended in spring 2020.

In September 2021, Cobblestone launched a study with a slightly modified design from what was used in 2019. Similar instruments and study procedures were used to conduct the current QED study. Minor revisions to the study instruments were required to assess the impact of COVID-19 procedures or protocols present in schools during the 2020–2021 and 2021–2022 school years. In addition, while treatment and control school sites were still included in the analyses, the study compared treatment and control schools overall based on an initial propensity score matching and baseline equivalency in addition to the site verification process. The 2021 QED study was designed to meet ESSA Tier 3 standards, pending a significant difference in the posttest variable of interest (2021–2022 STAAR reading scores) for treatment and control sites.

The purpose of the QED study is to determine the potential impact of the *Into Reading* program on student reading outcomes. The study was designed to compare state reading scores from *Into Reading* (treatment) schools to comparable non-*Into Reading* (control) schools across multiple districts in the state of Texas. The study was designed to meet ESSA Tier 3 Promising Evidence.

The purpose of the study was to answer one key research question:

- **Do schools using *Into Reading* outperform comparable schools that use another ELA program?**

STUDY DESIGN

To meet the **ESSA Tier 3 Promising Evidence** criteria, studies must include the following:

Source: Department of Education (n.d.)

To meet these requirements, the study included schools across the state of Texas. This state was selected because of the large number of school/districts who purchased the curriculum. These schools were initially included in the treatment group, while schools who did not purchase the curriculum were initially included in the comparison group. It should be noted that schools were not randomly assigned to conditions but rather selected based on their purchase and confirmed usage of the *Into Reading* program.

To detect potential effects of the *Into Reading* program, a power analysis determined that at least 100 (50 treatment and 50 control) schools were necessary to detect at least a medium effect size (.10). To bolster results and prepare for possible attrition, the sample goal was increased to 200 (100 treatment and 100 control) schools. In fall and winter 2021, the Cobblestone research team verified *Into Reading* program usage with each site in the sample.

STUDY SAMPLE

PROPENSITY SCORE MATCHING PROCESS

To determine the study sample, the Cobblestone research team began with HMH sales data to determine which districts in Texas had purchased *Into Reading* and intended to use the curriculum. To obtain the variables and information needed to conduct the propensity score matching process, sales data were combined with information from other publicly available data sources. These sources included Market Data Retrieval (MDR®) data purchased by HMH, Texas Education Agency (TEA) data, and information from the National Center for Education Statistics (NCES). In combination, these variables were used to create propensity scores for each school.

Propensity scores were in part used to select appropriate, comparable control group school sites. Propensity score matching is a technique used to balance meaningful covariates across treatment and control groups (Rosenbaum & Rubin, 1985). It is essential to ensure the covariates of treatment and control schools are balanced because having a larger proportion of schools with a particular covariate in one group could influence the overall results of the analysis. Thus, propensity score matching was used to match control schools to the treatment schools based on school classification (e.g., K–6), school enrollment, available demographic information (i.e., SES based on average family income and percentage of students with financial need, percentage of white and Hispanic/Latino students, and percentage of students enrolled in ELL), as well as neighborhood lifestyle (e.g., city centers), type of community (e.g.,

large city), and the percentage of students at or above state standard in reading for Grades 3–5 from 2019 (note, MDR student data was selected from 2018–2019 as it was the most recent data source at the time of the propensity score analysis). It should be noted that, as school classification, neighborhood lifestyle, and community type were categorical variables, they were set to require an exact match in the procedure.

There were 4,914 (2,369 treatment and 2,545 control) schools with MDR data, based on the latest available scores from 2019. However, some schools were missing data on the matching variables and were removed. Using the custom dialog, 553 treatment schools were matched with 553 control schools.

After matching, the Cobblestone research team examined the balance of all observed covariates, interactions among all covariates, and quadratic terms of covariates. The Cobblestone research team then selected 404 (202 treatment and 202 control) schools that represented the best matches—that is, those matches that had the smallest differences between their propensity scores. No significant differences were found across the treatment and control schools for any of the matching criteria. As such, selecting the schools that provided the best matches helped to ensure that the groups were as equivalent as possible at baseline.

Of note, when comparing the percentage of schools offering in-person, hybrid, or online modalities of instruction, there were significant differences in modality for both the 2020–2021 and 2021–2022 academic years. In 2020–2021, there were a greater proportion of treatment schools offering a hybrid modality than control schools ($\chi 2(2) = 7.18$, $p = .03$). In 2021–2022, there were a greater proportion of treatment schools offering in-person modality than control schools ($\chi 2(1) = 5.02$, $p = .03$). Therefore, modality of instruction was controlled for in the final analysis of outcomes. (See Appendix A for a detailed summary of the propensity score-matching process and Appendix C for a detailed summary of the baseline equivalence analysis.)

SITE VERIFICATION PROCESS

Although the study sample was initially based on *Into Reading* district-level HMH sales data, the Cobblestone research team deemed it necessary to verify that each site was using (or not using) the curriculum as planned, given that the actual usage of the program at a particular school could vary within a district. In addition to the verification protocol used in 2019, the site verification included assessing the modality of instruction (i.e., in-person, remote, hybrid).

In fall 2021, the Cobblestone research team began to verify *Into Reading* usage with each site in the sample and planned to reconfirm usage for treatment sites in spring 2022. However, since the COVID-19 pandemic, access to key contacts at school sites was challenging. For example, many schools experienced staff shortages in 2021, and often the school contact with the best knowledge of curriculum was not readily available to answer questions. The Cobblestone research team used multiple methods to contact school personnel, including through email and phone calls. The Cobblestone research team adapted to the low level of responses and sought additional ways to obtain confirmation data, such as public records information requests and adding verification questions to a survey for respondents to complete at their convenience. The verification process was extensive, with at least five phone calls and five email contacts per school, until site verification was completed in March 2022. Once usage was verified, 2022 STAAR reading data were collected to serve as an outcome measure.

FINAL STUDY SAMPLE

Through the verification process, the Cobblestone research team was able to properly place each site either into the treatment group (based on use of *Into Reading*) or the control group (based on use of a different reading program), or the team was able to remove the site (based on state closures or schools declining to participate in the verification process). Rather than removing schools based on their initial condition placement, school sites were moved to the appropriate group based on the information confirmed in the site verification process (e.g., the district bought *Into Reading* and decided to not implement the curriculum). Through this process, 316 treatment and control schools were verified, surpassing the requirement of 100 schools to detect a medium effect of at least .10. (See Appendix B for the final sample demographic characteristics.)

MEASURES

STATE OF TEXAS ASSESSMENTS OF ACADEMIC READINESS (STAAR)

The State of Texas Assessments of Academic Readiness (STAAR) has been in use since Spring 2012 to measure the Texas Essential Knowledge and Skills (TEKS) curriculum standards in math, reading, and language arts in grades 3–8, as well as an end-of-course assessment for Algebra I, English I, and English II. The tests are vertically scaled in grades 3–8 to allow for direct comparison of student test scores across grade levels within a content area.

The present study uses ELA assessment data from the STAAR. STAAR performance standards relate levels of test performance to the expectations defined in the TEKS. Cut scores established by the agency distinguish between performance levels, or categories (Masters Grade Level, Meets Grade Level, Approaches Grade Level, and Did Not Meet Grade Level).

RESULTS

The Cobblestone research team first examined whether there were overall demographic differences between treatment and control schools, including percentage of students with financial need, percentage of Hispanic/Latino students, percentage of English language learners, and total school enrollment numbers (see Table 1). Only total enrollment emerged as representing a significant difference, such that treatment schools had a lower overall enrollment ($M = 483.41$, $SD = 196.19$) than control schools ($M = 562.09$, $SD = 198.44$). Therefore, the analysis of school-level average scaled reading scores controlled for total enrollment. In addition, based on the findings from the baseline equivalency analyses (see Appendix C), teaching modality (i.e., in-person, virtual, or hybrid) was also controlled for in the outcome analyses of school-level average scaled reading scores.

TABLE 1. MEAN COMPARISON OF KEY DEMOGRAPHIC VARIABLES

Variable	Treatment (n = 140) Mean (SD)	Control (n = 176) Mean (SD)	t (194)	p	g
Hispanic/Latino	51.2% (30.3%)	46.7% (28.0%)	1.39	.17	.16
English Learner	14.1% (10.3%)	13.6% (8.2%)	0.50	.63	.06
Eligible for Free or Reduced-Price Lunch	63.2% (23.8%)	59.3% (27.2%)	1.33	.18	.15
School Enrollment*	483.4 (196.2)	562.1 (198.4)	3.52	<.001	.40

Note: p values less than .05 are considered to represent a statistically significant mean difference. Outcomes with a statistically significant difference are denoted with an asterisk (*).

COMPARISON OF STAAR SCALED READING SCORES

The purpose of the first set of outcome analyses was to examine potential changes in average scaled reading scores between 2020–2021 to the following academic year, 2021–2022. The analyses included a comparison of standardized STAAR scaled reading scores, from pretest (i.e., 2020–2021) to posttest (i.e., 2021–2022). School-level data was examined for two grade levels: (a) Grade 3 students' pretest scores (i.e., 2020–2021) compared to their posttest scores (i.e., 2021–2022, now Grade 4), and (b) Grade 4 students' pretest scores (i.e., 2020–2021) compared to their posttest scores (i.e., 2021–2022, now Grade 5).

An examination of 2021–2022 (i.e., posttest, Grade 4) STAAR school-level scaled reading scores across conditions, while controlling for 2020–2021 (i.e., pretest, Grade 3) scores, indicated that *the Into Reading* (treatment) schools had higher average scaled scores at posttest (F [1, 184] = 4.81, p = .03; see Table 2]. These differences were statistically significant, such that students in the *Into Reading* (treatment) schools had higher average scaled reading scores at posttest compared to students in the non-*Into Reading* (control) schools. These findings suggest that the *Into Reading* program significantly improved students' reading skills in comparison to other programs.

TABLE 2. MEAN COMPARISON OF SCALED STAAR READING SCORES, GRADES 3 & 4

Variable	Treatment (n = 132) Mean (SD)	Control (n = 160) Mean (SD)	F (1, 184)	p	n^2_p
Grade 3 Average Reading Scaled Score (2020–2021)	1416.25 (59.41)	1406.21 (63.12)	–	–	–
Grade 4 Average Reading Scaled Score (2021–2022)	1535.76 (50.93)	1530.58 (54.23)	4.81	.03	.016

Note: p values less than .05 are considered to represent a statistically significant mean difference; controlling for Grade 3 average reading scaled score, enrollment, 2020–2021 learning modality, and 2021–2022 learning modality.

In addition, an examination of 2021–2022 (i.e., posttest, Grade 5) STAAR school-level scaled reading scores across conditions, while controlling for 2020–2021 (i.e., pretest, Grade 4) scores, indicated that treatment schools had higher average scaled scores at posttest, although these differences were not statistically significant at the p =.05 level.

COMPARISON OF STAAR READING PROFICIENCY BY LEVEL

The Cobblestone research team also compared the rates of students who were proficient (i.e., Levels 2, 3, and 4 indicate "Pass") and not proficient (i.e., Level 1 indicates "Fail") across conditions for each grade level (i.e., Grades 3, 4, and 5). Results revealed there were marginally significant results (p values between .05 and .10) for Grade 4 (p = .09), such that the percentage of Grade 4 *Into Reading* (treatment) students who did not meet the standard was notably lower than the percentage of non-*Into Reading* (control) students who did not meet the standard.

CONCLUSION

The *Into Reading* QED study was designed to determine the potential impact of the *Into Reading* program on student reading outcomes. Using a well-implemented and designed QED study that included appropriate statistical measures to select appropriate comparison (i.e., initially using propensity score matching to determine the sample, then verifying condition membership through a site verification process), the Cobblestone research team designed a study that aimed to meet the ESSA Tier 3 standards (i.e., well-designed and implemented correlational study, with statistical controls for selection bias, that demonstrates a statistically significant positive effect, and has no strong negative findings from experimental or quasi-experimental studies). Based on the analysis of the percentage of students at or above the state standard in reading for Grades 3, 4, and 5 (2019 data) that represented 85,842 students across 316 school sites, the What Works Clearinghouse (WWC) baseline equivalence standard was met if the baseline variable was to be included in the final analytical model.

This report summarized two main outcome analyses conducted to assess potential differences between treatment and control conditions on the school-level 2021–2022 STAAR reading data. The first set of analyses examined (a) the differences between Grade 3 pretest scaled reading scores (2020–2021) at the school level and their posttest scores (2021–2022, Grade 4), and (b) the differences between Grade 4 pretest scaled reading scores (2020–2021) at the school level and their posttest scores (2021–2022, Grade 5). In addition, the rates of students who were proficient (i.e., proficiency Levels 2, 3, and 4) and not proficient (i.e., proficiency Level 1) were examined across conditions for each grade level. Results revealed the comparison of standardized STAAR scaled reading scores for Grade 3 *Into Reading* (treatment) students' pretest scores (i.e., 2020–2021) compared to their posttest scores (i.e., 2021–2022) was statistically significant. **These findings suggest that the *Into Reading* program significantly improved students' reading skills in comparison to other programs.**

It should be noted that the additional analysis that examined the rates of *Into Reading* students who were proficient (i.e., Levels 2, 3, and 4 indicate "Pass") and were not proficient (i.e., Level 1 indicates "Fail") across conditions for each grade level (i.e., Grades 3, 4, and 5) *were marginally significant* (p values between .05 and .10) for Grade 4 (p = .09). This finding is interesting, as the outcome variable of interest was 2021–2022 STAAR proficiency levels, given that most students in the *Into Reading* (treatment) schools in Grades 4 and 5 would have been exposed to the *Into Reading* program the year prior (2020-2021).

Although there was only one statistically significant finding across outcome analyses, the Cobblestone research team intentionally included additional information (e.g., school-level demographic information, teaching modality, enrollment rates) to account for contextual and environmental differences in assessing potential differences in student reading skills. It is noteworthy that the study was conducted after a long absence of normal schooling, brought on by the COVID-19 pandemic and subsequent changes to teaching modalities and lack of other normal practices. Consequently, we are far from understanding the longer-term impacts that occurred in schools as part of the COVID-19 pandemic, specifically as it relates to students' reading abilities and standardized test scores.

In the current study, results indicate a marginal significance for Grade 4 when examining the dichotomous pass/fail rates but not for Grades 3 or 5. It should also be noted that Grade 5 was approaching the threshold for marginal significance (p = .11). This may be due to younger students (i.e., Grade 3) being more sensitive to the effects of the COVID-19 pandemic as they learned to read compared the older students (i.e., Grades 4 and 5) who had already developed reading foundations. A recent study by Relyea and colleagues (2022) exploring the impact of COVID-19 on the reading achievement growth of Grade 3–5 students supports this notion. The researchers found that Grade 3–5 students had lower reading achievement gains during 2020–2021 compared to the pre-pandemic school year (2018–2019), with especially reduced reading gains among Grade 3 students. An alternative possibility is that there is a cumulative effect of the *Into Reading* curriculum that is not observed after one year (comparing Grades 3 and 4); however, the effect can be observed after two years, which would only be possible at Grades 4 and 5. Therefore, there is a need for additional research to examine the effects of age and grade level in light of the COVID-19 pandemic and other potential factors.

REFERENCES

Austin, P. C. (2010). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics*, *10*(2), 150–161. https://doi.org/10.1002/pst.433

Eddy, R. M., Alchehayed, A., Zucker, S., & Mendelsohn, D. (2023). *Houghton Mifflin Harcourt Into Reading Texas study report*. Cobblestone Applied Research & Evaluation, Inc.

Evidence for ESSA. (2019). *A2i professional support system*. https://www.evidenceforessa.org/programs/reading/elementary/a2i-professional-support-system-formerly-individualizing-student

National Center for Education Statistics. (2022). *National Assessment of Education Progress 2022 Report Card: 2022 NAEP Reading Assessment*. https://www.nationsreportcard.gov/reading/nation/groups/?grade=4

Relyea, J. E., Rich, P., Kim, J. S, & Gilbert, J. (2022). The COVID-19 impact on reading achievement growth of Grade 3–5 students in a U.S. urban school district: Variation across student characteristics and instructional modalities. *Reading and Writing*, *36*, 317–346. https://doi.org/10.1007/s11145-022-10387-y

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, *39*(1), 33–38. https://doi.org/10.2307/2683903

Thoemmes, F. (2012). *Propensity score matching in SPSS*. University of Tübingen. November 11, 2016, from http://arxiv.org/abs/1201.6385

U.S. Department of Education. (n.d.). *Using the WWC to find ESSA tiers of evidence*. Institute of Education Sciences, What Works Clearinghouse. https://ies.ed.gov/ncee/wwc/essa

Warner-Griffin, C., Liu, H., Tadler, C., Herget, D., & Dalton, B. (2017). *Reading achievement of U.S. fourth-grade students in an international context: First look at the Progress in International Reading Literacy Study (PIRLS) 2016 and ePIRLS 2016* (NCES 2018-017). U.S. Department of Education. https://nces.ed.gov/pubs2018/2018017.pdf

Wilkerson, S., & Savoy, M. (2013). *National Geographic Learning's Reach for Reading program: An efficacy study*. Magnolia Consulting, LLC.

APPENDIX A

PROPENSITY SCORE MATCHING PROCEDURE

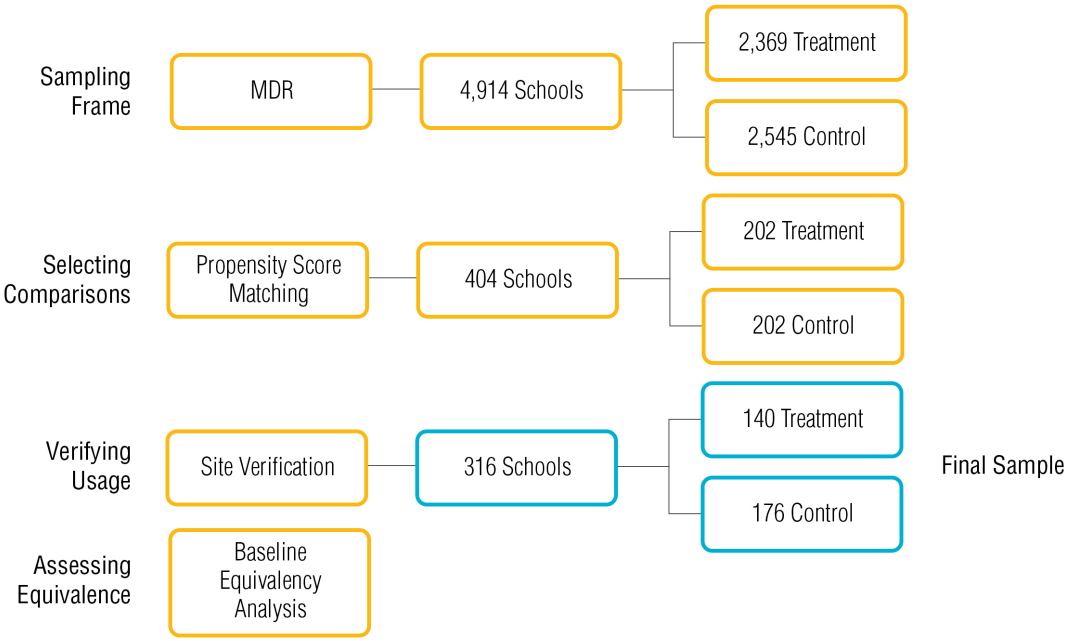
Propensity scores were in part used to select an appropriate, comparable control group. Propensity score matching is a technique used to balance meaningful covariates across treatment and control groups (Rosenbaum & Rubin, 1985). It is essential to ensure the covariates of treatment and control schools are balanced because having a larger proportion of schools with a particular covariate in one group could influence the overall results of the analysis. Thus, propensity score matching was used to match control schools to the treatment schools based on school classification (e.g., K–6), school enrollment, and available demographic information (i.e., SES based on average family income and percentage of students with financial need, percentage of white and Hispanic/Latino students, and percentage of English Language Learner students enrolled), as well as neighborhood lifestyle (e.g., city centers), type of community (e.g., large city), and the percentage of students at or above the state standard in Reading in Grades 3, 4, and 5, as well as the percentage of students at or above the state standard in Writing in Grade 4, from MDR. It should be noted that, as school classification, neighborhood lifestyle, and community type were categorical variables, they were set to require an exact match in the procedure.

There were 4,914 (2,369 treatment and 2,545 control) schools with MDR data. However, some schools were missing data on the matching variables and were removed. This left the procedure with 3,103 (1,530 treatment and 1,573 control) schools. Propensity score matching was conducted using the Propensity Score Matching for SPSS custom dialog, version 3.0 (Thoemmes, 2012). This custom dialog uses logistic regression as the estimation algorithm and uses nearest neighbor as the matching algorithm. Each treatment school was hoped to be matched to a control school. A caliper width of 0.2 standard deviations of the logit of the propensity score was set to exclude bad matches (Austin, 2010). Balance statistics produced by the SPSS custom dialog and chi-square tests of group differences were used to assess the balance of covariates across matched treatment and control schools.

It should be noted that, for the best possible outcomes for propensity score matching, a much larger pool of control schools should be present to match from. However, as we would be narrowing the results down to only 400 (200 treatment and 200 control) schools, it was deemed appropriate to try and match the entire sample. An overview of the sampling selection process can be found in Figure 1.

Using the custom dialog, 553 treatment schools were matched with 553 control schools. After matching, we examined the balance of all observed covariates, interactions among all covariates, and quadratic terms of covariates. We next selected the 404 (202 treatment and 202 control) schools that represented the best matches—that is, those matches that had the smallest differences between their propensity scores. Differences in propensity scores in each matched pair range from <.001 to .056 in this sample. Chi-square tests and ANOVAs were also run to assess the balance of covariates after matching. No significant differences were found across the treatment and control schools for any of the matching criteria. Although this is not a standard procedure, it is anticipated that many of these schools would not be included in the final sample. As such, selecting the schools that provided the best matches helps to ensure that the groups are as equivalent as possible at baseline. The Cobblestone research team still deemed it necessary to assess baseline equivalence once the final treatment and control schools were selected (see Appendix C).

Figure 1. Overview of Sample Selection Process



APPENDIX B

DEMOGRAPHIC INFORMATION, FINAL SAMPLE

TABLE 3. ETHNICITY BY GRADE

		GRADE LEVEL					
		3		4		5	
Condition & Number of Students		Treatment 11,802	Control 16,023	Treatment 12,063	Control 16,591	Treatment 12,470	Control 16,893
Ethnicity	Hispanic/Latino	52.8% (6,351)	44.8% (7,177)	53.5% (6,459)	45.5% (7,554)	53.6% (6,678)	45.5% (7,693)
	White (non-Hispanic)	33.7% (3,973)	32.4% (5,196)	33.4% (4,023)	31.9% (5,284)	33.4% (4,160)	31.8% (5,367)

Note: Total Students Grades 3–5 N = 85,842; 140 treatment schools; 176 control schools; weighted average by grade-level enrollment.

TABLE 4. ENGLISH LEARNER, SPECIAL EDUCATION, AND STUDENT NEED STATUS

Condition & Number of Students	Treatment 67,677	Control 98,928
English Learner Status	15.1% (10,219)	13.8% (13,652)
Special Education Status	10.4% (7,038)	10.4% (10,288)
Eligible for Free or Reduced-Price Lunch	67.0% (45,323)	59.6% (58,913)

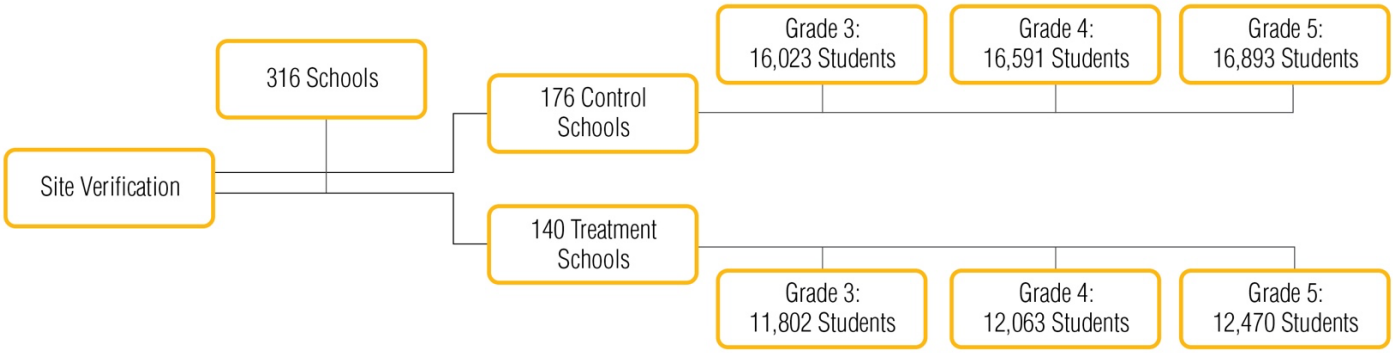
Note: Total School Enrollment N = 166,605; 140 treatment schools; 176 control schools. EL and SPED weighted averages by school-level enrollment; English learner and special education status provided at district level, estimated at the school level for this table. *Eligibility for free or reduced-price lunch provided at school level.

APPENDIX C

BASELINE EQUIVALENCE ANALYSIS

To assess the impact of *Into Reading* on student reading achievement, the Cobblestone research team sought to compare treatment and control school-level scores on the STAAR reading assessment. The outcome variable used was the 2019 percentage of students at or above the state standard in reading for Grades 3, 4, and 5 from the MDR. This variable was chosen for consistency, as it was used it in the propensity score match detailed in Appendix A. In addition, the *Into Reading* program was implemented in 2019–2020; therefore, using data from 2018–2019 provided a true baseline. Moreover, selecting data from 2018–2019 provided a baseline prior to the COVID-19 pandemic (in which teaching modality would also be controlled for in the outcome analysis). Finally, 2020–2021 school-level STAAR data was controlled for in the final outcome analyses. Across the 316 sites, school-level percentages of students at or above the state standard for 27,825 Grade 3 students, 28,654 Grade 4 students, and 29,363 Grade 5 students were analyzed. Two additional schools were only using the program in Grade 3 and so were only included in the Grade 3 analytic sample.

FIGURE 1. BASELINE EQUIVALENCE STUDY SAMPLE



Per the WWC standards, 2018–2019 baseline percentages of students at or above the state standard in reading were assessed to ensure that the treatment and control groups were equivalent prior to using *Into Reading*. WWC guidelines for assessing equivalence are based on absolute effect sizes, not statistical significance. To ensure that the analyses met WWC standards, the Cobblestone research team compared the effect size of the differences in pretest percentage of students at or above the state standard for reading (2018–19) between treatment and control to the WWC baseline equivalence thresholds. For the indicator of interest, if the effect size (Hedge's *g*) of the mean difference between treatment and control variables is less than 0.05, it is considered to have met the standard for baseline equivalence. If the effect size is greater than 0.05 standard deviations but less than 0.25 standard deviations, the baseline equivalence standard is met when the baseline measure is included in the final analytical model that assesses the relationship between condition and outcomes. The purpose of this is to control for differences between conditions at baseline. If the effect size is greater than 0.25, the baseline equivalence standard is not met.

GRADE 3 BASELINE EQUIVALENCE

A total of 140 treatment schools and 176 control schools with valid data were included in the baseline equivalence analysis. A *t*-test comparing conditions on the mean percent of student reading proficiency had an effect size of 0.13 (see Table 5). This means the baseline equivalence standard would be met if the baseline variable, 2018–2019 percentage of students at or above the state standard in reading, was included in the final analytical model.

GRADE 4 BASELINE EQUIVALENCE

A total of 140 treatment schools and 176 control schools with valid data were included in the baseline equivalence analysis. A *t*-test comparing conditions on the mean percent of student reading proficiency had an effect size of 0.13 (see Table 5). This means the baseline equivalence standard would be met if the baseline variable, 2018–2019 percentage of students at or above the state standard in reading, was included in the final analytical model.

GRADE 5 BASELINE EQUIVALENCE

A total of 140 treatment schools and 175 control schools with valid data were included in the baseline equivalence analysis (note, one Grade 5 control school was deleted from the control sample, as it was missing pretest data). A *t*-test comparing conditions on the mean percent of student reading proficiency had an effect size of 0.20 (see Table 5). This means the baseline equivalence standard would be met if the baseline variable, 2018–2019 percentage of students at or above the state standard in reading, was included in the final analytical model.

TABLE 5. BASELINE EQUIVALENCE RESULTS FOR GRADES 3–5, 2018–2019 STUDENT READING PROFICIENCY

Grade Level	Mean % Proficient (SD)		Hedges *g*
	Treatment	Control	
Grade 3	41.17 (15.90)	43.26 (15.67)	0.13
Grade 4	41.64 (15.90)	43.63 (15.72)	0.13
Grade 5	48.50 (16.90)	51.84 (16.21)	0.20

Note: Grade 3 treatment sample = 140; control sample = 176; Grade 4 treatment sample = 140; control sample = 176; Grade 5 treatment sample = 140; control sample = 175

EXPLORATORY ANALYSIS OF MODALITY

Potential differences in modality across the 2020–2021 and 2021–2022 academic years was also assessed. Specifically, the comparison included the percentages of schools that were offering in-person, hybrid, and online modalities. Chi-square analyses were conducted to explore whether these differences were statistically significant. The results indicated that there were significant differences in modality for both the 2020–2021 and 2021–2022 academic years. In 2020–2021, there were a greater proportion of treatment schools offering a hybrid modality than control schools, $\chi 2(2) = 7.18$, $p = .03$. In 2021–2022, there were a greater proportion of treatment schools offering in-person modality than control schools, $\chi 2(1) = 5.02$, $p = .03$.

BASELINE EQUIVALENCE SUMMARY

Based on the analysis of the 2018–2019 percentage of students at or above the state proficiency in reading, the WWC baseline equivalence standard is met, and the baseline variable could be included in the final analytical model assessing outcomes (i.e., 2022 STAAR reading scores). Thus, the design of the study was sufficient to meet ESSA Tier 3 standards, pending a significant difference in the posttest variable of interest (i.e., 2022 STAAR reading scores) for treatment and control sites.