Effectiveness of McGraw-Hill's *Treasures* Reading Program in Grades 3 – 5

October 21, 2010

Research Conducted by Empirical Education Inc.

Executive Summary

Background. Cognitive demands on student knowledge increase in middle elementary grades as students become primarily engaged in reading to learn, rather than learning to read (Chall, 1983). At the same time, children may lack general vocabulary as well as vocabulary related to academic concepts that enable them to comprehend what they are reading and to acquire content knowledge (Hart & Risley, 1995). Moreover, due to the Department of Education's emphasis on high-stakes testing and accountability in math, language arts, and science, the importance of reading comprehension and proficiency has never been greater for students, teachers, and school administrators. Because high-stakes, standardized tests are inherently reading tests, students who struggle with reading comprehension and proficiency will most likely struggle with these tests.

Treasures, developed by McGraw-Hill, is a "research-based, comprehensive" basal reading program for students in grades K – 6 that combines "explicit instruction and ample practice [to] ensure students' growth in reading proficiency….Each week's lesson integrates grammar, writing, and spelling for a total language arts approach" (McGraw-Hill, 2009). The *Treasures* Reading Program materials consist of leveled readers as well as online workbooks and activities. This report presents the first study of the effectiveness of *Treasures* based on an extensive sample of student performance records. This study includes grades 3 – 5 following prior research spanning grades K – 3.

This study tests two research questions:

- Do students in *Treasures* Reading Program schools achieve higher reading achievement scores than similar students in comparable schools who are not participating in the *Treasures* Reading Program?
- Are there discernible differences in the size of impact on children of different gender, ethnicity, and pre-test score?

Study Description.

Study Design. The study uses a quasi-experimental comparison group design to test the effectiveness of *Treasures* by comparing outcomes for students who used *Treasures* to those who did not use *Treasures*, adjusting for differences between the *Treasures* and comparison groups on baseline characteristics.

Sample Size and Selection. The *Treasures* group was formed through a multi-step process that began with multiple sales lists, provided by McGraw-Hill, of schools and districts that had purchased the *Treasures* Reading program. The list of potential *Treasures* schools was narrowed down to include only schools that participated in Northwest Evaluation Association's (NWEA) Measures of Academic Progress (MAP) Reading Test program and that began implementing *Treasures* during Fall 2006 or Fall 2007 in grades 3 – 5.

The comparison group was formed by selecting schools from the same or neighboring states as the *Treasures* schools that were well-matched to the *Treasures* schools in terms of characteristics recorded in the NCES database, including student demographics and community characteristics. The matching procedure resulted in a comparison group that differed from the *Treasures* group by no more than 0.075 standard deviation on each of the relevant background characteristics, including average pretest scores.

	Number of students	
	Treasures	**Comparison**
Grade 3	1531	1727
Grade 4	1757	2540
Grade 5	1740	2920
Total	**5028**	**7187**

The analytical sample was based on 35 *Treasures* and 48 comparison schools. Numbers of students in the final study analysis, by grade level, are shown in the table.

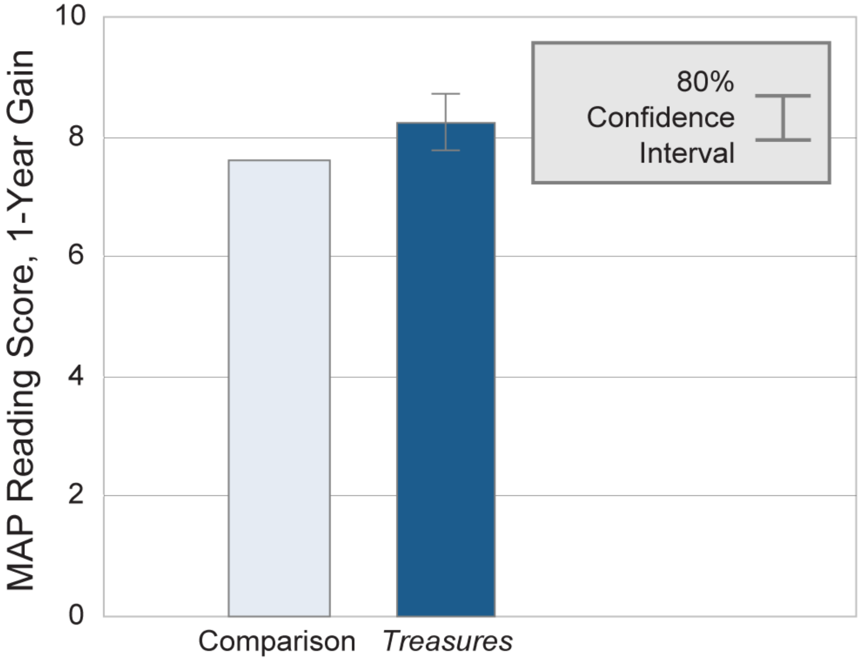
Outcome Measure and Analytical Methods. Findings of this study are based on data from NWEA's MAP Reading tests administered in 2006 – 2008. MAP Reading was selected because this standardized test facilitates consistent comparisons over time and across states and allows for producing a single estimate for participating grades at each *Treasures* and comparison school. The outcome measure used in this study was the score gain over the first year following the adoption of *Treasures*, specifically the difference between fall and spring MAP Reading scores

The data analysis was performed using analysis of covariance (ANCOVA), producing the estimates of *Treasures* impact adjusted for the differences in student-level covariates: gender, minority status, and pretest scores. In addition, moderator analyses were performed to explore subgroup differences in the effectiveness of *Treasures*.

Principal Findings. The study found that *Treasures* had a positive impact on overall elementary student literacy scores. The estimate of the program effect adjusted for the effects of covariates was found to be equal to 0.669 (effect size 0.082), with a *p* value of .031, which gives a high level of confidence in the result. A significant moderating effect of grade level was found; i.e., there was a significant difference in the effect of *Treasures* across grades, the strongest effect being observed for grade 5.

The figure below presents the estimated impact of *Treasures* comparing average outcomes for *Treasures* and comparison groups. The left bar in the diagram shows the actual average score gain for the comparison group and the right bar shows the estimated score gain for *Treasures* group adjusted for the differences between the two groups of students.

Impact of *Treasures*: Average outcomes for *Treasures* and control groups



No significant moderating effects of other student characteristics—gender, minority status, or the level of preparation (MAP Reading pretest)—were established.

This study shows that *Treasures* has a positive impact on student achievement in reading. Several factors may limit the generalizability and accuracy of the study results. The data were available for the first year of the program implementation so the study could not address longer term effects; numbers of third and fourth graders in the analytical sample were lower than the number of fifth graders; and program schools are located in only a few states, mostly in the Midwest.

Table of Contents

Background

Cognitive demands on student knowledge increase in middle elementary grades as students become primarily engaged in reading to learn, rather than learning to read (Chall, 1983). This is compounded by the possibility that children may lack general vocabulary as well as vocabulary related to academic concepts that enable them to comprehend what they are reading and to acquire content knowledge (Hart & Risley, 1995). Moreover, due to the Department of Education's emphasis on high-stakes testing and accountability in math, language arts, and science, the importance of reading comprehension and proficiency has never been greater for students, teachers, and school administrators. Because high-stakes, standardized tests are inherently reading tests, students who struggle with reading comprehension and proficiency will most likely struggle with these tests.

Treasures, a basal reading program for students in grades K – 6, was developed to address reading comprehension among students in the middle grades. *Treasures* is based on extensive research in vocabulary (Bear, D.R., 2004), comprehension (Dole, 2002; Paris, 2003), fluency (Hasbrouck, 1999), and phonics (Shanahan, 2001) and combines "explicit instruction and ample practice [to] ensure students' growth in reading proficiency" (McGraw-Hill, 2009). The *Treasures* Reading program materials consist of leveled readers, student anthologies, and listening libraries, among other products for the classroom. In addition, the program includes resources such as computer literacy lessons, spelling activities, and research and inquiry activities, which are accessible online (http://treasures.macmillanmh.com/national). "Each week's *Treasures* lesson integrates grammar, writing, and spelling for a total language arts approach" (McGraw-Hill, 2009).

This study follows a 2005 – 2006 study conducted by McGraw-Hill (MH) in association with Westat. This interrupted time series study took place in a single school, with students in grades K – 3. The Westat study, *Changes in Test Performance of Students Using the Treasures Program: A Look at an Inner City Schoo*l, found that students who used the *Treasures* Reading program made "significant gains in reading skills across the K – 3 grade range" and, according to the report, when data were examined on a "student by student basis, as opposed to the aggregate…the vast majority of students showed [this] pattern of gain (page 47)." The data for this analysis came from assessments routinely administered by the school in Fall 2005 and Spring 2006 (McGraw-Hill & Westat, 2007). Based on the study's positive results in grades K – 3, MH was interested in finding out whether a large scale research study focused on grades 3 – 5 would yield similar results. As such, MH contracted to conduct a one-year study.

The primary goal of the quasi-experimental study which is the subject of this report is to determine the effectiveness of the *Treasures* Reading Program for students across five states, in grades 3 – 5. To facilitate the comparison of students in *Treasures* schools to students in similar schools, researchers used Northwest Evaluation Association's (NWEA) Measures of Academic Progress (MAP) Reading test as the outcome measure. The MAP Reading test is vertically and horizontally scaled, which allows researchers to compare scores over time and across grades, as well as to combine scores across grades to produce a single estimate of average performance for each school. The MAP test also allows scores to be compared or aggregated across states.

This study tests two research questions:

- Do students in *Treasures* Reading Program schools achieve higher reading achievement scores than similar students in comparable schools who are not participating in the *Treasures* Reading Program?
- Are there discernible differences in the size of impact on children of different gender, ethnicity, and pre-test score?

Methods

Study Design

This study uses a quasi-experimental comparison group design, where the effectiveness of the *Treasures* Reading program is estimated by comparing achievement of students who used the *Treasures* Reading program to achievement of students who did not use the *Treasures* Reading program, adjusting for differences between the two student groups on baseline characteristics. The study focuses on reading comprehension performance of students in grades 3 – 5 as measured by their NWEA MAP Reading scores.

Quasi-experimental studies are used to estimate a program's impact in situations where the program has already been implemented and the data pertaining to its impact may already be available for analysis and/or when randomized assignment is impossible or undesirable. The major challenge in a quasi-experimental study is that program and comparison groups may systematically differ in terms of background characteristics that affect performance. Unless the estimate of the program effect is adjusted for the effects of these covariates, the result will be inaccurate, or biased. To minimize the bias in a quasi-experimental study, a sample selection procedure has been developed and implemented that ensures the program and comparison groups possess statistically similar properties. In addition, an analytical strategy is employed—analysis of covariance (ANCOVA)—that further reduces the potential for bias by controlling for the effects of important covariates, including pretests, that can lead to biased results if they are confounded with the program or comparison status of students and schools.

The unit of analysis in this study is the student. The outcome measure used in the study is the individual score gain over the first year following the adoption of the program; i.e., the difference between the fall and spring MAP Reading scores. The estimation of program impact was performed controlling for two pretest measures—the score gain over the pre-program year and the spring reading score (i.e., reading proficiency level achieved immediately prior to the program's implementation). The level of assignment was assumed to be the school; i.e., it is assumed that the program was adopted at the school level. The analysis therefore had to account for the clustering of students by school.

Sample Selection

Identification of the *Treasures* Group

Identifying and selecting the *Treasures* group was a multi-step process. In late 2008 and early 2009, MH provided researchers with multiple sales lists (one list for cohort 1 and one list for cohort 2) of districts and individual schools that had purchased the *Treasures* Reading program. Schools in cohort 1 had begun implementation of *Treasures* in Fall 2006 and schools in cohort 2 had begun in Fall 2007.

NWEA then identified the districts and schools from those lists that had administered the MAP Reading test to students in grades 3 – 5. From this narrowed list, the researchers selected potential districts and schools for cohort 1 and cohort 2 based on the criteria detailed in Table 1.

Table 1. Selection Criteria and Rationale for Potential *Treasures* Districts and Schools

Criteria	Rationale
States that contained both *Treasures* districts and comparison districts	More accurate and efficient matching of schools
Districts that had a high textbook/student ratio	Select only schools where all students could receive the *Treasures* reading intervention
Districts that had more than one school adopt the *Treasures* program	Minimize study costs
Districts that purchased *Treasures* for use in grades 3 – 5	MH chose to focus on the upper elementary grades
Only districts that purchased *Treasures* for implementation beginning in Fall 2006 and Fall 2007	The bulk of the available data were for the two most recent years of implementation.
Random sampling among districts from states that had a disproportionate number of *Treasures* districts	Prevent the impact of states with a large number of *Treasures* schools from dominating the analysis

Seventy-seven *Treasures* schools in cohort 1 (from nine districts in six states) and 66 *Treasures* schools in cohort 2 (from 22 districts in 10 states) remained in the selection pool after this process. Subsequently, researchers contacted the remaining *Treasures* districts (and schools, if necessary) to confirm the purchase date and the start date of implementation, and to confirm that *Treasures* was implemented in grades 3 – 5. Through this investigation, researchers found that 10 districts (45 schools) either did not use *Treasures* in grades 3 – 5 or had begun implementing *Treasures* on dates outside those required for this study.

Disclosure agreements were then sent to the remaining eligible *Treasures* districts to gain access to their de-identified student MAP Reading scores. Nine districts (50 schools) declined to authorize the release of their student test scores, which again reduced the number of participating *Treasures* schools.

Twenty-three *Treasures* schools for cohort 1 and 17 *Treasures* schools for cohort 2 were enlisted, for a total of 40 *Treasures* schools in five states. At this point the research team determined that, since there were no significant differences between the two cohorts, they would be combined for analysis.

The development of the *Treasures* group is detailed in Table 2.

Table 2. Development of the *Treasures* Group over Time

Event	No. of *Treasures* schools
Total schools possible	143
Potential *Treasures* schools called to ask if they had used *Treasures* and to request access to de-identified student data	143
Sample reduction due to denied access to student data	(58)
Sample reduction due to districts that did not implement *Treasures* during required date range or required grades	(45)
Final count *Treasures* schools with pretest and posttest data	40

Identification of the Pool of Comparison Schools

The pool of potential comparison schools consisted of all elementary schools located in the same regions where the recruited *Treasures* schools were located. To limit the possible bias that could arise from geographic mismatch, researchers intended to limit the set of potential matches for *Treasures* schools to those located in the same states: Colorado, Illinois, Indiana, and New Hampshire. However, all potential comparison schools in Indiana and a large number of potential comparison schools in Illinois declined to participate in the study. On the basis of geographic proximity and information included in National Assessment of Educational Progress (NAEP) State Profiles (2010), the research team determined that Michigan schools could be added to the pool of potential comparison schools for Illinois and Indiana.

Selection of comparison schools for the analysis was performed using the following matching procedure. For each *Treasures* and potential comparison school, a data profile was created that included the following data elements from the NCES database: type of school (magnet, charter, or regular), Title I eligibility, schools size (total enrollment in grades 3 – 5), characteristics of school location (urban/rural, community size, and distance from the nearest metropolitan area),[1] student-teacher ratio, number of students enrolled in the National School Lunch Program (proxy for socio-economic status), and demographic profile of students in grades 3 – 5 (percentages of males, African American, Hispanic, White/Non-Hispanic, Asian, and Native American students).

A proximity matrix was created for the potential comparison schools using Mahalonobis metric.[2] For each *Treasures* school, the six closest matches were selected from the pool of potential comparison schools located in the same region. To achieve the maximum baseline equivalence, additional constraints were imposed on the matches. First, considering that location characteristics serve as proxies for unobserved community characteristics, analysts retained only exact matches on these variables. Second, analysts removed the matches that differed from the matched

[1] This information was derived from a composite variable ULOCAL contained in the NCES database.

[2] Mahalanobis distance metric allows obtaining a measure of similarity between objects (schools in this study) that takes into account not only the differences between the values but also adjusts for correlation between the variables that enter into the calculation of distances to avoid double-counting of similar characteristics.

Treasures schools by more than 0.25 standard deviation on any single variable. The resulting set of matches contained 61 comparison schools.

NWEA then obtained passive consent and confirmed participation in the MAP assessment program from 59 of these 61 potential comparison schools. Active consent was subsequently obtained from 54 of the 59 potential comparison schools.

Finally, NWEA provided researchers with individual student data for all remaining schools: 40 *Treasures* schools and 54 comparison schools.

The development of the comparison group is detailed in Table 3.

Table 3. Development of the Comparison Group over Time

Event	No. of comparison schools
Total schools possible	1388
Excluded due to geographic mismatch	(911)
Excluded through matching	(416)
Schools statistically matched	61
Sample reduction due to no longer an NWEA member or opted out during passive consent	(2)
Potential comparison schools called to request access to de-identified student data	**59**
Sample reduction due to denied access to data during active consent	(5)
Final count of comparison schools with pretest and posttest data	**54**

Data Sources and Collection Methods

The data for this study consist of demographic information obtained from NCES and NWEA's MAP Reading test outcomes. In addition, researchers asked districts to verify that schools had implemented *Treasures* for each cohort, the dates that implementation began, and the grades in which *Treasures* was implemented.

NCES Supplied Information

Researchers acquired the following data by district (unless noted otherwise) from the NCES website.

- Number of English Language Learners
- Number of students enrolled in the National School Lunch Program (proxy for socio-economic level)
- Number of African American, Hispanic, White/Non-Hispanic, Asian, and Native American students

- School size
- Type of school (i.e., magnet, charter)
- Category of urbanization
- Title I school status and Title I student status
- Longitude and latitude (geographical location)
- State in which district is located

In addition, researchers calculated the student-teacher ratio for each school based on the teacher full-time equivalent and whole school population provided by NCES.

Achievement Measures

A single outcome measure was used in this study: NWEA MAP Reading test scores. There is one MAP Reading test for students in grades 2 – 5: Reading Goals Survey 2 – 5 (versions two through four); and one MAP Reading test for students in grade 6: Reading Goals Survey 6+ (versions two through four). The various versions represent only changes for adherence to the Illinois content standards. Scores from different versions can be compared because they have been put on the same (RIT) scale (NWEA, personal communication June 22, 2010).

Testing Schedule

The pretests were administered in Fall 2006 and Fall 2007, and the posttesting took place in Spring 2008 and Spring 2009, as shown in Table 4.

Table 4. Testing Schedule

Milestone	Date
Cohort 1 NWEA MAP Reading pretest	Fall 2006
Cohort 2 NWEA MAP Reading pretest	Fall 2007
Cohort 1 NWEA MAP Reading posttest	Spring 2008
Cohort 2 NWEA MAP reading posttest	Spring 2009

Researchers requested NWEA MAP Reading test data from multiple testing periods for every student in each condition and cohort. However, only a minority of participating schools test their students during the winter testing period. By contrast, fall and spring testing was performed systematically. Therefore, only fall and spring test scores were retained in the analytical dataset. Table 5 below shows the MAP Reading test data that were requested from NWEA, by testing period.

Table 5. MAP Reading Test Data Requested, by Testing Period

Cohort 1	Cohort 2
Fall 2006	Fall 2007
Winter 2006	Winter 2007
Spring 2007	Spring 2008
Fall 2007	Fall 2008
Winter 2007	Winter 2008
Spring 2008	Spring 2009

Analytical Methods

Data analysis was performed using analysis of covariance (ANCOVA), a commonly used method for quasi-experimental designs. ANCOVA allows estimating the effect of a program on student performance, adjusting for systematic differences between the program and comparison groups on baseline covariates. The adjustment was performed by including in the analysis the following individual characteristics (covariates): student gender, minority status, and two pretest measures: MAP Reading score gain in the year prior to the adoption of *Treasures* and Spring MAP Reading test score prior to the adoption of *Treasures*.

The outcome measure used in this study was the score gain over the first year following the adoption of the program; i.e., the differences between fall and spring MAP Reading test scores. This approach allows adjusting for unobserved differences between students by eliminating ("differencing out") the confounding effects of student and school-level factors that do not change over the period of the study. The score gain approach makes efficient use of the information contained in all four individual measurements recorded (for most students) in the NWEA data that were made available to the study team: fall and spring test scores in the years before and after the adoption of the program.

Program impact estimates were obtained from a linear mixed model as implemented in *R*-language package *lme4* (Bates, 2010). The model had hierarchical (two-level) structure to account for clustering of students (the unit of analysis) in schools (the unit of program assignment). Variation in the statistical relationship between the program effect, covariates, and the outcome measure across schools was modeled using a random effects approach.

In addition to the estimation of the average treatment effect, moderator analyses were performed to explore whether there are subgroup differences in the effectiveness of *Treasures*. Four moderator variables were considered: grade level, gender, minority status, and pre-program spring score. The following sections report estimates of the average program effect and the results of moderator analyses.

Analytical Sample Size and Characteristics

The analytical strategy based on the use of score gains required that four test scores were available for each student. In addition, only students who normally progressed from one grade level to the next were included in the sample to limit the impact of anomalous cases on the estimates. These additional requirements, in conjunction with limitations in data availability (missing data),[3] reduced the sample used in the analysis to 35 *Treasures* and 48 comparison schools with a total of 12,215 students (7187 and 5028 respectively).[4] The breakdown of students in the analytical sample is given in Table 6.

Table 6. Size of the Analytical Sample

	Number of students	
	Treasures	**Comparison**
Grade 3	1531	1727
Grade 4	1757	2540
Grade 5	1740	2920
Total	**5028**	**7187**

This analytical sample is characterized by the baseline equivalence according to WWC guidelines (Institute of Educational Studies, 2008, p.15): the mean differences between student characteristics in the *Treasures* and comparison groups were under 0.05 of their standard deviations in all cases except one—percent minority—where it exceeded 0.05 standard deviation but was still well under the threshold level of 0.25 standard deviation (see Table 7).[5] Both unadjusted and adjusted estimates of the program effect are reported. The latter of these statistically adjusts for the small imbalance between the *Treasures* and comparison group on the covariates, and thereby eliminates any inaccuracy attributable to this imbalance.

[3] In use here is listwise deletion of cases (student records) that have missing values for the outcome measure or any of the covariates. This is one of the standard and well-performing approaches to handling missing data.

[4] At the study design stage, the assumption was made that it would be possible to find three matched comparison schools for each program school. With this proportion of schools in each condition, it was calculated that 37 Treasures schools and 111 comparison schools would be required to detect an effect size as small as .15 standard deviation units in magnitude, assuming the covariates account for 64% of the variance in the school average of student posttest scores, an unconditional intraclass correlation coefficient of .20, 150 students per school, a type-1 error rate of .05, and statistical power of 80%. Although the sample sizes achieved were smaller than anticipated, the statistical power was not adversely affected: assuming the other parameter values do not change, with the achieved sample size, the minimum detectable effect size is 0.18. In fact, as the results presented in this report demonstrate, it was possible to detect a smaller effect size.

[5] There was no need to model the grade level explicitly. If there is any imbalance on grade, it will not bias the result because the pretest is vertically scaled and the effect of the pretest is factored into the analysis.

Table 7. Characteristics of the Analytical Sample

	Mean, treatment	Mean, comparison	Standard deviation, pooled	Difference between means in units of standard deviation
Percent male	50.38	52.08	49.98	0.034
Percent minority	35.72	32.16	47.24	0.075
Average pretest, fall	193.56	193.75	16.93	0.011
Average pretest, spring	202.85	203.00	14.62	0.010

Results

Average Program Effect

The study found that *Treasures* had a positive impact on the overall elementary school student literacy scores.

The unadjusted average program effect was 0.717 test score points, which corresponds to 0.088 effect size. [6] The *p* value for this result was less than .001, which means that we are very unlikely to observe a difference this large due to chance variation when there is no difference. One can have a high level of confidence that there is a positive impact of the program and that the observed differences are not just due to chance.

Adjusted program effect was found to be equal to 0.669 (effect size 0.082), with the *p* value of.031,[7] which also implies a high level of confidence in the result. We performed a series of sensitivity analyses to determine if the estimate varies substantially depending on the specification of the model (such as inclusion of additional interaction terms and nonlinear functions of covariates) and the exclusion of one or more states from the analytical sample and found that the estimate remains positive and statistically significant.

[6] Following WWC guidelines (Ibid, p.17), Hedges effects size are reported, which equals to the raw treatment effect divided by the unadjusted within-group standard deviation of the outcome measure (which equals 8.17 for the outcome measure used in this study). Effect size is a measure that is independent of the units of measurement of the outcome and compares the program effect against the magnitude of the variability of the outcomes.

[7] This p value is obtained from one-tail *t* test and equals the probability of observing an effect of this magnitude or greater when the true effect is zero or negative. Generally, the lower the p value the higher the confidence that the effect is not zero or negative. It is customary to report *p* values from two-tail *t* tests which measure the probability of observing an effect equal to or greater than the magnitude of the one observed when the true effect size is zero. One-tail *t* tests are appropriate in program evaluation studies in which the primary interest is whether there is a positive effect of the program, and where zero or negative outcomes have the same implications for the decision-making process (a new program should not be adopted in either case) and when prior research suggests that the program should have a positive effect. Both conditions are met in this study.

Table 8. Average Effects of *Treasures*

	Effect size	Percentile gain	*p* value (1-tail *t* test)	*p* value (2-tail *t* test)
Unadjusted	0.088	3	<.001	<.001
Adjusted	0.067	3	.031	.062

Both the adjusted and unadjusted effect size estimates correspond approximately to a three percentile point gain. In other words, one year after the adoption of *Treasures*, a student who performs at the median of the score distribution would score three percentile points higher if she had been in the *Treasures* group rather than in the comparison group.

Figure 1. Impact of *Treasures*: Average outcomes for *Treasures* and control groups

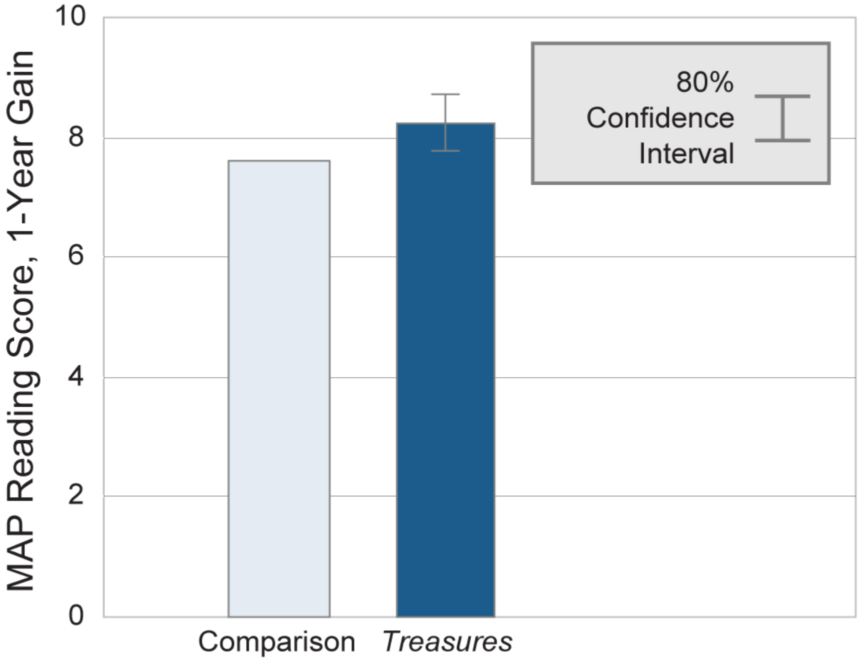


Figure 1 presents the estimated impact of *Treasures* comparing average outcomes for *Treasures* and comparison groups. The left bar in the diagram shows the actual average score gain for the comparison group and the right bar shows the estimated score gain for *Treasures* group adjusted for the differences between the two groups of students. [8]

Program Effects by Grade Level

Analysis of the outcomes by grade level showed that the program effect differs across the grade levels.[9] The program effect on fifth graders was the highest and had the lowest *p* value. Estimates for grades 3 and 4 have lower size effects (and higher *p* values) than those for grade 5, which may be due to the smaller numbers of third and fourth graders in the analytical sample. One may still have some confidence that *Treasures* has a positive effect on the fourth grade students. The results of the grade-level analysis are presented in Table 9.

Table 9. Grade-Level Effects of *Treasures*

	Effect size	Percentile gain	*p* value (1-tail *t* test)	*p* value (2-tail *t* test)
Grade 3	0.039	2	.232	.464
Grade 4	0.063	3	.103	.206
Grade 5	0.114	5	.011	.022

Figure 2 presents the estimates of differential impact of *Treasures* on students in grades three through five.

[8] The height of the *Treasures* group bar can be interpreted as the potential achievement of the comparison group students if they were all enrolled in *Treasures* classes. The confidence interval shown in the diagram represents the likely boundaries for the estimate for the impact of *Treasures*. The lower boundary of the 80% confidence interval located higher than the top of the comparison group bar indicates that *Treasures* has a positive effect on student achievement with at least 80% probability. The probability that the estimated positive effect is due to chance is at most 20%.The precise probability of positive effect (p value) is reported in Table 8.

[9] This analysis was performed using grade level as moderator of the program effect. The model also allowed for possible differences in the statistical relationship between pre- and post-test student achievement.

Figure 2. Impact of *Treasures*: Outcomes by Grade

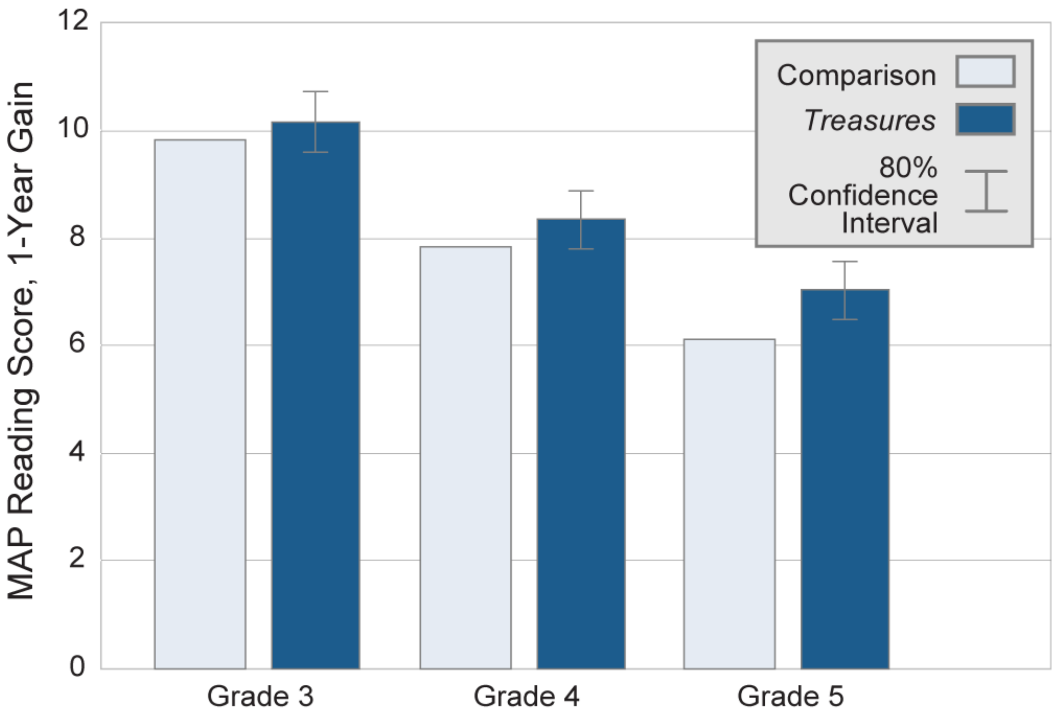


Figure 2 presents the estimated effects of *Treasures* by grade level. For each of the grade levels three through five, the left (light blue) bars show the actual average score gains for the comparison groups and the right (dark blue) bars show the estimated score gains for *Treasures* group adjusted for the differences between the two groups of students at the appropriate grade level.[10]

Moderating Effects of Student-Level Covariates

Additional moderator analyses tested whether program effects varied by gender, ethnicity (minority status), and pretest. The magnitude of the program effect did not vary by gender, minority status, or pretest. *Treasures* is therefore likely to serve differing student populations equally well.

[10] Similar to Figure 1, the heights of the *Treasures* group bars in Figure 2 can be interpreted as the potential achievement of the comparison group students if they were all enrolled in *Treasures* classes. The confidence intervals shown in the diagram represent the likely boundaries for the estimates for the impact of *Treasures* by grade. The lower boundaries of the 80% confidence intervals located higher than the top of the appropriate comparison group bars (grades four and five) indicate that *Treasures* has a positive effect on student achievement in grades four and five with at least 80% probability. For grade three, the lower boundaries of the 80% confidence intervals located lower than the top of the appropriate comparison group bar indicates that *Treasures* may have a positive effect on student achievement in grade three with a lower than 80% probability . The precise probabilities of positive effect (p values) are reported in Table 8.

Discussion

This study shows that *Treasures* has a positive impact on student achievement in the first year of its implementation and that the program effect increases with grade level, with the highest effect at the fifth grade.

The differences between the estimates of grade-level effects should be taken with caution because the number of third and fourth graders in the analytical sample was lower than the number of fifth graders, which resulted in the lower accuracy of the estimates for the earlier grade levels.

The accuracy of the estimates was possibly negatively affected by the absence of important and potentially variable student characteristics such as disability and English learner status in the dataset provided by NWEA.

The data were available for the first year of the program implementation so the study could not address longer term effects. Furthermore, geographical representativeness of this study is practically limited to the Midwest, since a vast majority of observations (all *Treasures* schools except four) in the sample came from Illinois and Indiana.

A study with a greater geographical coverage, more balanced representation of grade levels, and with a larger number of years after the program adoption would have the potential to produce more accurate estimates of the positive impact of *Treasures* on reading achievement in upper elementary schools.

References

Bates, D. (2010, June). *Linear mixed model implementation in lme4*. University of Wisconsin-Madison. Retrieved from http://cran.r-project.org/web/packages/lme4/vignettes/Implementation.pdf

Bear, D. R., & Helman, L. (2004). Word study for vocabulary development in the early stages of literacy learning: Ecological perspectives and learning English. In J. F. Baumann & E. J. Kame'enui, (Eds.), *Vocabulary instruction: Research to practice*. New York: Guilford Press.

Bloom, H. S., (2005). Randomizing groups to evaluate place-based programs. In H. S. Bloom (Ed.), *Learning More from Social Experiments*. New York, NY: Sage.

Chall, J. (1983). *Stages of Reading Development.* Fort Worth, TX: Harcourt-Brace.

Dole, J. A. (2002a.). Comprehension strategies. In B. Guzzetti (Ed.), *Literacy in America: An encyclopedia, Volume I* (pp. 85-88). New York: ABC-CLIO.

Ehri, L. C., Nunes, S. R., Willows, D. M., Schuster, B. V., Yaghoub-Zadeh, Z., & Shanahan, T. (2001). Phonemic awareness instruction helps children learn to read: Evidence from the National Reading Panel's meta-analysis. *Reading Research Quarterly, 36,* 250–287.

Hart, B., & Risley, T. R. (1995). *Meaningful Differences in the Everyday Experience of Young American Children*. Baltimore, MD: Brooks.

Hasbrouck, J. E., Ihnot, C. & Rogers, G. H. (1999). Read naturally: A strategy to increase oral reading fluency. *Reading Research & Instruction*, *39*(1), 27–38.

MacMillan McGraw-Hill (2009). Retrieved March 2009 from http://activities.macmillanmh.com/reading/treasures/

National Assessment of Educational Progress (2010). Retrieved June 2010 from http://nces.ed.gov/nationsreportcard/states

National Center for Education Statistics (2009). Retrieved April 2009 from http://www.nces.ed.gov/datatools/index.asp?DataToolSectionID=4

Northwest Evaluation Association (2009). Retrieved April 23, 2009 from http://www.nwea.org/

Paris, A.H., & Paris, S.G. (2003). Assessing narrative comprehension in young children. *Reading Research Quarterly*, *38*(1), 36–76.

Institute of Educational Studies (2008, May). *What Works Clearinghouse Evidence Standards for Reviewing Studies, Version 1.0*, (p. 15). Retrieved from http://ies.ed.gov/ncee/wwc/pdf/wwc_version1_standards.pdf

Appendix

The NWEA MAP Reading Test

The outcome measure is a student-level score from the NWEA MAP Reading test. The MAP test is a computerized adaptive comprehensive assessment aligned to each state’s measurement scales and content standards, which is designed to measure growth over time. The MAP test draws on a subset (specific to state content standards) of an item bank that contains questions of varied difficulty (Northwest Evaluation Association, 2009).

NWEA tests are scored on a Rasch unIT (RIT) scale, a measurement scale developed to simplify the interpretation of test scores. This scale is used to measure student achievement and student growth on an equal-interval scale so that a change of one unit indicates the same change in growth, regardless of the actual numerical values. RIT scores typically range from about 150 to 300 and indicate a student's current achievement level along a curriculum scale for a particular subject. Since this is a continuous scale, third-grade student scores are usually found lower on the scale, whereas fifth-grade scores typically appear higher along the scale.