Effectiveness of McGraw-Hill's *Treasures* Reading Program in Grades 3–5: *A Study in Osceola Schools*

August 23, 2011

Research Conducted by Empirical Education Inc.

Executive Summary

Background. *Treasures*, a basal reading program for students in grades K–6, was developed to address reading comprehension among students in the middle elementary grades. *Treasures* is based on extensive research in vocabulary (Bear & Helman, 2004), comprehension (Dole, 2002; Paris, 2003), fluency (Hasbrouck, 1999), and phonics (Ehri et al., 2001) and combines "explicit instruction and ample practice [to] ensure students' growth in reading proficiency" (McGraw-Hill, 2009). The *Treasures* reading program "integrates grammar, writing, and spelling for a total language arts approach" (McGraw-Hill, 2009). This evaluation study sought to understand the impact of *Treasures* on reading achievement for students in grades 3-5 in one Florida school district. The results inform future implementation and development of the *Treasures* reading program.

Study Design. This study uses an interrupted time series design to study the effect of *Treasures* on student achievement in reading. The impact of *Treasures* was evaluated using student reading scores from the Florida Comprehensive Assessment Test (FCAT). In addition, moderator analyses were performed to explore whether there were subgroup differences in the effectiveness of *Treasures*. Five moderator variables were considered: grade level, gender, ethnicity, disability, and English learner status.

Source Data and Sample Selection. The primary data for this study were provided by the Osceola school district and consist of demographic information, FCAT test scores, and information on student transfers during the year (between schools within the districts and from other districts). The dataset covered five consecutive school years from 2005-06 to 2009-10, including two years prior to introduction of the intervention and three years after the introduction. To improve its quality, the sample was reduced by 13%. This reduction in sample size, however, still allowed for sufficient power to detect small program effects. The sample students in grades 3-5 including 10,192 students total in the Treasures group and 8,911 total in the control group.

Results. The study results show that *Treasures* has a positive impact on reading achievement in grades 3-5, and this result has a strong statistical significance. Moreover the study revealed that *Treasures* has a positive impact on subgroups of students for which we had data available: each grade level; boys and girls; students with and without disabilities; and English learners and native speakers. *Treasures* shows a much stronger positive impact on students with disabilities and English learners than the rest of the student population. *Treasures* also shows a stronger positive impact on girls than on boys.

Conclusion. This study demonstrates that adoption of the *Treasures* reading program benefits students in middle elementary grades. Subgroups that are deemed underperforming, namely students with disabilities and English learners, appear to especially benefit, suggesting that *Treasures* could reduce the achievement gap between these relatively underperforming groups and the rest of the student population. *Treasures* also shows a stronger positive impact on girls than on boys, actually increasing the gender achievement gap.

Table of Contents

Objective

The primary goal of this quasi-experimental study is to determine the effectiveness of the *Treasures* reading program for students in grades 3–5 in Florida's School District of Osceola County.

This study seeks to answer two research questions:

- Do students achieve higher reading scores after the introduction of the *Treasures* reading program in their schools starting in 2007-2008 school year?
- Are there discernible differences in the size of the impact of *Treasures* on the scores of students belonging to various demographic subgroups (grade level, gender, ethnicity, disability, and English learner status)?

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 mathematics, 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 elementary grades. *Treasures* is based on extensive research in vocabulary (Bear & Helman, 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. "Each week's *Treasures* lesson integrates grammar, writing, and spelling for a total language arts approach" (McGraw-Hill, 2009).

This research follows two earlier studies of *Treasures*. The first study was conducted in 2005–2006 by McGraw-Hill in association with Westat. This interrupted time series study took place in a single school, with students in grades K–3. Researchers found that students who used the *Treasures* reading program made significant gains in reading skills across the K–3 grade range (McGraw-Hill & Westat, 2007).

The second larger scale study, *Effectiveness of McGraw-Hill's* Treasures *Reading Program in Grades 3–5,* was conducted by Empirical Education Inc. (McGraw-Hill, 2010). Researchers used MAP Reading test scores from NWEA's national database, matching student records based on geographic location, student demographics, and community characteristics (NWEA, 2009). The study found that *Treasures* had a positive impact (effect size: 0.082, *p* value: .031) on student literacy scores. Additionally, there was a significant difference in the effect of *Treasures* across grades, with the strongest effect in grade 5.

Study Design and Methods

This study uses a quasi-experimental approach which is suitable when the program has already been implemented so that the data pertaining to its impact is already available for analysis while randomized assignment is impossible. *Treasures* was introduced simultaneously in all elementary schools in the Osceola school district starting in 2007-08. Therefore, this study uses an interrupted time series

design, where the effectiveness of the *Treasures* reading program is estimated by comparing student achievement in the cohorts after the introduction of the *Treasures* reading program to student cohorts in the two years prior to that. The study uses individual student records and relies on the FCAT test administered by the state of Florida as the outcome measure.

In order to adjust for differences between the successive cohorts of students and for student, teacher, and school-level effects, the estimation of the program effect was performed in the framework of a hierarchical linear model.[1] In addition to the estimation of the average treatment effect, moderator analyses were performed to explore whether there were subgroup differences in the effectiveness of *Treasures*. Five moderator variables were considered: grade level, gender, ethnicity, disability, and English learner status.

In the presentation of the results, we used the following technique to demonstrate the gains from the introduction of *Treasures*. When showing the achievement or treatment group, we used actual average achievement in years following the program adoption. For the lack of a true comparison group, we constructed a counterfactual by calculating the predicted (from the estimated model) average scores that would be achieved by the actual students had *Treasures* not been adopted. In the case of moderator analyses, this counterfactual outcome was calculated separately for each appropriate subgroup of student population.

Data Sources and Collection Methods

Sources and Collection Methods

The student-level data for this study were provided by the Osceola school district and consist of demographic information, FCAT test scores, and information on student transfers during the year (between schools within the districts and from other districts). Additionally, we inferred information on transfers between years from comparing school information in successive records for each student. Both types of transfers exhibit significant negative impact on student achievement.

The following student characteristics were provided by Osceola.

- Date of birth
- Gender
- Ethnicity (African American, Hispanic, White/non-Hispanic, Asian, Native American, or mixed)
- English proficiency (English language learner status)
- Disability status
- Enrollment in the National School Lunch Program (proxy for socio-economic status)

The lunch program enrollment data proved to be unreliable (the data for two years showed almost uniform program enrollment against the backdrop of wide variability in the other three years) and was

[1] The outcome variable (FCAT developmental score) was modeled at the student level as a function of demographic variables, average class characteristics, information on student transfers, and a binary treatment indicator set to one for the 2007-08 and later school years and set to zero for all prior years. Unobserved differences among the units of analysis were modeled using an appropriate random effect structure comprising two-level teacher-within-school effects and individual student effects. The latter was particularly instrumental in increasing the precision of estimates in this study by accounting for unobserved differences between students through the use of multiple observations (two or three successive test scores for most students). Estimation was performed using an implementation of linear mixed models in *R*-language package *lme4* (Bates, 2010).

not included in the analysis. All remaining variables were included at the student level and used to calculate class averages needed to control for possible peer effects.

In addition, we used data on state average FCAT scores over the years of the study (Florida Department of Education, 2010) and determined that statewide averages exhibited a significant upward drift. Average scores in the district exhibited a similar drift. We therefore calculated a score deflator using the state averages and applied it to the recalculation of individual scores in the dataset.

The dataset covered five consecutive school years from 2005-06 to 2009-10, including two years prior to the introduction of the intervention and three years after.

Analytical Sample Size and Characteristics

The data records were analyzed for completeness and consistency. Several minor problems were detected in the dataset including missing test scores for some of the variables used in the analysis. In addition, we excluded data for students who either repeated a grade or skipped a grade which created an ambiguity in the association of a student with a cohort. These problems resulted in reduction in the overall sample size by 13 percent. The remaining number of records allowed for sufficient power to detect small program effect.[2] The size of the analytical sample broken down by grade level is presented in Table 1.

Table 1. Size of the Analytical Sample

	Number of students	
	Comparison	*Treasures*
Grade 3	2,916	3,458
Grade 4	2,928	3,372
Grade 5	3,067	3,362
Total	8,911	10,192

Table 2 presents average characteristics of the students in the sample compared to the state averages in 2009-10. It appears that the district has a sufficiently diverse student population to allow for subgroup (moderator) analyses, although it is not representative of the Florida student population in general. The average program effect estimated in this study is therefore relevant for student populations with greater than average percentages of minority students, English learners, and students with disabilities.

[2] Post factum power analysis showed that fewer than 1000 student observations would have been sufficient to detect the lowest effect size estimated in this study.

Table 2. Characteristics of the Analytical Sample

	Grade 3		Grade 4		Grade 5	
	Study Data	State Average	Study Data	State Average	Study Data	State Average
White (%)	29.30	42.15	30.09	43.57	30.31	44.18
Black (%)	9.91	23.40	9.77	22.52	9.93	22.24
Hispanic (%)	50.46	27.55	44.20	27.17	48.94	27.06
Male (%)	50.14	51.35	49.89	50.83	50.13	51.04
Students with disabilities (%)	10.95	12.74	12.20	13.50	12.55	14.12
ELL (%)	17.62	10.50	19.06	8.75	14.68	6.75
FCAT 2010 Developmental Scale Score (mean)	1337	1386	1512	1601	1566	1649

Note. “Study data” is the average over the five years included in this study; “state averages” are for 2010 FCAT takers.

Results

Average and Grade-Level Program Effects

Estimation results show that *Treasures* has a positive effect over all and on all subgroups of students for which moderator analysis was performed. All estimates have very low *p* values—less than .001—indicating that we have very strong confidence in the results. The *p* value corresponds to the likelihood that a difference this large may have occurred when there is no actual difference. In this study, probability of this sort of error is practically zero.

Table 3 presents estimates of the average effect and grade-level effects from the analysis with grade level as a moderator variable. The lowest program impact for the third graders is consistent with the finding of the earlier multi-state study of *Treasures* effectiveness (McGraw-Hill, 2010). Unlike that earlier study, the effect estimated in this study is statistically significant.

Table 3. Estimated Effect of *Treasures:* by Average and by Grade

	Estimate	Effect size	*p* value
Average	20.67	0.06	< .001
Grade 3	13.91	0.04	< .001
Grade 4	31.93	0.11	< .001
Grade 5	18.31	0.06	< .001

The bar graph in Figure 1 demonstrates the effect of *Treasures* by comparing average actual student achievement to a counterfactual comparison group constructed as described earlier in the section "Methods and Techniques." Bar graphs in Figure 2 present a similar comparison with a breakdown by grade level. Right (dark blue) bars in each of the two diagrams show the actual average FCAT developmental scores and the left (lighter) bars show predicted average FCAT scores had *Treasures* not been adopted. Brackets on top of the bars corresponding to the *Treasures* group represent the 80% confidence intervals for the program effect estimates, demonstrating the precision of the estimate.

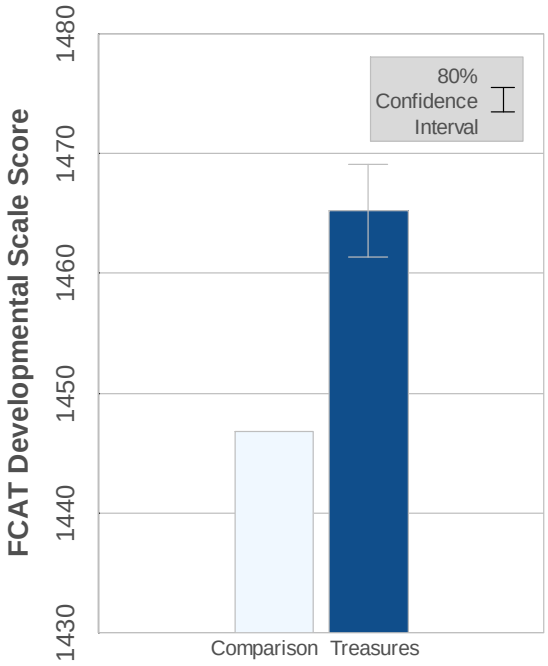

Figure 1. Average Impact of *Treasures* on FCAT Reading Assessment

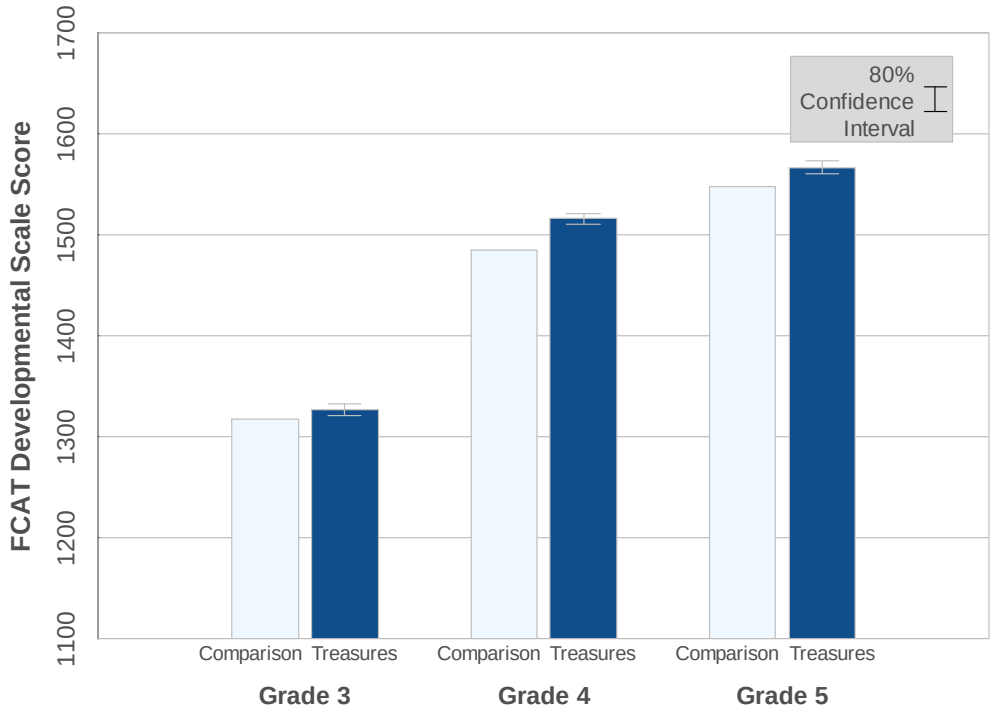

Figure 2. Impact of Treasures on FCAT Reading Assessment by Grade Level

The effect size reported in Table 3 is calculated by dividing the impact estimate (leftmost column in the table) by the standard deviation of student scores. This is a conventional approach for producing estimates that are comparable across studies, but it may not be as informative in a time series study as it is in a comparison group study. An alternative metric that allows evaluating the practical significance of program impacts can be produced by using the vertical alignment of the FCAT developmental scores. Since the test is administered only once a year in the Spring, the difference between fourth-year and third-year test scores is a measure of achievement gain in grade four; the difference between the fifth- and fourth-year tests measures achievement in grade five. These average gains, 175 and 54 points respectively for the data in the study[3], can be related to the program impact estimates in Table 3—31.9 and 18.3 respectively. Dividing the estimates by the average score gains produces estimates of the acceleration in learning due to adoption of *Treasures*—18% and 34% respectively. These numbers are very crude estimates because: a) they are based on deflated scores averaged over five years, b) they do not adjust for changes in the student population, and c) the test itself may not be perfectly aligned. They are, however, instrumental in showing that the *Treasures* impact in the fifth grade is unlikely to be lower and is *possibly higher* than the effect on fourth graders if we take into account the uneven pace of learning as expressed in FCAT developmental scores. A similar analysis cannot be performed for the third grade because of the lack of pretest; FCAT testing is not performed in the second grade.

[3] These numbers can be calculated from the FCAT numbers in the last row of Table 2.

Moderator Analysis

Moderator analyses of student demographic characteristics yielded statistically significant results for gender, disability, and English learner status. Table 4 shows the estimated *Treasures* impact moderated by these student characteristics. Ethnicity did not appear to produce a significant moderator effect and therefore, is not reported here.

Table 4. Estimated Effect of *Treasures*: Moderated by Student Gender, Disability and English Learner Status

	Estimate	Effect Size	*p* value
Disability			
Students w/o disabilities	18.2	0.06	< .001
Students with disabilities	42.2	0.09	< .001
Gender			< .001
Male	12.9	0.04	< .001
Female	28.2	0.09	< .001
English learner status			
Native and fluent English speakers	10.1	0.03	< .001
English learners	92.0	0.26	< .001

Results in Table 4 show that *Treasures* has a substantially larger impact on lower-achieving subgroups—students with disabilities and English learners—and therefore, contributes to reducing the achievement gap between these groups of students and the rest of the student population. This is the first time that this kind of result could be obtained because earlier studies had no access to the appropriate data.

At the same time, *Treasures* exhibits a greater positive effect on female students, who generally tend to achieve higher reading scores than boys. (Buchmann, DiPrete, &,McDaniel, 2008; Entwisle et al., 2007; Machin & McNally, 2006; Willingham & Cole, 1997; Cornwell, Mustard, & Van Parys, 2011; Parekh, 2011). In this dataset, boys scored about 11 percentile points lower on FCAT reading in the two years preceding the adoption of *Treasures*. This study, therefore, does not provide evidence that *Treasures* offsets the well-documented recent trend of a growing gender gap in elementary grades reading achievement. Despite these differences, the results of moderator analyses show unequivocally that *Treasures* has a significant positive impact on every student subgroup, for which the data are available.

The following bar graphs present subgroup achievement comparisons using the same approach as in the section above: right (dark blue) bars in each of the two diagrams show the actual average FCAT developmental scores and the left (lighter) bars show predicted average FCAT scores had *Treasures* not been adopted. Brackets on top of the bars, corresponding to the *Treasures* group, represent the 80% confidence intervals for the program effect estimates, demonstrating the precision of the estimate.

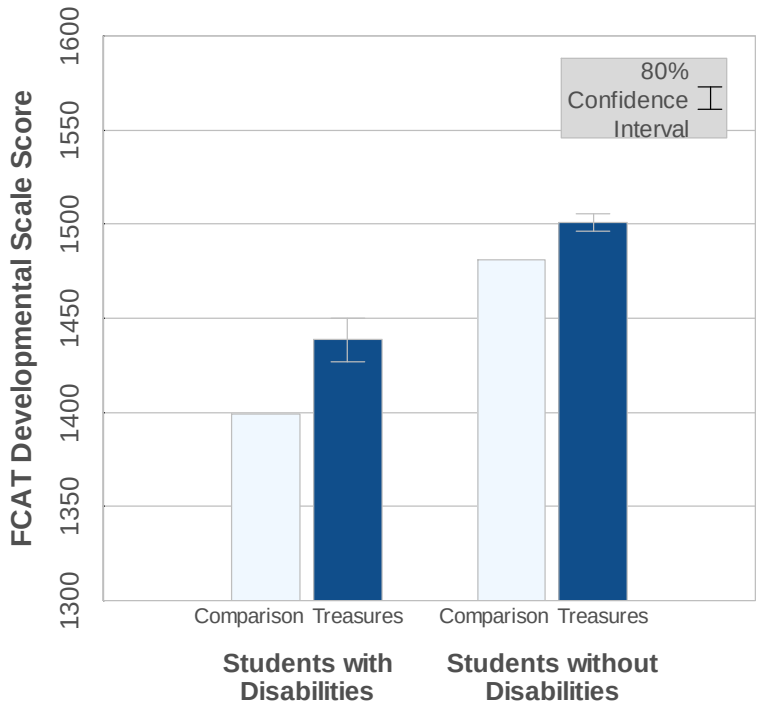

Figure 3. Estimated Effect Moderated by Disability Status

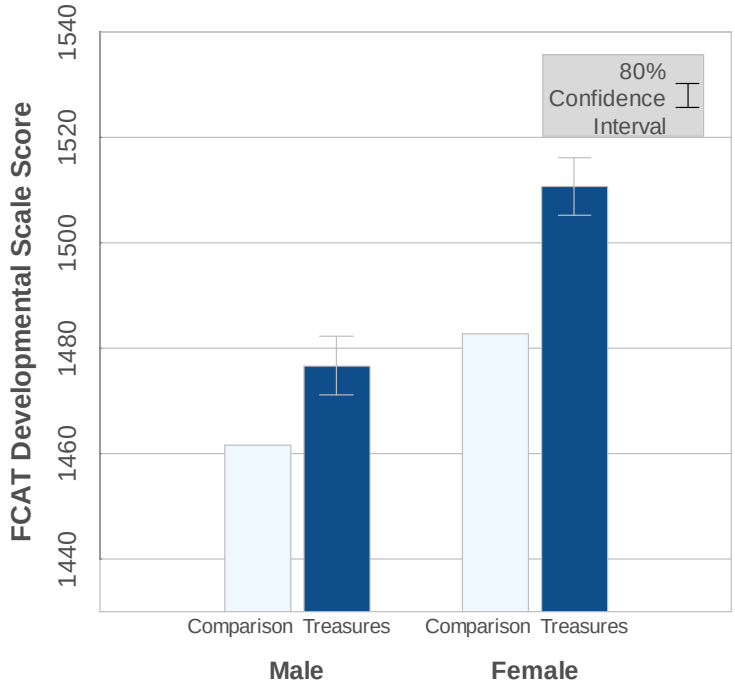

Figure 4. Estimated Effect Moderated by Gender

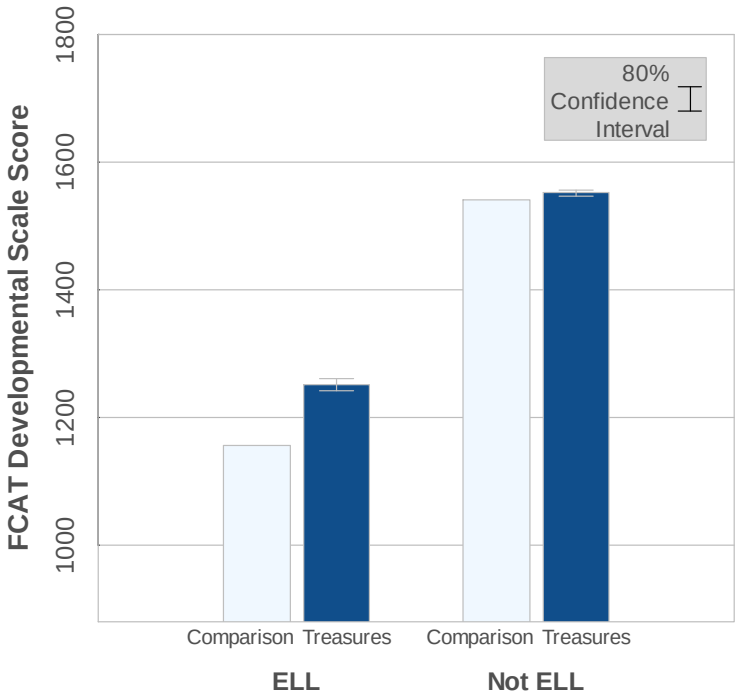

Figure 5. Estimated Effect Moderated by English Learner Status

Conclusion

This study shows that *Treasures* has a positive impact on reading achievement in grades 3-5, and this result has a strong statistical significance. Moreover, the study finds that *Treasures* has a positive impact on subgroups of students for which we had data available: each grade level, boys and girls, students with and without disabilities, and English learners and native speakers. *Treasures* shows a much stronger positive impact on students with disabilities and English learners, suggesting that its adoption could reduce the achievement gap between these relatively underperforming groups and the rest of the student population. *Treasures* also shows a stronger positive impact on girls than on boys, actually increasing the gender achievement gap.

We must be cautious in generalizing these results because they use data from only one district, which despite the diversity of its population, is not representative of the whole elementary school population—statewide or nationwide. It does provide valuable insight into the effectiveness of *Treasures* among populations with high proportions of English learners. We also need to be cautious in interpreting the findings, because the way the program was implemented in the district—simultaneously across all schools—dictated the use of an interrupted times series without a comparison group. The results may therefore be confounded by changes in district-wide parameters not reflected in the data. A study with a greater geographical coverage would have the potential to produce more accurate estimates of the positive impact of *Treasures* on reading achievement in the upper elementary grades.

References

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.

Buchmann,C., DiPrete,T., & McDaniel, A. (2008). Gender inequalities in education. *Annual Review Sociology, 34*,319–37.

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

Cornwell, C., Mustard, D., & Van Parys, J. (March, 2011). *The gender gap in academic achievement among primary -school children: Test scores, teacher grades, and importance of non-cognitive skills.* Paper presented at the annual conference of the Association for Education and Finance, Seattle, WA.

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.

Entwisle, D.R., Alexander ,K.L., & Olson, L.S. (2007). Early schooling: the handicap of being poor and male. *Sociology of Education. 80*(2), 114–38.

Florida Department of Education. (2010). *Florida Comprehensive Assessment Test (FCAT): Sunshine State Standards, State Report of District Results* https://app1.fldoe.org/FCATDemographics/Selections.aspx?level=State&subj=Reading

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.

Machin, S., & McNally, S. (2006). *Gender and student achievement in English schools*. London: Center for the Economics of Education. Retrieved from http://eprints.lse.ac.uk/4666/1/Gender_and_Student_Achievement_in_English_Schools.pdf

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

McGraw-Hill Companies. (2010, October). *Effectiveness of McGraw-Hill's* Teasures *Reading Program in Grades 3–5.* (Empirical Education Rep. No. Empirical_MGH-6027-FR1-Y1-O.2). Palo Alto, CA: Empirical Education Inc.

McGraw-Hill, & Westat. (2007). *Changes in Test Performance of Students Using the Treasures Program: A Look at an Inner City School.* New York: Macmillan/McGraw-Hill Glencoe.

Northwest Evaluation Association [NWEA]. (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.

Parekh, C. (2011, March). *How Do Boys and Girls Do? New Evidence on the Gender Gap in New York City Public Schools*. Paper presented at the annual conference of the Association for Education and Finance, Seattle, WA.

Willingham, W. W., & Cole, S.E. (1997). *Gender and Fair Assessment*. Mahwah, NJ: Lawrence Erlbaum.