Educational Research
Institute of America
www.eriaonline.com • email: info@eriaonline.com
An affiliate of Beck Evaluation & Testing Associates, Inc.

A STUDY OF THE INSTRUCTIONAL EFFECTIVENESS OF Collections © 2017

Report Number 531

July 2017

Contents

ABSTRACT

To help school students read, analyze, compare, and communicate their understanding of various literary texts. Houghton Mifflin Harcourt has published, ***Houghton Mifflin Harcourt Collections © 2017*** for students in grades 6 to 12***. Houghton Mifflin Harcourt Collections*** supports the Common Core State Standards for English Language Arts, provides complex texts including fiction, nonfiction, and informational texts, and enhances online collaboration with interactive Common Core writing lessons.

In order to evaluate the program's effectiveness, Houghton Mifflin Harcourt contracted with the Educational Research Institute of America (ERIA) to conduct a full school year study to test the effectiveness of the program. The study was conducted with students in grades 7 to 10 during the 2016-2017 academic year.

Pretest and post-test assessments were developed to assess the program objectives and the Common Core State Standards. The assessments were focused on having students read, analyze, compare, and communicate their understanding of various literary texts.

The increases were statistically significant at all grades and the effect sizes were substantively important and classified as medium at all grades. The results also showed ***Houghton Mifflin Harcourt Collections © 2017*** was effective with both higher and lower pretest scoring students at all grades. The small sample size that resulted when the grade 7 students were divided into two groups prevented an analysis of low and high pretest scoring students. Group sizes were adequate at grades 8, 9, and 10 for statistical analyses. The results at those three grades showed that the low pretest students increased their average scores statistically significantly and the effect sizes were substantively important and were classified as large. The high pretest students at all three grades increased their scores statistically significantly and the effect sizes were substantively important and were classified as medium at all three grades.

Overview of the Study

This report describes a 2016-2017 academic year study with students in grades 7 to 10 to determine the impact of the ***Houghton Mifflin Harcourt Collections © 2017*** program for students in grades 6 to 12.

Houghton Mifflin Harcourt Collections © 2017 transforms English Language Arts instruction to focus on mastery of the Common Core state standards in language arts. Organized into topical or thematic cross-genre collections of literary and informative texts including media, the Student Edition delivers standards instruction either in print or digitally. The program has been designed to help students develop abilities to analyze complex texts, determine evidence, reason critically, and communicate thoughtfully.

Houghton Mifflin Harcourt contracted with the Educational Research Institute of America (ERIA) to conduct a full year study during the 2016-2017 academic year to determine the program's effectiveness. The ***Houghton Mifflin Harcourt Collections © 2017*** was the primary instructional program in the tryout classes.

The program is described by the publisher on the Houghton Mifflin Harcourt web site as follows:

> *Collections is an innovative, new English Language Arts program for students in grades 6-12. Built to meet the rigorous expectations of the Common Core State Standards (CCSS), Collections propels the traditional literature anthology into the future with a multifaceted digital approach to prepare students for college, career and beyond. At each grade level, Collections is organized into six thematic groups of multi-genre, complex texts that provide a foundation in all aspects of Common Core instruction. Complemented by flexible digital components that deepen students' knowledge, reinforce key skills and create personalized learning environments, the program includes an interactive writing and editing workspace, a companion website offering current and curated media resources on key Collections topics, and personalized user dashboards for progress monitoring and planning.*
>
> *Collections places instructional focus on analysis, drawing inferences and conclusions, and producing evidence-based writing. Complex anchor texts and performance tasks challenge students to analyze and synthesize fiction, literary nonfiction, informational texts and other media.*

Research Questions

The following research questions guided the design of the study and the data analyses:

1. Is ***Houghton Mifflin Harcourt Collections © 2017*** effective in increasing the skill and knowledge of grade 7 to 10 students to analyze complex texts, determine evidence, reason critically, and communicate thoughtfully?
2. Is ***Houghton Mifflin Harcourt Collections © 2017*** effective in increasing the skill and knowledge of grade 7 to 10 students who scored higher or lower on the pretests?

Design of the Study

The program's efficacy was evaluated using a pretest/post-test design. The study took place during the 2016/2017 academic year in three states in five different schools. There were 3 different teachers at grades 7, 9, and 10. At grade 8 there were 4 different teachers.

Pre-tests and post-tests were administered at the beginning and end of the school year. The tests modeled the assessments developed for the Collections program. The test carefully matched the standards that were the focus of the instructional program. The classroom teachers administered the pretests and post-tests. All tests were returned to ERIA for scoring and analyses.

Timeline and Program Use

The teachers used the ***Houghton Mifflin Harcourt Collections© 2017*** text as their primary instructional program. The teachers reported using the program an average of 3 days per week and for an average of about 35 minutes per day over the entire academic year. Pretests were administered mid-September, 2016 and posttests were administered mid-June, 2017.

Description of the Research Sample

Table 1 provides the demographic characteristics of the schools included in the study. It is important to note that the school data does not provide a description of the make-up of the classes that participated in the study. However, the data does provide a general description of the school and, thereby, an estimate of the make-up of the classes included in the study.

The percentage of students enrolled in free/reduced lunch programs ranged from 47% to 71% and averaged 60% across the sample of schools. By comparison, the reported national average for students enrolled in free/reduced lunch programs in public schools is reported as 48.1%.

The percentage of students classified as minority students (non-Caucasian) ranged

from 1% to 81% with an average of 47%. By comparison, 49.8% of the students enrolled in U.S. public schools were classified as non-Caucasian.[1]

Table 1
Schools Included in the Study: Demographic Characteristics

School	State	Location	Grades	Enrollment	% Non-Caucasian	% FRLP*
1	MT	Rural	9 to 12	137	12%	47%
2	MT	Rural	7 to 8	55	1%	56%
3	WA	Town	7 to 12	211	62%	60%
4	WI	Rural	6 to 8	87	77%	71%
5	WI	Rural	9 to 12	129	81%	68%
Averages				124	47%	60%

**Free and Reduced Lunch Program*

Description of the Assessments

The pretest and post-test used in the study were developed to assess the literary analysis of various texts. Based on these standards 30 item multiple-choice assessment pre/post tests were developed focusing on students' abilities to analyze complex texts, determine evidence, reason critically, and communicate thoughtfully as taught in the ***Collections*** program.

Table 2 provides the statistical results for the administration of the post-test for grades 7 to 10. The KR 20 reliability and the Standard Error of Measurement for the post-test indicates that the post-test score results were reliable for arriving at decisions regarding the achievement of the students to whom the tests were administered.

Table 2
Post-Test Test Statistics

Test	Reliability*	SEM**
Grade 7 Post-test	.71	2.15
Grade 8 Post-test	.73	1.97
Grade 9 Post-test	.70	2.13
Grade 10 Post-test	.69	2.37

**Reliability computed using the Kuder-Richardson 20 formula.*
*** SEM is the Standard Error of Measurement.*

[1] *The National Center for Educational Statistics* (NCES) reported that for the 2011–2012 school year, 48.1% of public school students were enrolled in free/reduced lunch programs. No free/reduced lunch data were available for the 2012–2013 school year. Also, the NCES reported that for the 2012–2013 school year, 49.8% of public school students were classified as minority (non-Caucasian) students.

Test Item Discrimination

In addition to determining the reliability and standard error of measurement of a test the quality of a test can be evaluated by computing the discrimination of each test item. Test item discrimination is an easy concept to understand.

The calculation of item discrimination can range from -1.0 to +1.0. If the discrimination of a test is above 0 it means that the students who scored higher on the test answered the item correctly more often than students who scored lower on the test. If the discrimination is below 0 it would have a negative discrimination meaning that the students who scored lower on the test answered the question correctly more often than students who scored higher on the test.

All tests will have a range of item discriminations. We can, however, examine a test to see how many good items there are on a test. The average discrimination of all the items on a test should be above +.15. The highest discriminations are rarely above +.50.

A scale that can be used to evaluate the discrimination of test items and the number of items for each of the four tests used in this study is provided in Table3. The table shows that for grades 7 to 10 the percentage of acceptable, good or excellent test items ranges from a low of 83% to a high of 96% with an average across the 4 grades of 88%.

Table 3

Test Item Discrimination for Collections Post-test Assessments

		Test Items in each Category			
Item Discrimination	Discrimination Values	Grade 7 Post-test	Grade 8 Post-test	Grade 9 Post-test	Grade 10 Post-test
Below 0	Poor test items (should be replaced)	2	1	1	1
+.01 to +.10	Weak test items (revise items)	3	0	3	4
+.11 to +.20	Acceptable	4	8	6	4
+.21 to +.30	Good items	7	12	8	2
+.30	Excellent test items	14	9	12	19

Data Analyses

Standard scores were developed to provide a more normal distribution of scores. The standard scores were a linear transformation of the raw scores. A mean raw score was translated to a mean standard score of 300 and the standard deviation of the raw scores was translated to 50. Standard scores were then used for the statistical analyses.

Data analyses and descriptive statistics were computed for the standard scores from the ***Collections*** assessments. The $\leq .05$ level of significance was used as the level at which increases would be considered statistically significant for all the statistical tests.

The following statistical analyses were conducted to compare students' pretest scores to post-test scores:

- A paired comparison *t*-test was used to compare the pretest mean standard scores with the post-test mean standard scores for all students.
- The students were split into two groups based on pretest scores. Paired comparison *t*-tests were used with the group that scored higher and the group that scored lower on the pretest to determine if the program was equally effective with students who had lower and higher pretest scores.

Descriptive statistics were also used to compare pretest and post-test standard test scores for the total group as well as the higher and lower pretest score groups.

An effect-size analysis was computed for each of the paired *t*-tests. Cohen's *d* statistic was used to determine the effect size. This statistic provides an indication of the strength of the effect of the treatment regardless of the statistical significance. Cohen's *d* statistic is interpreted as follows:

.2 = small effect
.5 = medium effect
.8 = large effect

Analysis Results

Grade 7 Analyses

Researchers at ERIA conducted a paired c**o**mparison *t*-test to determine if the difference from pretest standard scores to post-test standard scores was statistically significant. For this analysis, researchers could match the pretest and post-test scores for 55 students. Students who did not take both the pretest and the post-test were not included.

Table 4 shows that the average standard score on the pretest was 288, and the average standard score on the post-test was 312. The increase was statistically significant and the effect size was substantively important and is classified as medium.

Table 4
Paired Comparison *t*-test Results
Pretest/Post-test Comparison of Standards Scores

Test	*Number Students*	*Mean Standard Score*	*SD*	*t-test*	*Significance*	*Effect Size*
Pretest	55	288	53.0	4.289	≤.0001	.51
Post-test	55	312	44.2			

Higher and Lower Scoring Students

An analysis was planned to determine if students who scored lower on the pretest made gains as great as those students who scored higher on the pretest. The sample size for conducting this analysis would have resulted in groups at each level with fewer than 30 students in each group. This was too small for a reliable statistical analysis. However, samples were large enough in grades 8, 9, and 10 to complete the analysis of higher and lower pretest scoring groups.

Grade 8 Analyses

Researchers at ERIA conducted a paired c**o**mparison *t*-test to determine if the difference from pretest standard scores to post-test standard scores was statistically significant. For this analysis, researchers could match the pretest and post-test scores for 79 students. Students who did not take both the pretest and the post-test were not included.

Table 5 shows that the average standard score on the pretest was 282, and the average standard score on the post-test was 318. The increase was statistically significant and the effect size was substantively important and is classified as medium.

Table 5
Paired Comparison *t*-test Results
Pretest/Post-test Comparison of Standards Scores

Test	*Number Students*	*Mean Standard Score*	*SD*	*t-test*	*Significance*	*Effect Size*
Pretest	79	282	50.7	7.675	≤.0001	.77
Post-test	79	318	42.4			

Higher and Lower Scoring Students

An additional analysis was conducted to determine if students who scored lower on the pretest made gains as great as those students who scored higher on the pretest. For this analysis students were ranked in order based on their pretest standard scores. The group of 79 students was divided into two approximately equal sized groups of 39 and 40 students. The first group included those students who scored lower on the pretest and the second group included those who scored higher on the pretests.

Pretest-to-posttest comparisons are shown in Table 6 for the lower and higher pretest scoring students. Scores were analyzed using a paired comparison *t*-test to determine if both groups made significant gains.

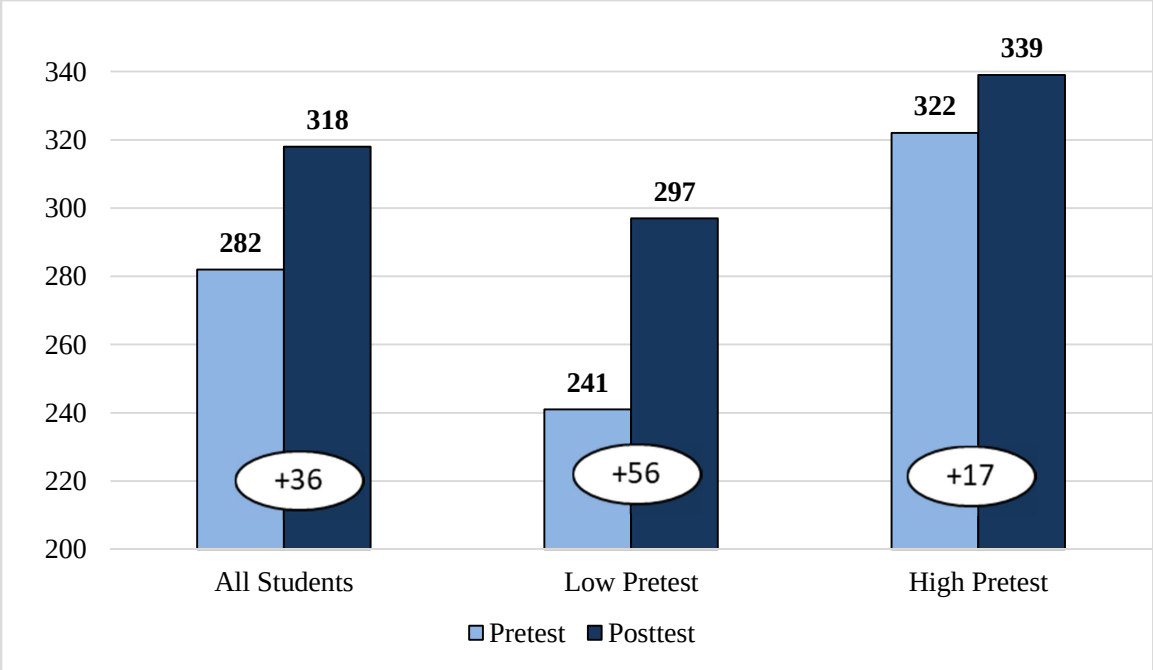
For both the higher and the lower scoring groups, the average scores increased statistically significantly. The effect sizes for both groups were substantively important and were classified as large for lower pretest scoring group and medium for the higher pretest scoring group.

Table 6
Paired Comparison *t*-test Results for Pretest/Posttest Standard Scores for the High- and Low-Scoring Pretest Groups

Test Form	*Number Students*	*Standard Score*	*SD*	*t-test*	*Significance*	*Effect Size*
Lower Scoring Group						
Pretest	39	241	32.6	8.458	≤.0001	1.61
Posttest	39	297	38.8			
Higher Scoring Group						
Pretest	40	322	28.3	3.258	≤.002	.51
Posttest	40	339	37.2			

Figure 1 provides a graphic representation of the gains achieved by the grade 8 students. The average scores for the total group increased 36 standard score points. The low pretest scoring students increased their average standard scores by 56 points and the high pretest scoring increased by 17 points.

Figure 1
Grade 8 Pretest Posttest Gain Comparison
All Students, Low Pretest Students, High Pretest Students



Grade 9 Analyses

Researchers at ERIA conducted a paired c**o**mparison *t*-test to determine if the difference from pretest standard scores to post-test standard scores was statistically significant. For this analysis, researchers could match the pretest and post-test scores for 83 students. Students who did not take both the pretest and the post-test were not included.

Table 7 shows that the average standard score on the pretest was 283, and the average standard score on the post-test was 318. The increase was statistically significant and the effect size was substantively important and is classified as medium.

Table 7
Paired Comparison *t*-test Results
Pretest/Post-test Comparison of Standards Scores

Test	*Number Students*	*Mean Standard Score*	*SD*	*t-test*	*Significance*	*Effect Size*
Pretest	83	283	46.7	6.415	≤.0001	.75
Post-test	83	318	47.2			

Higher and Lower Scoring Students

An additional analysis was conducted to determine if students who scored lower on the pretest made gains as great as those students who scored higher on the pretest. For this analysis students were ranked in order based on their pretest standard scores. The group of 83 students was divided into two approximately equal sized groups. The first group included 41 students who scored lower on the pretest. The higher scoring group included 42 students.

Pretest-to-posttest comparisons are shown in Table 8 for the lower and higher pretest scoring students. Scores were analyzed using a paired comparison *t*-test to determine if both groups made significant gains.

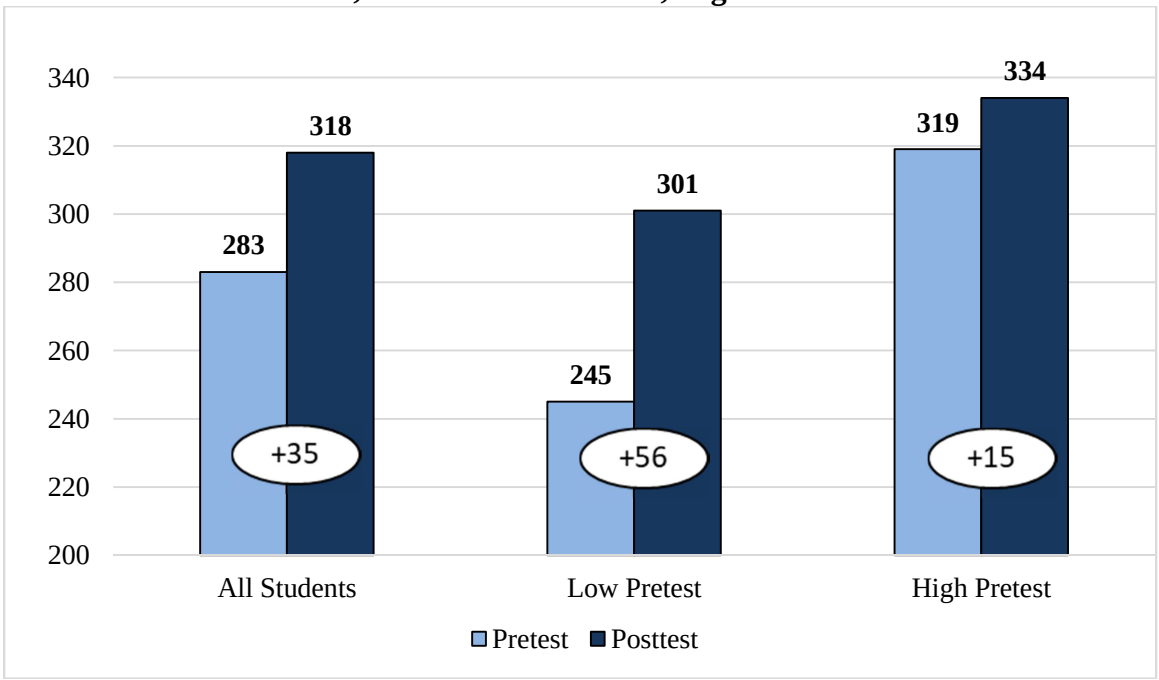
For both the higher and the lower scoring groups, the average scores increased statistically significantly. The effect sizes for both groups were substantively important and were classified as large for lower pretest scoring group and small for the higher pretest scoring group.

Table 8
Paired Comparison *t*-test Results for Pretest/Post-test Standard Scores for the High- and Low-Scoring Pretest Groups

Test Form	*Number Students*	*Standard Score*	*SD*	*t-test*	*Significance*	*Effect Size*
Lower Scoring Group						
Pretest	41	245	30.4	6.703	≤.0001	1.39
Post-test	41	301	48.3			
Higher Scoring Group						
Pretest	42	319	26.2	2.627	≤.0001	.44
Post-test	42	334	40.4			

Figure 2 provides a graphic representation of the gains achieved by the grade 9 students. The average scores for the total group increased 35 standard score points. The low pretest scoring students increased their average standard scores by 56 points and the high pretest scoring increased by 15 points.

Figure 2
Grade 9 Pretest Posttest Gain Comparison
All Students, Low Pretest Students, High Pretest Students



Grade 10 Analyses

Researchers at ERIA conducted a paired c**o**mparison *t*-test to determine if the difference from pretest standard scores to post-test standard scores was statistically significant. For this analysis, researchers could match the pretest and post-test scores for 79 students. Students who did not take both the pretest and the post-test were not included.

Table 9 shows that the average standard score on the pretest was 282, and the average standard score on the post-test was 318. The increase was statistically significant and the effect size was substantively important and is classified as medium.

Table 9
Paired Comparison *t*-test Results
Pretest/Post-test Comparison of Standards Scores

Test	*Number Students*	*Mean Standard Score*	*SD*	*t-test*	*Significance*	*Effect Size*
Pretest	79	282	50.7	7.675	≤.0001	.75
Post-test	79	318	42.4			

Higher and Lower Scoring Students

An additional analysis was conducted to determine if students who scored lower on the pretest made gains as great as those students who scored higher on the pretest. For this analysis students were ranked in order based on their pretest standard scores. The group of 79 students was divided into two approximately equal sized groups. The first group included 39 students who scored lower on the pretest. The higher scoring group included 40 students.

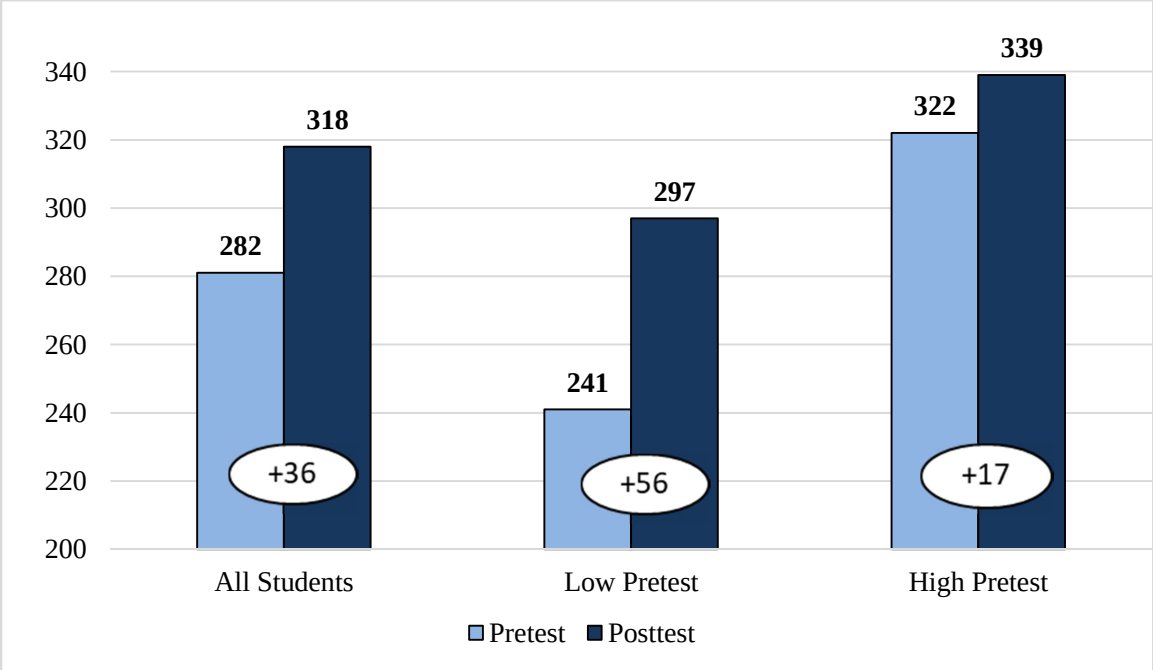
Pretest-to-posttest comparisons are shown in Table 10 for the lower and higher pretest scoring students. Scores were analyzed using a paired comparison *t*-test to determine if both groups made significant gains. For both the higher and the lower scoring groups, the average scores increased statistically significantly. The effect size for the lower pretest scoring group was large and for the higher pretest scoring students the effect size was medium.

Table 10
Paired Comparison *t*-test Results for Pretest/Post-test Standard Scores for the High- and Low-Scoring Pretest Groups

Test Form	*Number Students*	*Standard Score*	*SD*	*t-test*	*Significance*	*Effect Size*
Lower Scoring Group						
Pretest	39	241	32.6	8.458	≤.0001	1.61
Post-test	39	297	36.8			
Higher Scoring Group						
Pretest	40	322	28.3	3.258	≤.002	.51
Post-test	40	339	37.2			

Figure 3 provides a graphic representation of the gains achieved by the grade 10 students. The average scores for the total group increased 36 standard score points. The low pretest scoring students increased their average standard scores by 56 points and the high pretest scoring increased by 17 points.

Figure 3
Grade 10 Pretest Posttest Gain Comparison
All Students, Low Pretest Students, High Pretest Students



Conclusions

This study sought to determine the effectiveness of ***Houghton Mifflin Harcourt Collections © 2017,*** a grade 6 to 12 literature program published by Houghton Mifflin Harcourt. The study was carried out with classes at grades 7 to 10. The teachers were using the program for the first time and received no special instruction in using the program.

Two research questions guided the study:

Question 1: *Is Houghton Mifflin Harcourt Collections effective in increasing the skill and knowledge of grades 7 to 10 students to analyze complex texts, determine evidence, reason critically, and communicate thoughtfully?*

Pretests and post-tests were developed to match the standards of the Collections program. The assessments covered the objectives of the program as well as the Common Core State Standards. For students at all four grades the test scores increased statistically significantly. The effect sizes were substantively important at all four grades and were classified as medium.

Question 2: *Is Houghton Mifflin Harcourt Collections effective in increasing the skill and knowledge of grades 7 to 10 at higher and lower pretest scoring levels?*

At grade 7 the sample sizes were too small to conduct valid analyses. For grades 8, 9, and 10 both high and low pretest scoring groups increased their scores from pretesting to post-testing statistically significantly. The effect sizes for the lower pretest scoring groups was substantively important and were classified as large at all three grades. The increase from pretesting to post-testing for the higher pretest scoring group were statistically significant for all three groups. The effect size was substantively important and classified as medium for all three grades.

Based on this study, both research questions can be answered positively.

- ***The Houghton Mifflin Harcourt Collections program is effective in improving the ability of students in grades 7 to 10 to analyze complex texts, determine evidence, reason critically, and communicate thoughtfully.***

- ***The Houghton Mifflin Harcourt Collections program is effective in improving the ability of lower performing as well as higher performing students in grades 8 to 10.[i] Students at all three grade levels showed significant improvement in their ability to analyze complex texts, determine evidence, reason critically, and communicate thoughtfully.***

[i] Analyses were not conducted at grade 7 due to the small sample size when the group was split.