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 507

July 2016

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 study to test the effectiveness of specific program units from the program. Teachers who were already using the Collections program were contacted to determine if they would like to participate in the study. Each teacher was asked to select a unit that they would be teaching sometime during the beginning of the second semester. A total of six different units were included in the study. Each unit took between 6 to 8 weeks for teachers to complete. A unit pretest was administered before teachers began using the unit and a post-test was administered after the teacher completed the unit. Each teacher maintained a somewhat different schedule based on their schedule for the particular unit being taught. The instruction took place during the second half of the 2015-2016 academic year.

Pretest and post-test assessments were developed to assess the program objectives for the particular unit that had been selected. The assessments were focused on having students read, analyze, compare, and communicate their understanding of various literary texts.

Test results were translated to standard scores. For each of the three grades the analyses showed that the ***Houghton Mifflin Harcourt Collections*** classes made statistically significant gains at each grade. The effect sizes were large for grades 6 and 7 and medium at grade 8.

Results for the three grades were combined and the results were statistically significant for the total group and for the pretest high and low scoring groups. The effect size for the whole group and for the low pretest scoring group were large. For the high pretest scoring group the effect size was medium.

Overview of the Study

This report describes a second semester study conducted during 2015-2016 academic year with students in grade 6, 7, and 8 to determine the impact of individual units of the ***Houghton Mifflin Harcourt Collections © 2017*** program for students in grades 6 to 12. The program has been designed to help students develop their abilities to analyze complex texts, determine evidence, reason critically, and communicate thoughtfully. 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.

Houghton Mifflin Harcourt School Publishers contracted with the *Educational Research Institute of America* (ERIA) to conduct a study of specific units of study during the second semester of the 2015/2016 academic year to determine the program's effectiveness. The ***Houghton Mifflin Harcourt Collections*** was the primary instructional program in the tryout classes. Participating teachers were asked to select a unit from the program that would be scheduled for use at the beginning or middle of the second semester. The decision as to which unit to use was each teacher's independent decision.

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

> *Collections© 2017 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*** effective in increasing the skills and knowledge of sixth, seventh, and eighth grade students to analyze complex texts, determine evidence, reason critically, and communicate thoughtfully?
2. Is ***Houghton Mifflin Harcourt Collections*** equally effective in increasing the skills and knowledge of sixth, seventh, and eighth grade students scoring at higher and lower pretest score levels?

Design of the Study

The program's efficacy was evaluated using a quasi-experimental pretest/posttest design. The study took place during the second semester of the 2015/2016 academic year in two different states in three different schools. The program included 6 different teachers. Each of the six teachers selected a different unit to teach as part of the study. The grade levels and units include:

Grade 6	1 teacher	Unit 4
Grade 7	3 teachers	Units 1, 4, 6
Grade 8	2 teachers	Units 2, 4

Pre-tests and post-tests were administered at the beginning and end of the tryout of each unit. The tests carefully matched the standards that were the focus of the instructional unit being taught. Pretest and post-test administration was under the direction of the classroom teacher. All tests were returned to ERIA for scoring and analyses.

Timeline and Program Use

The teachers used the ***Houghton Mifflin Harcourt Collections*** text as their primary instructional program. They were asked to select a unit from the program to be included in the study. The teachers reported teaching the unit from 6 to 8 weeks for an average of 4 to 5 days per week, and for 35 to more than 50 minutes per day.

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.

Table 1
Schools Included in the Study: Demographic Characteristics

School	State	Location	Grades	Enrollment	% Minority	% Free/Reduced Lunch
1	WA	Suburban	7 to 8	713	16%	28%
2	WA	Suburban	7 to 8	744	18%	23%
3	IL	Suburban	7 to 8	1037	98%	94%
Averages				831	44%	48%

Description of the Assessments

The pretest and posttest 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 program.

Table 2 provides the statistical results for the administration of the post-test for the 6 different post-tests used. The KR 20 reliability and the Standard Error of Measurement for the post-test indicates the posttest 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 6 Unit 4	.82	2.19
Grade 7 Unit 1	.68	2.21
Grade 7 Unit 4	.66	2.40
Grade 7 Unit 6	.77	2.12
Grade 8 Unit 2	.84	2.18
Grade 8 Unit 4	.75	2.56

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

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 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. It would be best, however, if a test had no negative discriminating items and all positive discriminating items were above +.10.[1] However, that is very seldom the case with any test. We can, however, examine a test to see how many "psychometrically" 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 seldom above +.50.

A scale that can be used to evaluate the discrimination of test items and the number of items for each of the two tests used in this study is provided in Table3. The table shows that the percentage of acceptable, good, or excellent items ranges from 83% to 97%. The average test item discriminations for all of the tests are excellent.

Table 3

		Grade 6	**Grade 7**			**Grade 8**	
Item Discrimination	**Discrimination Values**	**Unit 4**	**Unit 1**	**Unit 4**	**Unit 6**	**Unit 2**	**Unit 4**
Below 0	Poor items	2	1	4	1	3	1
+.01 to +.10	Weak items	1	4	3	0	1	2
+.11 to +.20	Acceptable	4	3	2	2	2	3
+.21 to +.30	Good	2	3	9	6	4	4
+.30	Excellent	21	19	12	21	20	20
% of Items Acceptable, Good or Excellent		***90%***	***83%***	***77%***	***97%***	***87%***	***90%***
Average Test Item Discriminations		***+.40***	***+.31***	***+.27***	***+.35***	***+.31***	***+.36***

[1] Item discrimination is determined by the quality of the test item but also by the effects of instruction and the performance level of students to whom the test is being administered.

Data Analyses

Standard scores were developed in order 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 assessments. The ≤.05 level of significance was used as the level at which increases would be considered statistically significant for all of the statistical tests.

The total group of grade 6, 7, and 8 were analyzed together and then each grade level was analyzed separately. The following statistical analyses were conducted to compare students' pretest scores to posttest scores:

- A paired comparison *t*-test was used to compare the pretest mean standard scores with the posttest 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 pretests 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

All Grades and Units

A paired c**o**mparison *t*-test was used to determine if the difference from pretest standard scores to posttest standard scores was statistically significant across all six units at each of three grades. For this analysis, researchers were able to match the pretest and posttest scores for 306 students. Each of the six units was taught by a different teacher. Students who did not take both the pretest and the posttest were not included in the analyses.

Table 4 shows that the average standard score on the pretest was 299, and the average standard score on the posttest was 310. The increase was statistically significant (≤.0001), and the effect size was large.

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

Test	*Number Students*	*Mean Standard Score*	*SD*	*t-test*	*Significance*	*Effect Size*
Pretest	306	299	12.33	14.623	≤.0001	.84
Posttest	306	310	14.31			

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 on the basis of their pretest standard scores. The group of 306 students was divided into two equal sized groups of 153 students.

Pretest-to-posttest comparisons are shown in Table 5 and were analyzed using a paired comparison *t*-test to determine if both groups made significant gains.

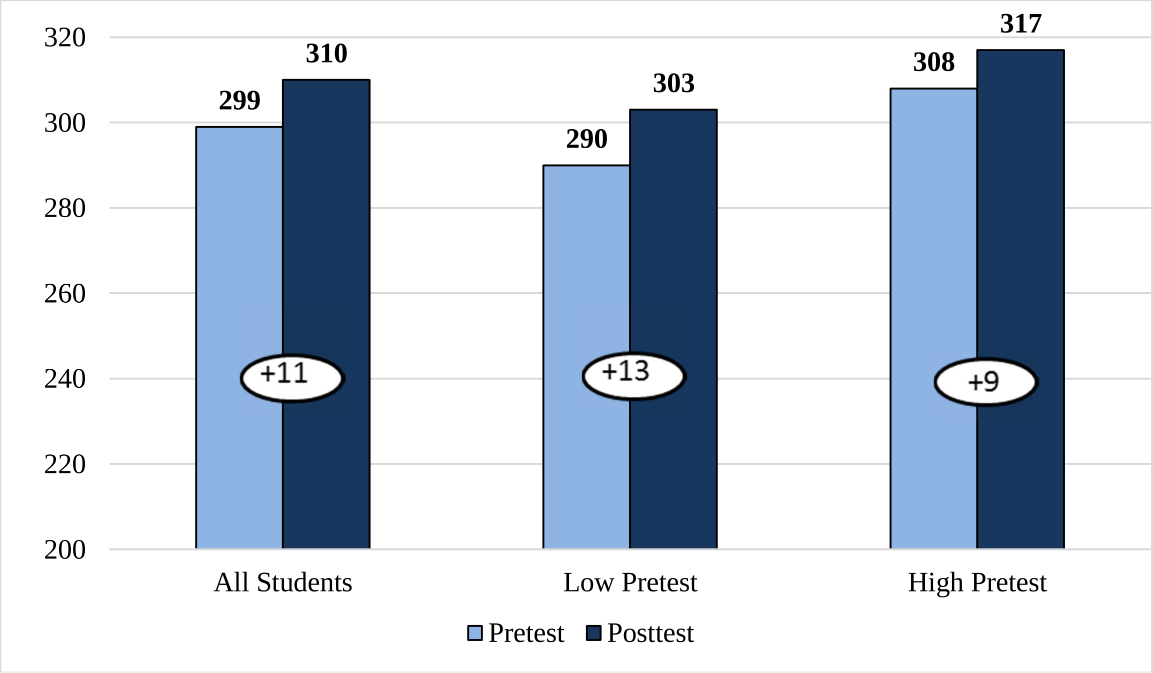
For both the higher and the lower scoring groups, the average scores increased statistically significantly (≤.0001). The effect size for the lower pretest scoring group was large and for the higher pretest scoring the effect size was medium.

Table 5
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	153	290	5.91	12.522	≤.0001	1.40
Posttest	153	303	12.00			
Higher Scoring Group						
Pretest	153	308	10.82	8.426	≤.0001	.75
Posttest	153	317	13.31			

Figure 1 shows that the average scores for the total group increased 11 standard score points, the low pretest scoring students increased by 13 points, and the high pretest scoring increased by 9 points.

Figure 1
Grade 9 Pretest Posttest Gain Comparison
All Students, Low Pretest Students, High Pretest Students



Grade 6 Unit 4

Researchers at ERIA conducted a paired comparison *t*-test to determine if the difference from pretest standard scores to posttest standard scores was statistically significant. For this analysis, researchers were able to match the pretest and posttest scores for 24 students. Students who did not take both the pretest and the posttest were not included. The sample size was small as only unit 4 unit was tried out at grade 6.

Table 6 shows that the average standard score on the pretest was 299, and the average standard score on the posttest was 311. The increase was statistically significant (≤.0001), and the effect size was large.

Table 6
Paired Comparison *t*-test Results
Pretest/Posttest Comparison of Standards Scores

Test	*Number Students*	*Mean Standard Score*	*SD*	*t-test*	*Significance*	*Effect Size*
Pretest	24	299	9.9	5.124	≤.0001	.81
Posttest	24	311	17.5			

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 on the basis of their pretest standard scores. The group of 24 students was divided into two equal sized groups of 12 students.

Pretest-to-posttest comparisons are shown in Table 7 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.

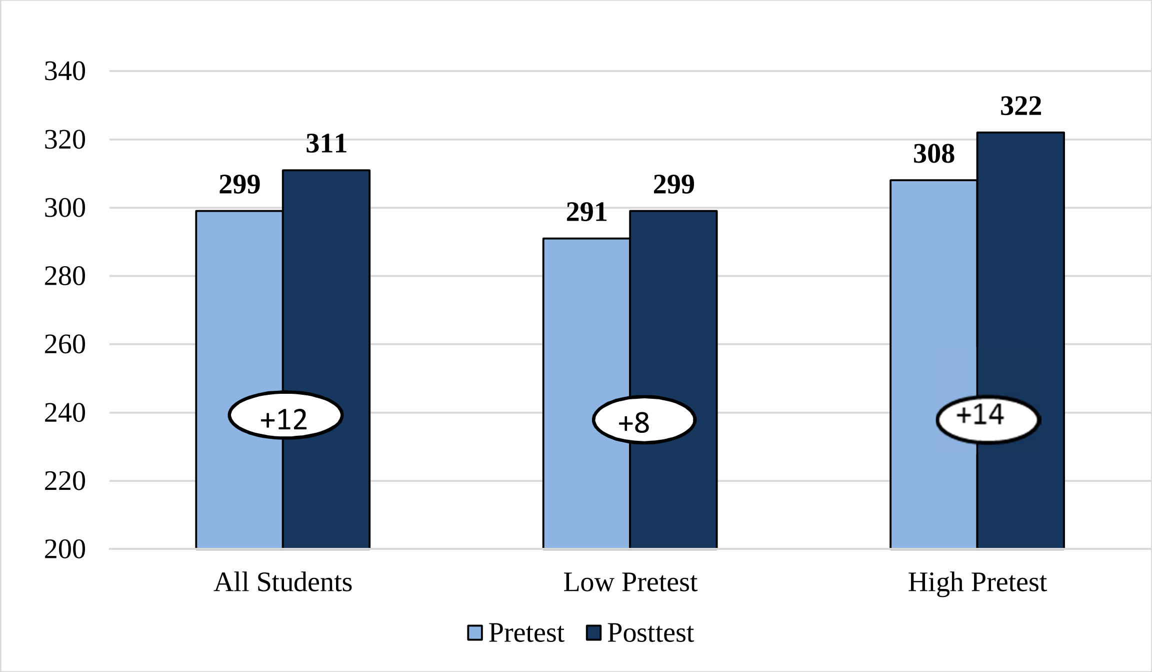
For both the higher and the lower scoring groups, the average scores increased statistically significantly (≤.0001). The effect size for the lower and higher pretest scoring group was large.

Table 7
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	12	291	3.34	3.368	≤.006	1.17
Posttest	12	299	9.45			
Higher Scoring Group						
Pretest	12	308	6.51	4.060	≤.002	1.20
Posttest	12	322	16.04			

Figure 2 provides a graphic representation of the gains achieved by the grade 6 students. The average scores for the total group increased 12 standard score points. The low pretest scoring students increased their average standard scores by eight points and the high pretest scoring students average standard scores increased by 14 points.

Figure 2
Grade 6 Unit 4 Pretest/Posttest Gain Comparison
All Students, Low Pretest Students, High Pretest Students



Grade 7 Units 1, 4, and 6

Researchers at ERIA conducted a paired comparison *t*-test to determine if the difference from pretest standard scores to posttest standard scores was statistically significant. For this analysis, researchers were able to match the pretest and posttest scores for 156 students. Units 1, 4, and 6 were each tried out with a different teacher for each unit. Students who did not take both the pretest and the posttest were not included.

Table 8 shows that the average standard score on the pretest was 298, and the average standard score on the posttest was 311. The increase was statistically significant (≤.0001), and the effect size was large.

Table 8
Paired Comparison *t*-test Results
Pretest/Posttest Comparison of Standards Scores

Test	*Number Students*	*Mean Standard Score*	*SD*	*t-test*	*Significance*	*Effect Size*
Pretest	156	298	12.93	10.967	≤.0001	.92
Posttest	156	311	15.24			

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 on the basis of their pretest standard scores. The group of 156 students was divided into two equal sized groups of 78 students.

Pretest-to-posttest comparisons are shown in Table 9 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.

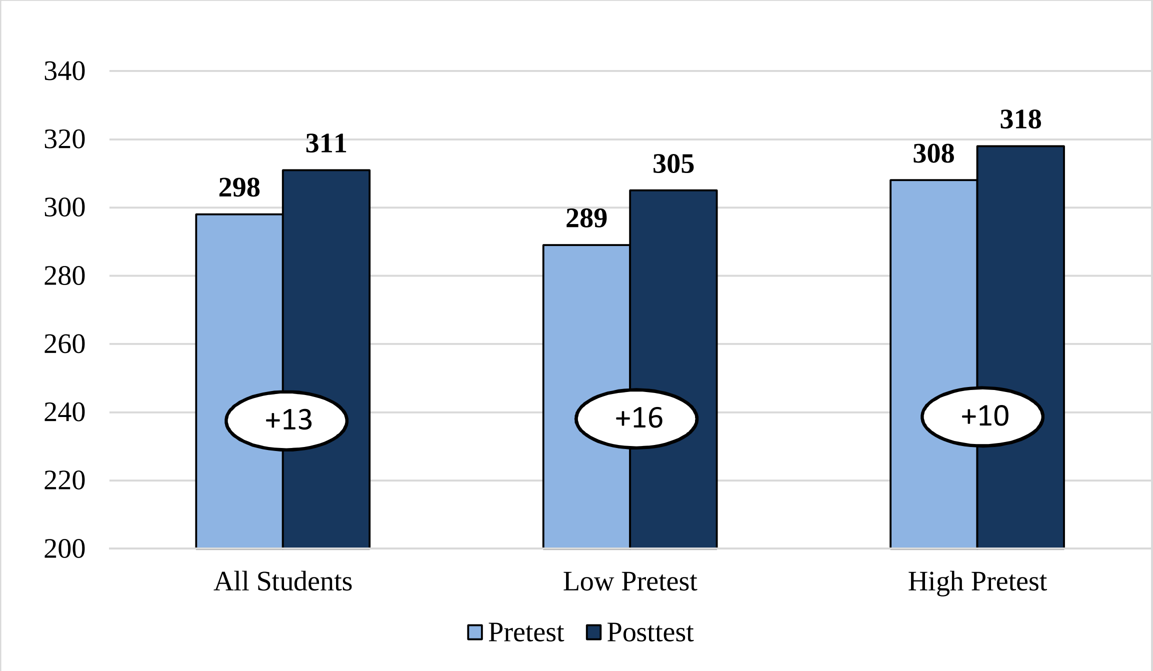
For both the higher and the lower scoring groups, the average scores increased statistically significantly (≤.0001). The effect size for the lower pretest scoring group was large and for the higher pretest scoring the effect size was medium.

Table 9
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	78	289	4.03	10.443	≤.0001	1.58
Posttest	78	305	14.19			
Higher Scoring Group						
Pretest	78	308	11.39	5.647	≤.0001	.77
Posttest	78	318	13.64			

Figure 3 provides a graphic representation of the gains achieved by the grade 7 students. The average scores for the total group increased 13 standard score points. The low pretest scoring students increased their average standard scores by 16 points and the high pretest scoring students increased by 10 points.

Figure 3
Grade 7 Pretest/Posttest Gain Comparison
All Students, Low Pretest Students, High Pretest Students



Grade 8 Units 2 and 4

Researchers at ERIA conducted a paired comparison *t*-test to determine if the difference from pretest standard scores to posttest standard scores was statistically significant. For this analysis, researchers were able to match the pretest and posttest scores for 126 students. Units 2, and 4 were each tried out with a different teacher for each unit. Students who did not take both the pretest and the posttest were not included.

Table 10 shows that the average standard score on the pretest was 300, and the average standard score on the posttest was 308. The increase was statistically significant (≤.0001), and the effect size was medium.

Table 10
Paired Comparison *t*-test Results
Pretest/Posttest Comparison of Standards Scores

Test	*Number Students*	*Mean Standard Score*	*SD*	*t-test*	*Significance*	*Effect Size*
Pretest	126	300	12.03	8.671	≤.0001	.72
Posttest	126	308	12.24			

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 on the basis of their pretest standard scores. The group of 126 students was divided into two equal sized groups of 63 students.

Pretest-to-posttest comparisons are shown in Table 11 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.

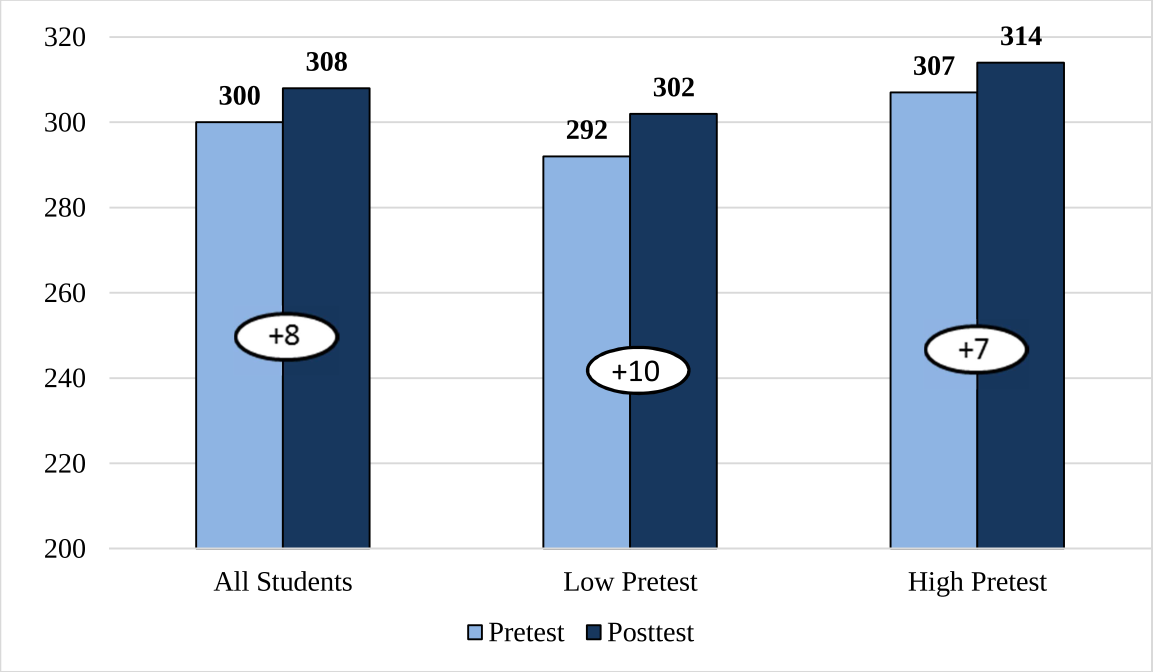
For both the higher and the lower scoring groups, the average scores increased statistically significantly (≤.0001). The effect size for the lower pretest scoring group was large and for the higher pretest scoring the effect size was medium.

Table 11
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	63	292	7.58	6.885	≤.0001	1.22
Posttest	63	302	8.86			
Higher Scoring Group						
Pretest	63	307	10.85	5.375	≤.0001	.65
Posttest	63	314	12.04			

Figure 4 provides a graphic representation of the gains achieved by the grade 8 students. The average scores for the total group increased eight standard score points. The low pretest scoring students increased their average standard scores by 10 points and the high pretest scoring increased by seven points.

Figure 4
Grade 8 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 6, 7, and 8. The teachers were using the program for the first time. They were asked if they would like to be included in a study of one unit from the program. The teachers were asked to select a unit they would be teaching during the second semester. A total of seven teachers agreed to participate. The teachers independently selected the unit to be studies. Each teacher selected a different unit.

Two research questions guided the study:

Question 1: *Is* **Houghton Mifflin Harcourt Collections** *effective in increasing the skill and knowledge of grade 6, 7 and 8 students to analyze complex texts, determine evidence, reason critically, and communicate thoughtfully?*

Pretests and post-tests were developed to match the standards of the each of the six units included in the study. Despite the short duration of the study student scores combined across all three grades increased statistically significantly and the effect size was large. The analysis of each grade separately also resulted in statistically significant increases.

Question 2: *Is* **Houghton Mifflin Harcourt Collections** *effective in increasing the skill and knowledge of grade 6, 7, and 8 at higher pretest scoring and lower pretest scoring students?*

For the combined group of students at grades 6, 7, and 8 and for each grade level analyzed separately, the increases were statistically significant. The effect sizes for the low pretest scoring group was large for the combined group as well as for students analyzed separately at each of the three groups. For the higher scoring total group and for the students analyzed separately at each grade the effect size was medium.

On the basis of this study, both research questions can be answered positively. In fact, the study results are remarkably positive and consistent across three grades for a study of limited duration and relatively small sample sizes.

- ***The Houghton Mifflin Harcourt Collections program is effective in improving the ability of grade 6, 7, and 8 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 at grades 6, 7, and 8 to analyze complex texts, determine evidence, reason critically, and communicate thoughtfully.***