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 © 2015

Report Number 504

July 2015

## 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 © 2015*** 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 and 9 during the 2014-2015 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 results showed that the ***Houghton Mifflin Harcourt Collections*** classes made statistically significant gains at both grades 7 and 9 over the course of the full year study. The increases at both grades were statistically significant and the effect sizes were medium at grade 7 and small at grade 9. The results also showed the ***Houghton Mifflin Harcourt Collections*** program proved effective with both higher and lower pretest scoring grade 7 and grade 9 students. Those students increased their average scores statistically significantly and the effect sizes at grade 7 were large for the lower pretest scoring students and medium for the higher pretest scoring students. For grade 9 students the effect size for the lower pretest scoring students was medium and for the higher pretest scoring students the effect size was small.

## Overview of the Study

This report describes a 2014-2015 academic year study with students in grade 7 and 9 to determine the impact of the ***Houghton Mifflin Harcourt Collections © 2015*** program for students in grades 6 to 12. The English Language Arts instruction in ***Houghton Mifflin Harcourt Collections © 2015*** focuses 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.

*Houghton Mifflin Harcourt School Publishers* contracted with the *Educational Research Institute of America* (ERIA) to conduct a full year study during the 2014/2015 academic year to determine the program's effectiveness. The ***Houghton Mifflin Harcourt Collections © 2015*** was the primary instructional program in all classes.

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

> *Collections© 2015 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 skill and knowledge of grade 7 and grade 9 students to analyze complex texts, determine evidence, reason critically, and communicate thoughtfully?
2. Is ***Houghton Mifflin Harcourt Collections*** equally effective in increasing the skill and knowledge of grade 7 and 9 student scoring higher and lower at pretest to analyze complex texts, determine evidence, reason critically, and communicate thoughtfully?

## Design of the Study

The program's efficacy was evaluated using a pretest/posttest design. At grade 7, the program was used by 14 teachers in 8 different schools located in 4 different states. At grade 9 there were 8 teachers in 5 different schools located in 4 different states.

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. Most questions were changed from the original questions included with those tests. The test carefully matched the standards that were the focus of the instructional program. 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. 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 the end of August/beginning of September, 2014 and posttests were administered the end of May/beginning of June, 2015.

## 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 |
|---|---|---|---|---|---|---|
| **Grade 7 Schools** | | | | | | |
| 1 | NJ | Suburban | 6 to 8 | 463 | 4% | 30% |
| 2 | MT | Rural | 6 to 8 | 656 | 13% | 33% |
| 3 | FL | Suburban | 7 to 8 | 509 | 6% | 26% |
| 4 | FL | Suburban | 7 to 12 | 1340 | 13% | 22% |
| 5 | FL | Urban | 7 to 8 | 709 | 39% | 65% |
| 6 | IN | Suburban | 7 to 8 | 766 | 3% | 45% |
| 7 | IN | Urban | 6 to 8 | 860 | 28% | 51% |
| 8 | IN | Urban | 6 to 8 | 758 | 45% | 71% |
| Averages | | | | 758 | 19% | 43% |
| **Grade 9 Schools** | | | | | | |
| 1 | MT | Rural | 9 to 12 | 1506 | 11% | 18% |
| 2 | FL | Urban | 9 to 12 | 1724 | 43% | 51% |
| 3 | FL | Suburban | 9 to 12 | 1449 | 13% | 24% |
| 4 | IN | Rural | 9 to 12 | 2400 | 15% | 33% |
| 5 | NJ | Suburban | 9 to 12 | 638 | 4% | 23% |
| Averages | | | | 1543 | 17% | 30% |

## 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, a 45 item multiple-choice assessments, at each grade level, 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 pretest and the post-test for both grades 7 and 9. The KR 20 reliability and the Standard Error of Measurement for the post-test indicates both the pretest score results and the posttest score results were reliable for arriving at decisions regarding the achievement of the students to whom the tests were administered.

**Table 2**
**Pretest and Post-Test Test Statistics**

| Test | Reliability* | SEM** |
|---|---|---|
| Grade 7 Pretest | .79 | 2.91 |
| Grade 7 Post-test | .82 | 2.71 |
| Grade 9 Pretest | .83 | 2.99 |
| Grade 9 Post-test | .84 | 2.79 |

**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 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. 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 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 two tests used in this study is provided in Table3. The table shows that both the grade 7 and grade 9 posttests have a large percentage of acceptable, good or excellent test items grade 7 (87%) grade 9 (89%). The average test item discriminations for grade 7 and grade 9 are excellent.

[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.

**Table 3**
**Test Item Discrimination for Collections Post-test Assessments**

| | | Test Items in each Category | |
|---|---|---|---|
| **Item Discrimination** | **Discrimination Values** | **Grade 7 Posttest** | **Grade 9 Posttest** |
| *Below 0* | Poor test items (should be replaced) | 1 | 0 |
| *+.01 to +.10* | Weak test items (revise items) | 5 | 5 |
| *+.11 to +.20* | Acceptable | 4 | 2 |
| *+.21 to +.30* | Good items | 6 | 5 |
| *+.30* | Excellent test items | 29 | 33 |
| ***Average*** | | ***+.32*** | ***+.36*** |

## 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 ***Collections*** assessments. The $\leq .05$ level of significance was used as the level at which increases would be considered statistically significant for all of the statistical tests.

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 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 posttest standard scores was statistically significant. For this analysis, researchers were able to match the pretest and posttest scores for 904 students. Students who did not take both the pretest and the posttest were not included.

Table 4 shows that the average standard score on the pretest was 290, and the average standard score on the posttest was 314. The increase was statistically significant (≤.0001). The effect size was medium.

**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 | 904 | 290 | 47.2 | 19.125 | ≤.0001 | .51 |
| Posttest | 904 | 314 | 47.5 | | | |

### 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 904 students was divided into two equal sized groups of 452 students. The first group included those students who scored lower on the pretest with a mean of 252 with scores ranging from 154 to 288. The higher scoring group scored an average standard score on the pretest of 328 with scores ranging from 288 to 422.

Pretest-to-posttest comparisons are shown in Table 5 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.

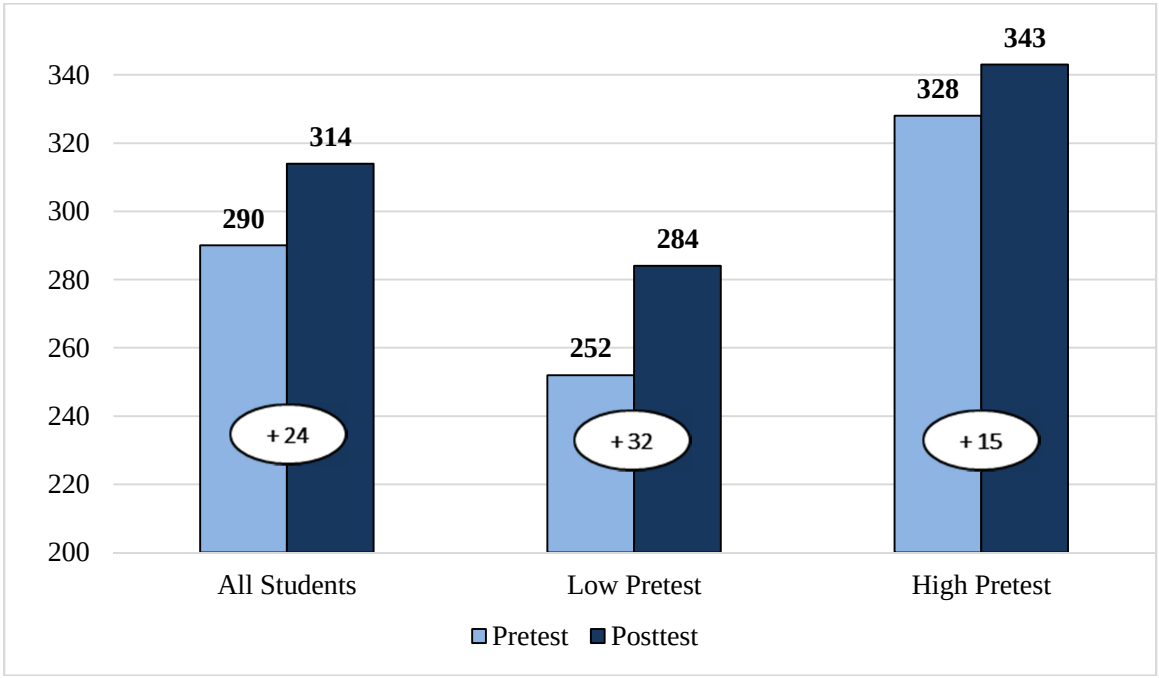
For both the higher and the lower scoring groups, the average scores increased statistically significantly (≤.0001). The effect size for the lower pretest scoring group was large and for the higher pretest scoring the effect size was medium.

**Table 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 | 452 | 252 | 28.5 | 17.069 | ≤.0001 | .92 |
| Posttest | 452 | 284 | 40.3 | | | |
| Higher Scoring Group | | | | | | |
| Pretest | 452 | 328 | 27.3 | 10.088 | ≤.0001 | .50 |
| Posttest | 452 | 343 | 33.9 | | | |

Figure 1 provides a graphic representation of the gains achieved by the grade 7 students. The average scores for the total group increased 24 standard score points. The low pretest scoring students increased their average standard scores by 32 points which was an increase 100% higher than the high pretest scoring students whose average standard scores increased by 15 points.

**Figure 1**
**Grade 7 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 posttest standard scores was statistically significant. For this analysis, researchers were able to match the pretest and posttest scores for 366 students. Students who did not take both the pretest and the posttest were not included.

Table 6 shows that the average standard score on the pretest was 292, and the average standard score on the posttest was 308. The increase was statistically significant (≤.0001). The effect size was small.

**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 | 366 | 292 | 50.3 | 9.799 | ≤.0001 | .32 |
| Posttest | 366 | 308 | 48.3 | | | |

## 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 366 students was divided into two equal sized groups of 183 students. The first group included those students who scored lower on the pretest with a mean of 251, ranging from 150 to 295. The higher scoring group scored an average standard score on the pretest of 332, ranging from 302 to 413.

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.

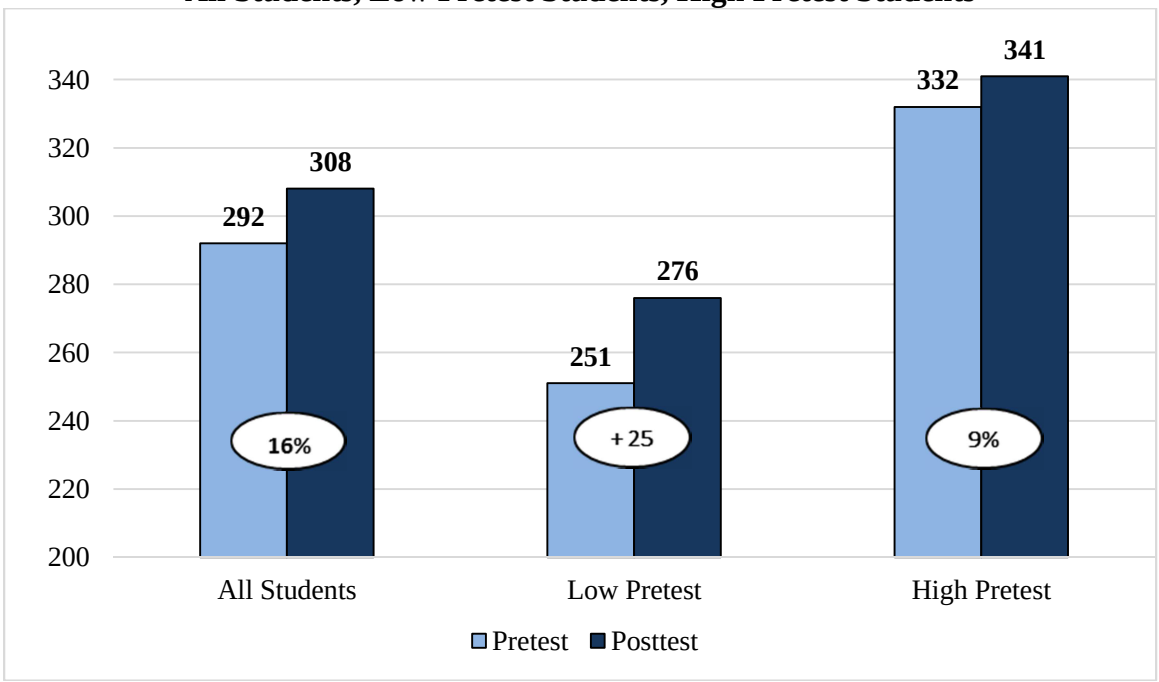
For both the higher and the lower scoring groups, the average scores increased statistically significantly (≤.0001). The effect size for the lower pretest scoring group was medium and for the higher pretest scoring the effect size was small.

**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 | 183 | 251 | 35.0 | 8.857 | ≤.0001 | .64 |
| Posttest | 183 | 276 | 42.2 | | | |
| Higher Scoring Group | | | | | | |
| Pretest | 183 | 332 | 24.6 | 4.880 | ≤.0001 | .34 |
| Posttest | 183 | 341 | 28.8 | | | |

Figure 2 provides a graphic representation of the gains achieved by the grade 9 students. The average scores for the total group increased 16 standard score points. The low pretest scoring students increased their average standard scores by 25 points and the high pretest scoring increased by 9 points.

**Figure 2**
**Grade 9 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 © 2015,*** a grade 6 to 12 literature program published by Houghton Mifflin Harcourt. The study was carried out with classes at grades 7 and 9. The teachers were using the program for the first time and received no special instruction.

Two research questions guided the study:

***Question 1:*** *Is Houghton Mifflin Harcourt Collections effective in increasing the skill and knowledge of grade 7 and grade 9 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 the grade 7 students, statistical analyses of students' scores showed that the students increased their scores statistically significantly and the effect size was medium. For the grade 9 students, statistical analyses of students' scores showed that the students increased their scores statistically significantly and the effect size was small.

***Question 2:*** Is ***Houghton Mifflin Harcourt Collections*** equally effective in increasing the skill and knowledge of grade 7 and 9 student scoring higher and lower at pretest to analyze complex texts, determine evidence, reason critically, and communicate thoughtfully?

At grade 7 the analysis of the low scoring and high scoring pretest students showed that both groups increased statistically significantly. The effect size for the lower pretest scoring group was large and for the higher scoring pretest group, the effect size was medium. For grade 9 students the analysis of the low scoring and high scoring pretest students showed that both groups increased their scores statistically significantly and the effect size for the lower pretest scoring group was medium. The grade 9 high pretest scoring group increased statistically significantly from pretesting to post-testing and the effect size was small.

On the basis of this study, both research questions can be answered positively.

- ***The Houghton Mifflin Harcourt Collections program is effective in improving the ability of grade 7and 9 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 grade 7 and 9 students to analyze complex texts, determine evidence, reason critically, and communicate thoughtfully.***