Measuring the Average Impact of an iPad Algebra Program

A REPORT OF FINDINGS FROM AN RCT IN FOUR SCHOOL DISTRICTS CONSIDERING ONE AS A SPECIAL CASE

March 2012

Andrew Jaciw
Megan Toby
Boya Ma
Garrett Lai
Li Lin
Empirical Education Inc.

Acknowledgements

We are grateful to the people in Fresno Unified School District, Long Beach Unified School District, Riverside Unified School District, and San Francisco Unified School District, for their assistance and cooperation in conducting this research. The research was sponsored by Houghton Mifflin Harcourt which provided Empirical Education Inc. with independence in reporting the results.

About Empirical Education Inc.

Empirical Education Inc. is a Palo Alto, California-based research company that provides rigorous and independent evidence to inform school system decisions. The company brings its expertise in research, data analysis, engineering, and project management to customers that include the US Department of Education, educational publishers, foundations, leading research organizations, and state and local education agencies.

Measuring the Average Impact of an iPad Algebra Program

A Report of Findings from an RCT in Four School Districts Considering One as a Special Case

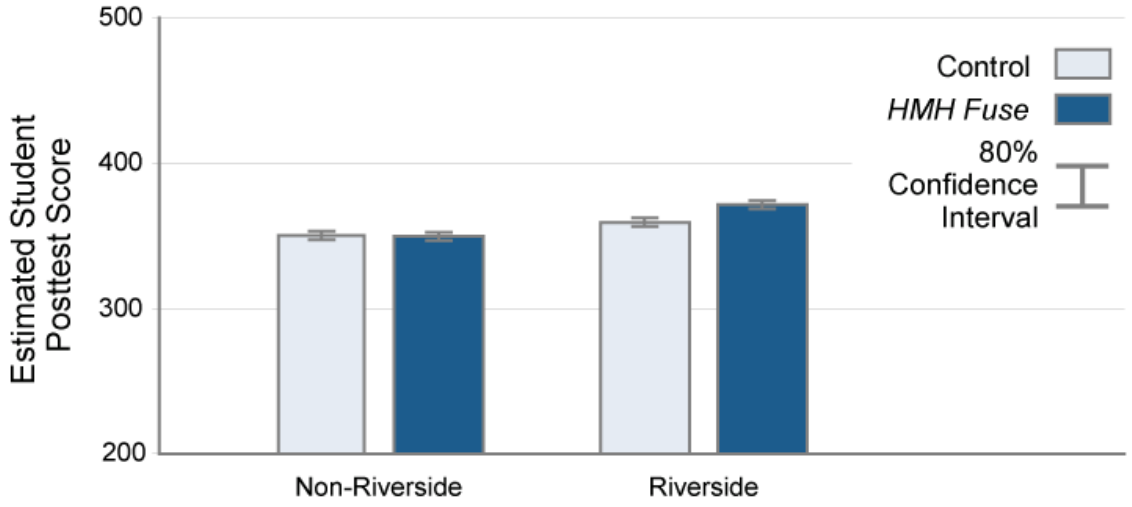
Reference for this report: Jaciw, A., Toby, M., Ma, B., Lai, G., & Lin, L. (2012, March). *Measuring the Average Impact of an iPad Algebra Program: Findings from an RCT in Four School Districts Considering One as a Special Case*. (Empirical Education Rep. No. Empirical_Fuse120330-FR1-YR1_O.3). Palo Alto, CA: Empirical Education Inc.

Executive Summary

Introduction. In spring 2010, Houghton Mifflin Harcourt (HMH) began planning a pilot of an application for the Apple iPad, *HMH Fuse: Algebra 1*, which was then in development. The application was to be piloted in four California school districts during the 2010-2011 school year. HMH contracted with Empirical Education Inc. to conduct a one-year randomized experiment aimed at producing evidence of the effectiveness of *HMH Fuse* for increasing algebra achievement and student attitudes toward math for seventh and eighth grade students.

HMH Fuse for the Apple iPad contains the content of the Holt McDougal Algebra 1 2011© text and includes interactive lessons, explanations, quizzes, and problem solving. In addition, *HMH Fuse* comes with the 300+ videos that are also available online to students using the traditional print version of the text. We compared classes using *HMH Fuse* on the iPad with classes using the conventional text containing the same content.

Findings. We found no impact of *HMH Fuse* on the primary measure of algebra achievement, the California Standards Test (CST), on average across the four districts. One of the school districts, Riverside Unified, initiated its own investigation of the data for the participating students and found what appeared to be a strong impact. We used the same statistical modeling approach to examine impacts for this district and for the other three. For the other three, and consistent with the overall results, there was no discernible difference between *HMH Fuse* and control. For Riverside, however, we found a substantial impact equivalent to a nine-point increase in percentile standing (p = .023). The following figure represents the differential effect of *HMH Fuse* in the other three districts compared to the effect in Riverside.



MODERATING EFFECT OF MEMBERSHIP IN RIVERSIDE ON THE IMPACT OF HMH FUSE ON THE CST

It is also noteworthy that the teachers in Riverside were selected for the pilot on the basis of their experience with technology innovations and reported more time instructing with *HMH Fuse* than reported by most of the other teachers in the study.

On average across the districts, we found no impact on a second measure, the End of Course Assessment. We have some confidence of a positive impact on student attitudes toward math as

measured by the Student Attitude Questionnaire. It is notable that students with positive attitudes toward math were found to achieve higher scores on the CST.

We also gathered implementation data via student and teacher surveys. Conditions for implementation were generally good across both groups; teachers received the necessary materials within the first few weeks of school although many teachers reported technical difficulties. We have some confidence in an impact of *HMH Fuse* on time spent on the algebra program outside the class, number of videos watched, and student attitude towards math. At the end of the school year, nine of the eleven teachers would choose to continue teaching with *HMH Fuse* over the control curriculum.

Research Methods. This was a randomized control trial (RCT) in which we randomly assigned one algebra period for each participating teacher to the program condition, in which they use *HMH Fuse*. Each teacher's remaining algebra sections formed the control group assigned to use the regular text version of the program. Across the four districts we had six schools and 11 teachers. In the control group there were 23 sections of Algebra 1 and 625 students with CST posttests. In *HMH Fuse* group there were 11 sections (one per teacher) and 318 students with CST posttests. Riverside had two teachers with seven control sections with 197 students and two *HMH Fuse* sections with 64 students. Because randomization was blocked by teacher, the two teachers and nine sections in Riverside constituted a very small, yet independent RCT. Statistical modeling took full advantage of the pretest and demographic information to provide appropriate controls and adjustments were made for clustering of students in sections.

Conclusion. After a one-year pilot implementation with *HMH Fuse,* we do not have evidence of a generalizable effect of the program on algebra achievement. We did find clear evidence that the effect was dependent on local conditions. For two teachers in one school—selected for the study on the basis of experience with technology innovations—there was an impact. While we cannot generalize the results beyond these two teachers, the study is suggestive of approaches that may lead to success with applications such as *HMH Fuse*.

Table of Contents

Introduction

In spring 2010, Houghton Mifflin Harcourt (HMH) began planning a pilot of an application for the Apple iPad, *Houghton Mifflin Harcourt Fuse: Algebra 1 (HMH Fuse)*, which was then in development. The application was to be piloted in four California school districts during the 2010-2011 school year. HMH contracted with Empirical Education Inc. to conduct a one-year randomized control trial (RCT) aimed at producing evidence of the effectiveness of *HMH Fuse* for seventh and eighth grade students. We report here on the research conducted in Long Beach Unified School District, Riverside Unified School District, Fresno Unified School District, and San Francisco Unified School District

HMH Fuse for the Apple iPad contains the content of the Holt McDougal Algebra 1 2011© text and includes interactive lessons, explanations, quizzes, and problem solving. In addition, *HMH Fuse* comes with the 300+ videos that are also available online to students using the traditional print version of the text. We compared classes using *HMH Fuse* on the iPad with classes using the conventional text containing the same content.

The specific research questions we addressed are as follows.

- Is *HMH Fuse* effective at increasing mathematics achievement of students in Algebra 1 classes?
- Does the impact of *HMH Fuse* vary for students with different characteristics (i.e. depending on previous achievement, English language learner (ELL) status, and grade level)?
- Are impacts on mathematics achievement associated with impacts on mediating variables (i.e. the amount of time program students spend with materials, number of videos watched)?
- Are differences in the level of use of *HMH Fuse* associated with differences in student mathematics achievement?

For this experimental study, we worked with HMH to recruit 11 middle school teachers who teach two or more Algebra 1 sections. The size of the sample was constrained by the number of iPads available for the study. We randomly assigned one algebra period for each participating teacher to the program condition, in which they use *HMH Fuse*. Each teacher's remaining algebra sections formed the control group assigned to use the regular text version of the Holt McDougal Algebra 1 2011 program.

A randomized control trial eliminates the variety of biases that could otherwise compromise the validity of the research. However, random assignment to experimental conditions does not assure that we can generalize the results beyond the districts where the research was conducted. In the case of this study, differences among the districts were brought to our attention by the independent work by one of the participating districts. This work demonstrated to us that averaging the results across the districts in the study masked critical differences. After seeing these results, which were made public by that district and HMH (Houghton Mifflin Harcourt, n.d.), Empirical researchers conducted an additional analysis, which we report in the final section of the results. Even without this examination of the subgroup results, we would caution that results of an RCT should not be considered to apply to school districts with practices and populations different from those in the experiment. This report provides a rich description of the conditions of implementation to provide the reader with an understanding of the context for our findings.

Methods

Our experiment results in a comparison of outcomes for classes where *HMH Fuse* is in place and classes using the print edition of the Holt McDougal Algebra 1 text. The outcomes of interest are student scores on the algebra California Standards Test (CST) and the Holt Algebra 1 End of Course Assessment.

This section details the methods used to assess, at a specific level of confidence, the size of the difference in outcomes and whether the introduction of *HMH Fuse* is responsible for those differences. We begin with a description and rationale for the experimental design and go on to describe the intervention, the research sites, the sources of data, the composition of the experimental groups, and finally the statistical methods used to generate our conclusions about the impact of *HMH Fuse.*

Experimental Design

Due to the challenges inherent in recruiting schools and the voluntary nature of any experimental study, the sample was largely one of convenience. Thus, for this study, any inferences beyond the sample should be made on strong heuristic grounds, if at all. It is important to recognize that the study results could change if we were to select a new sample. The design of the experiment is based on our best understanding of the amount of variability in outcomes that we expect that is not attributable to the program, and we attempt to detect the stable signal (the effect) if it exists by limiting the effect of this variation. There is always a level of uncertainty and an associated level of imprecision in our estimates of the effects of the program. We relate the uncertainty to the likelihood that we would obtain a different result if we took a new sample of sections from the same larger population. Our design attempts to efficiently deploy the available resources to reduce uncertainty and improve precision—in other words—to reduce the likelihood that we would obtain a different result if we tried the experiment again.

Before beginning the experiment, we created a design or plan in which we establish the specific questions to be answered.

First, before seeing the results, we identify where we expect to see an impact and which factors we expect will moderate or mediate the impact. In other words, we specify the research questions up front. In this way, we avoid fishing for results in the data, a process that can lead to mistaking chance differences for differences that are probably important as a basis for decisions. Because some effects will be big simply by chance, mining the data in this way can capitalize on chance—concluding that there is an effect when really we're just picking the outcomes that happen to be big as a result of chance variation. We can still explore the data after the fact, but this is useful mainly for generating ideas about how the new program worked; that is, as hypothesis-generating efforts for motivating future study, rather than as efforts from which we make firm conclusions from our existing study.

Second, an experimental design will include a determination of how large the study should be in terms of units such as students, teachers, or schools in order to get to the desired level of confidence in the results. In the planning stage of the experiment, we calculate either how many cases we need to detect a specifically sized difference between the *HMH Fuse* and control groups, or how big a difference we can detect given the sample size that is available. Technically, this is called a power

analysis. We will explain several aspects of the design and how they influence the sample size needs for the experiment.

Randomization

Since we want to know the impact of *HMH Fuse*, we have to isolate its impact from all the other factors that might make a difference for how or what teachers and students do. We want to determine whether *HMH Fuse* caused a difference. Randomization ensures that, on average, characteristics—other than the program—that affect the outcome are equally distributed between program and control groups. This distribution prevents us from confusing the program's effects with some other factors—technically called confounders—that, because they also affect the outcome, would lead to bias if they are unevenly distributed between the groups.

There are various ways to randomize to experimental conditions. Our research works within the organization of schools, not disrupting the existing hierarchy in which students are grouped within sections that are nested under teachers and schools. Randomizing students individually would not be feasible given how the program is implemented. The level in the hierarchy at which we conduct the randomization is generally determined on the basis of the kind of program being tested. We attempt to identify the lowest level at which the program can be implemented without unduly disrupting normal processes or inviting sharing or "contamination" between control and program units. For example, school-wide reforms call for a school-level randomization while a professional development program can use a teacher-level randomization.

For this experiment, we randomized classes under teachers who volunteered for participation to the *HMH Fuse* and control groups. Because classes, instead of students, were assigned to *HMH Fuse* or the control materials, this kind of experiment is often called a "group randomized trial."

What Factors May Moderate the Impact of *HMH Fuse*?

The selected design allows us to measure the differential effectiveness of *HMH Fuse* for specific subgroups of students. We planned to compare the program's effectiveness based on pretest scores, whether students are classified as English language learners, and possibly grade level. These are variables that were measured before the experiment started, and that we had reason to believe would affect the magnitude of the effect of *HMH Fuse*. Technically, these are called potential moderators, because they may moderate (increase or decrease) the impact of *HMH Fuse*. We measure the effect of the interaction between each potential moderator and the variable indicating assignment (i.e., to *HMH Fuse* or control); that is, we measure whether the effect of *HMH Fuse* changes across levels of each moderator. The study is not specifically designed to detect differential impacts, therefore, if we do not observe an effect, this may indicate that we do not have a large enough sample (i.e., sufficient power) to detect the difference.

What Factors May Mediate Between *HMH Fuse* and the Outcome?

We also identified variables that we believed would facilitate the effect of *HMH Fuse* on student outcomes and that could only be measured after the experiment had started. In this experiment, we measured the number of Holt 1 Algebra videos students watched, student attitudes toward mathematics, and time spent with algebra materials. Technically, these are called mediators and are themselves intermediate outcomes, measurable in both assignment groups, which may be impacted by *HMH Fuse*.

We usually think of a "mediator" as a factor in *how* the program has an impact. Based on the nature of the intervention, we identified process variables or mediators that were likely to facilitate the

overall impact of the program. We pre-identified these mediators and tested whether there was a difference between the program and control group processes. We then used this information to draw further conclusions about whether the difference in the final outcome was facilitated through an impact of the program on the mediating process. Because we don't assign cases to levels of the mediator, we cannot be sure whether it is a proxy for an intermediate effect that we have not identified. The study is not specifically designed to detect mediating effects, therefore, if we do not observe an effect, this may indicate that we do not have a large enough sample (i.e., sufficient power).

How Large a Sample Do We Need?

A process called power analysis was used to plan the number of classes that the experiment would need in order to say with specific levels of confidence that the intervention has an impact. This is an important part of experimental design, and here we walk through the factors considered.

How Small an Impact Do We Need?

The size of the sample required for a study depends on how small an effect we need to detect. Experiments require a larger sample to detect a smaller impact. It is important to know the smallest potential impact that would be considered educationally useful in the study's particular setting. It is necessary to decide in advance on this value as part of the power analysis since the number of cases needed in the sample is related to how small an effect we need to detect. Conversely, if we had a fixed number of cases to work with, we would want to know how small an effect we could detect—the so-called "minimum detectable effect size" (MDES). Once the sample size is determined based on the MDES, or once the MDES is established based on the available sample, it remains possible that effects exist that we cannot detect because they are smaller than the MDES. We designed this experiment to detect an effect of 0.25, measured in standardized effect size units.

How Much Variation Exists Between Classes

When we randomize at the class level but the outcome of interest is a test score of students associated with those classes, we pay special attention to the differences among classes in class average scores. The greater the variation in class averages of student scores, the more classes we need in the experiment to detect the impact of the intervention. This is because the extra variation among classes adds noise to our measurement which makes the effect of the intervention—the signal—harder to detect. A summary statistic that tells us how the variation is divided up among levels of analysis is the intraclass correlation coefficient (ICC). Technically, it is the ratio of the variation in the class averages of students' scores to the total variation in students' scores. A larger ICC means between-class differences in student posttest scores contribute more noise to our program effect estimate. A larger sample of classes is then needed to dampen the noise to acceptable levels. We determine a value of the ICC before randomization. For this experiment we assumed a fairly conservative intraclass correlation of .10.

How Much Value Do We Gain From a Pretest and other Covariates?

In order to estimate effects of interest with additional precision, we make use of other variables that we know will impact performance. These are called covariates. By including covariates in the analysis, we increase the precision of our effect estimates by removing variation in the results that is accounted for by the covariates. Technically, a covariate-adjusted analysis is called an analysis of covariance (or ANCOVA). In our experiments, a student's score on a pretest (which may be a test in a subject that is closely related to the outcome measure rather than the same test but given earlier) is

almost always the covariate most closely associated with the outcome. In almost all of our analyses, we adjust for the effect of the pretest, which is a strong predictor of posttest performance. We may include other covariates as well. In this experiment, we assumed a fairly substantial correlation between the pre- and posttests (.80).[1] In a power analysis determining the number of classes we will need, greater capacity of covariates to predict posttest performance yields greater precision and thereby requires fewer classes to detect the same level of impact.

Are There Subgroups of Particular Interest?

Often we are interested in whether a program has more of an impact for a particular subgroup than for others. Usually we have more statistical power for detecting differential impacts if the subgroups exist within the randomized cases (e.g., subgroups of students within sections), than if the subgroups are identified at the level of randomization (e.g., different types of sections). In the latter case we will need to include more units of randomization in the experiment in order to have enough cases of each type. In the current experiment, we are interested in whether the impact varies depending on the pretest scores and on whether students are classified as English language learners. To examine whether impacts vary by grade level, we would have to compare impacts for subsets of cases randomized. The number of cases in each subgroup would be too small to allow an adequately powered analysis, even as an exploration.

How Much Confidence Do We Want to Have in our Results?

We want to be certain that if we conclude there is no impact that, in fact, there is no impact (we want to limit the possibility of drawing a false negative conclusion). Also, we want to be certain that if we conclude there is an impact, that, in fact, there is an impact (we want to limit the possibility of drawing a false positive conclusion). Conventionally, researchers have given priority to avoiding false positive conclusions, requiring differences large enough that they would be seen 5% of the time in the absence of an effect before concluding that there is an effect; while at the same time, allowing a conclusion of no effect when in fact there is an effect, 20% of the time. For the power analysis we adhere to these criteria. However, our conclusions reached about the presence of an effect are expressed in terms of levels of confidence (strong, some, limited or none) rather than as a yes-or-no declaration. As we describe later, we interpret results in terms of whether they give a lot, some, limited, or no confidence that there is a true impact.

Sample Size Calculation for This Experiment

Taking all the above factors into consideration, and with the number of teachers and sections that were available for this study, we estimated that the smallest effect size that we can detect is an absolute difference of ten percentile points for algebra for a student who performs at the median of the distribution: this effect size is what we would see if we took a student who performs at the 50th percentile of the distribution of posttest performance for the program group and found that student's score to be 10 percentile points higher (i.e., at the 60th percentile) or 10 percentile points lower (i.e., at the 40th percentile) than the median score for the control distribution. As we explain later in this section, we can also express this difference as a standardized effect size, that is, as a

[1] That is, we assume that .80*.80=.64 is the proportion of variance in the outcome (i.e., the R-squared) that is accounted for by the covariate in either condition.

proportion of the standard deviation of posttest performance. In terms of that metric, the MDES for algebra is 0.25.

Our research design assumed that we would report the results for the four districts combined. The sample size calculation was conducted using Optimal Design (Raudenbush, Spybrook, Liu, & Congdon, 2006), a software program developed for this purpose.

The power analysis reflects the numbers that were expected prior to the start of the trial. In actuality, 34 sections were randomized, rather than the anticipated 35. The effect on statistical power of having one fewer unit than expected is small.

Participant Recruitment and Site Descriptions

How the Sample was Identified

How the participants for the study are chosen largely determines how widely the results can be generalized. In this case, HMH recruited into the study four school districts that were using another edition of the Holt Algebra 1 print textbook. Therefore, our sample here is one of convenience. Districts that agreed to participate then selected the schools and teachers to be included in the research project. This was left to the district's discretion and individual districts did not necessarily use the same processes or criteria to identify participants.

Long Beach Unified School District

Long Beach Unified School District (LBUSD) is located in Long Beach, California. Long Beach is a large city located approximately 20 miles south of downtown Los Angeles. The city's total population is 494,709 (California Department of Finance, 2011). LBUSD's operating budget was $967,896,000 in 2008 and the per-pupil expenditure was $9,646, ranking Long Beach 329th out of 979 districts in the state for per-pupil expenditures (Federal Education Budget Project, 2011). LBUSD has 91 schools with a total enrollment of 87,509 students. Table 1 provides information about the entire district including the schools that participated in the study.

TABLE 1. DEMOGRAPHICS OF LONG BEACH UNIFIED SCHOOL DISTRICT

Long Beach Unified School District	
Total schools	91
Total full-time equivalent teachers	3,897.3
Student to teacher ratio	22.5
Student population	87,509
ELL students	23.7%
White	16.1%
Black	17.1%
Hispanic	51.6%
Asian	8.1%
Pacific Islander	1.9%
Filipino	3.7%
American Indian/Native Alaskan	0.2%
Multi racial/No response	1.2%

Source: California Department of Education 2009-2010 school year

Note. Percentages may not add up to100% due to rounding of decimals

Riverside Unified School District

Riverside Unified School District (RUSD) is located in Riverside, California. Riverside is a large city located approximately 60 miles east of Los Angeles. The city's total population is 304,051 (California Department of Finance, 2011). RUSD's operating budget was $407,551,000 in 2008 and the per-pupil expenditure was $8,268, ranking Riverside 722nd out of 979 districts in the state for per-pupil expenditure. RUSD has 48 schools with a total enrollment of 43,336 students (Federal Education Budget Project, 2011). Table 2 provides information about the entire district including the schools that participated in the study.

TABLE 2. DEMOGRAPHICS OF RIVERSIDE UNIFIED SCHOOL DISTRICT

Riverside Unified School District	
Total schools	48
Total full-time equivalent teachers	1,886.5
Student to teacher ratio	23
Student population	43,336
ELL students	18.7%
White	30.4%
Black	9.2%
Hispanic	53.6%
Asian	3.3%
Pacific Islander	0.6%
Filipino	1.2%
American Indian/Native Alaskan	0.6%
Multi racial/No response	1.1%

Source: California Department of Education 2009-2010 school year

Fresno Unified School District

TABLE 3. DEMOGRAPHICS OF FRESNO UNIFIED SCHOOL DISTRICT

Fresno Unified School District	
Total schools	105
Total full-time equivalent teachers	3,831
Student to teacher ratio	20
Student population	76,621
ELL students	26.0%
White	13.9%
Black	10.7%
Hispanic	60.1%
Asian	13.4%
Pacific Islander	0.4%
Filipino	0.4%
American Indian/Native Alaskan	0.7%
Multi racial/No response	0.3%

Source: California Department of Education 2009-2010 school year

Fresno Unified School District (FUSD) is located in Fresno, California. Fresno is a large city located in the center of the San Joaquin Valley. The city's total population is 502,303 (California Department of Finance, 2011). FUSD's operating budget was $851,431,000 in 2008 and the per-pupil expenditure was $10,053, ranking Fresno 266th out of 979 districts in the state for per-pupil expenditure (Federal Education Budget Project, 2011). FUSD has 105 schools with a total enrollment of 76,621 students. Table 3 provides information about the entire district including the schools that participated in the study.

San Francisco Unified School District

San Francisco Unified School District (SFUSD) is located in San Francisco, California. San Francisco is a large city located in northern California. The city's total population is 896,095 (California Department of Finance, 2011). SFUSD's operating budget was $597,176,000 in 2008 and the per-pupil expenditure was $9,711, ranking San Francisco 320th out of 979 districts in the state for per-pupil expenditure (Federal Education Budget Project, 2011). SFUSD has 112 schools with a total enrollment of 55,183 students. Table 4 provides information about the entire district including the school that participated in the study.

TABLE 4. DEMOGRAPHICS OF SAN FRANCISCO UNIFIED SCHOOL DISTRICT

San Francisco Unified School District	
Total schools	112
Total full-time equivalent teachers	2,985.1
Student to teacher ratio	18.5
Student population	55,183
ELL students	30.5%
White	10.8%
Black	12.3%
Hispanic	23.1%
Asian	41.3%
Pacific Islander	1.3%
Filipino	5.8%
American Indian/Native Alaskan	0.6%
Multi racial/No response	4.8%

Source: California Department of Education 2009-2010 school year

Note. Percentages may not add up to 100% due to rounding of decimals

HOUGHTON MIFFLIN HARCOURT FUSE: ALGEBRA 1

HMH Fuse consists of an iPad preloaded with an application containing the content of the Holt McDougal Algebra 1 program, plus additional interactive features. Participating teachers received a one-day training on the program and were offered technical support throughout implementation.

Training/Professional Development

Participating teachers were invited to a one-day initial training to learn the basic operation and features of *HMH Fuse*. Trainings for the four districts occurred separately at each site and were attended by all participating teachers. In most cases, district administrators and other local observers were also present, outnumbering the teachers in the room. All four training sessions took place during the first two weeks of September 2010. Representatives from Apple, Houghton Mifflin Harcourt, EduSoft, and Empirical Education conducted the various components of the day's training. The agenda held the following format.

- Meet and Greet and Introduction (HMH)
- iPad in the Classroom (Apple) – Instances of the iPad as implemented in various educational settings
- How to use the iPad (Apple) – Basic functions of the iPad

- *Houghton Mifflin Harcourt Fuse: Algebra 1* (HMH) – Navigating the application and its various features
- Introduction to the Research Study (Empirical Education) – Introducing the research design and participant responsibilities; obtaining participant information and consent
- Assessment (EduSoft) – Assessment plan and protocols
- Wrap-up and Q&A (all presenters)

***Houghton Mifflin Harcourt Fuse: Algebra 1* Materials**

Houghton Mifflin Harcourt Fuse: Algebra 1 is an application for the Apple iPad that contains the complete content of the Holt McDougal Algebra 1 text. In addition, the application provides interactive lessons, explanations, quizzes, and problem solving. *HMH Fuse* comes preloaded with the 300+ videos that are available online to students using the print version of the text. *HMH Fuse* contains a variety of interactive tools such as Graphing Equations, Quadratic Explorer, Linear Explorer, and Algebra Tiles which allow students to manipulate variables and see the results. The note-taking feature allows students to type in notes, color code for organization, and leave themselves recorded voice messages. By touching a vocabulary word within a lesson, students are brought to the glossary where they are provided with a definition for the term. Are You Ready? quizzes test specific skills before a student begins a chapter and are accompanied by a scratchpad to help with calculations or writing notes. Students are prompted to review, practice, and retest skills. Icons on the sidebar provide tips and links for support, such as view-in-motion explanations or videos that address concepts from a different approach. The application also contains a Search function.

Control Materials

The control materials consist of the print edition of the Holt McDougal Algebra 1 2011 program and online access to videos. Each lesson within the textbook includes levels of skill development for differentiated instruction. The print text also provides Are You Ready? and Ready to Go On? quizzes. Participating classrooms received new sets of the textbook at the beginning of the school year.

Common to Both Assignment Groups

All teachers received the Holt McDougal Algebra 1 Assessment Resource (Guide) which contains tests for all 11 chapters, as well as a placement test and an End of Course Assessment. Teachers sent completed student assessments to the testing company which scored the assessments and provided online access to results. Teachers received a login to retrieve report data for students in both assignment groups. The Class List, Performance Band, and Student Performance reports provide information on individual student performance, overall performance, group averages, and per standard scores. These data were available for all internal Holt McDougall Algebra 1 assessments. Both the print textbook and *HMH Fuse* have 11 chapters.

Expectations of Implementation

Teachers were expected to use *HMH Fuse* as the core math program in classrooms assigned to the *HMH Fuse* group. Similarly, teachers were expected to use print edition of the Holt McDougal Algebra 1 2011 program as the core math program in classes assigned to the control group.

Schedule of Major Milestones

Teachers were provided with iPads for the training but were required to return the devices at the day's end. In the following week, media events occurred in each of the four districts. During the event, presenters from HMH, as well as one of the textbook's authors, Dr. Ed Burger, introduced the research study and *HMH Fuse*. Students received iPads to try out and follow along during the author's presentation. These devices were also returned at the end of the media event. Deployment teams returned to each district in the following weeks to distribute the devices and ensure everything worked properly. While two of the districts experienced some delay in receiving the iPads, all participating classrooms had program materials by September 20, 2010.

Table 5 lists the major project milestones and associated dates.

TABLE 5. RESEARCH MILESTONES

Date	Milestone
August 30 – September 9, 2010	Teacher trainings in four districts
September 8 – 10, 2010	Media events in four districts
September 13 – 17, 2010	Deployment teams distribute devices in four districts
September 15 – November 1, 2010	Teachers administer student placement test and student attitude questionnaire (pre)
September 2010	Initiation of chapter tests
September 2010	Begin collecting parental consent for participating students
September 2010	Obtain district agreements and approval to conduct research
September 24, 2010	Initiation of monthly teacher surveys
November 15, 2010	Begin collecting district rosters, demographics, and CST pretest data
May 13, 2010	Final monthly teacher survey deployed
May 2010	Teachers begin to administer student attitude questionnaire (post)
June 2010	Teachers begin to administer End of Course Assessment
August 10, 2011	Request CST posttest data

Data Sources and Collection

The data for this study consist primarily of CST algebra scores, as well as assessment scores from the Holt McDougal Algebra 1 program. In addition, we collected student survey data for each of the 11 chapters, a pre/post measure of student attitudes toward mathematics, nine web-based teacher surveys, and log data from the *HMH Fuse* application to track student usage of the device.

Class Rosters and Demographic Data

Researchers collected class roster and demographic data in order to conduct balance checks, to analyze student data nested within section within teacher within school, and to conduct moderator and mediator analyses. Specifically, the districts were asked to provide the following student data.

- Name

- Unique identifier
- Gender
- Ethnicity
- English proficiency status
- Disability status (whether or not student has a disability or is in special education, but not the specific condition)
- Age
- Grade
- Classroom teacher
- Course name and section
- School
- CST scores

All student and teacher data having any individually identifying characteristics were stripped of such identifiers, and the data were stored using security procedures consistent with the provisions of the Family Educational Rights and Privacy Act (FERPA).

Outcome Measures

We employed two measures of student achievement to determine whether *HMH Fuse* is effective at increasing mathematics achievement of students in Algebra 1 classes: the California Standards Test and the Holt McDougal Algebra 1 End of Course Assessment. Researchers also used the Student Attitude Questionnaire to determine if *HMH Fuse* has an impact on students' attitudes toward math. Chapter test data were collected for exploratory analyses investigating whether *HMH Fuse* has an impact on chapter test performance.

The California Standards Test

According to California's Standardized Testing and Reporting (STAR) website:

> The CSTs are a major component of the STAR program. The CSTs are developed by California educators and test developers specifically for California. They measure students' progress toward achieving California's state-adopted academic content standards in English language arts (ELA), mathematics, science, and history–social science, which describe what students should know and be able to do in each grade and subject tested (California Department of Education, 2009).

Because the test is linked to California's standards, there is no national comparison. Students receive a scale score between150 and 600. Based on this scale score, California uses five performance levels to report student achievement on the CSTs: Advanced, Proficient, Basic, Below Basic, and Far Below Basic (California Department of Education, 2011). Students take a mathematics assessment in all grades 2 through 11; in grades 8 through 11, student testing is dependent on the mathematics course in which the student is enrolled. CST scores from the 2009-2010 school year were used as a pretest measure, and scores from the 2010-2011 school year—the year of *HMH Fuse* implementation—were

used for the posttest. At the start of the trial, HMH identified the CST as the primary outcome measure for this project.[2]

Holt McDougal Algebra 1 End of Course Assessment

The Holt McDougal Algebra 1 Assessment Resource Guide contains a placement test, which researchers used as a pretest, and an End of Course Assessment, which researchers used as a posttest. The assessment book accompanies the California edition of the Holt McDougal Algebra 1 text.

Student Attitude Questionnaire

According to HMH, increased student engagement is a key intermediate outcome. That is, students who use *HMH Fuse* will become more engaged in mathematics lessons, and engagement will then lead to improved student achievement. Therefore, students participated in a survey designed to measure their attitudes about math. Participating students responded to a baseline measure at the beginning of the school year and the same questionnaire again at the end of the school year. The goal was to see if student attitudes about math changed over time and if this change is associated with group assignment. This measure was created using five subscales from the Motivated Strategies for Learning Questionnaire and two subscales from the Attitudes toward Mathematics Inventory. Researchers obtained active parental consent prior to administering the survey.

Questions from both main scales were included in the Student Attitude Questionnaire, but the results were analyzed separately by scale. That is, we decided to not combine the two scales and produce a total score based on the combined items, because we do not have information to indicate that it would be technically correct to merge the scales. Doing so could compromise the constructs intended to be measured by each scale.

Motivated Strategies for Learning Questionnaire
The Student Attitudes Questionnaire included the following five subscales from the Motivated Strategies for Learning Questionnaire: Self-Efficacy, Intrinsic Value, Test Anxiety, Cognitive Strategy Use and Self-Regulation. We examine the impact of *HMH Fuse* on the combined score of the five subscales as well as on the individual subscales.

Attitudes toward Mathematics Inventory
The Attitudes toward Mathematics Inventory included two subscales: Self-Confidence and Enjoyment. We were unable to look at the impact for the two scales combined because an error in the construction of the questionnaire caused student responses to the final questions of the measure to be uninterpretable: the questionnaire that students received contained no Question 47 - the measure skipped from Question 46 directly to Question 48 - but the answer sheet did have a Question 47, which some students filled out. We could not be sure how students addressed this problem. Some may have noticed the discrepancy and skipped Question 47 on the answer sheet, while other students may have entered the intended response for Question 48 into Question 47 on the answer sheet. Because Questions 47-51 on the answer sheet were rendered unusable in analysis, we can only report on the items prior to Question 47. Those questions consist of two (out of three

[2] In the terminology established for Institute of Education Sciences research (Schochet, 2008), analyses of average impacts on CST are considered confirmatory.

total) items from the Self-Confidence subscale. None of the items from the Enjoyment subscale, which all occurred after Question 47, could be analyzed.

Holt McDougal Algebra 1 Chapter Tests

The assessment book included 11 chapter tests which teachers were expected to administer upon completion of each chapter taught. Teachers are not expected to teach the chapters in order, and the teachers do not always teach all of the chapters. Since we have a number of outcome variables, and since we do not adjust for multiple comparisons, we determined at the outset that this scale would be considered exploratory.

Testing Schedule

For the CST, we used scores available from the previous year's spring testing (2009-2010 school year) as a pretest measure and the scores from spring 2011 as the outcome measure.

HMH assessment kits were delivered to the participating teachers at their schools within the first week of class. The delivery date was different for each district, but generally fell between the middle of August and early September 2010. The first assessment kit contained the pretest, Student Attitude Questionnaire (pre), and tests for chapters one through seven. The remaining tests were delivered later in the school year in two separate shipments: 1) chapters eight and nine; 2) chapters 10 and 11, along with the End of Course Assessment and Student Attitude Questionnaire (post). Teachers returned completed scantrons in Federal Express envelopes provided to them by the testing company (EduSoft). The testing company then translated the student test data into electronic form and transferred the data to the research company. Teachers were required to administer the pre-assessment first but were permitted to select their own sequence for teaching the chapters and administering the subsequent assessments (once they had received the assessment materials). Therefore, not all classrooms taught and tested the chapters in the same order.

The timeframes listed in Table 6 reflect the dates when the majority of tests were received.

TABLE 6. ASSESSMENT COMPLETION DATES

Date	Assessment
September 15 – November 1, 2010	Placement Test
September 2, 2010 – January 25, 2011	Student Attitude Questionnaire (pre)
September 10, 2010 – October 20, 2010[a]	Chapter 1
October 12, 2010 – December 16, 2010[a]	Chapter 2
October 14, 2010 – June 1, 2011	Chapter 3
October 21, 2010 – January 25, 2011	Chapter 4
November 16, 2010 – January 24, 2011[a]	Chapter 5
January 10, 2011 – March 8, 2011[a]	Chapter 6
September 22, 2010 – May 17, 2011	Chapter 7
March 2, 2011 – June 3, 2011	Chapter 8
March 15, 2011 – June 3, 2011	Chapter 9
May 17, 2011 – June 3, 2011	Chapter 10
June 1 – 9, 2011	Chapter 11
June 1 – 9, 2011	End of Course Assessment
May 9 – 15, 2011	Student Attitude Questionnaire (post)

[a] Dates noted with an asterisk had a small number of tests (fewer than 15 students) received outside of these timeframes, all at the end of the academic year in June.

Program Implementation Measures

In addition to achievement and student attitudes data, we also collect implementation data over the entire period of the experiment, beginning with the teacher trainings and ending with the academic calendar of the district in June 2011. Data collected through training observations, multiple teacher surveys, student surveys, the *HMH Fuse* device log, email exchanges, and phone conversations are used to provide both descriptive and quantitative evidence of the implementation.

Teacher Training Observation

Researchers observed the initial teacher training in three of the four districts and asked additional questions about the initial training through the teacher online surveys.

Teacher Survey Data

Surveys were deployed to participating teachers beginning in September 2010 and continued on a monthly basis through May 2011. Table 7 outlines the survey schedule and final response rates.

TABLE 7. SURVEY SCHEDULE

Survey	Deployment	Response Rate
Survey 1	September 24, 2010	100%
Survey 2	October 22, 2010	100%
Survey 3	November 12, 2010	100%
Survey 4	December 10, 2010	100%
Survey 5	January 14, 2011	100%
Survey 6	February 18, 2011	91%
Survey 7	March 18, 2011	91%
Survey 8	April 8, 2011	82%
Survey 9	May 13, 2011	100%

The survey topics were developed to account for the various aspects of teacher and student actions associated with instruction and learning. In order to characterize the average time teachers and students spent using *HMH Fuse* and control materials, we used a repeated question strategy. Questions inquiring about the number of videos watched were also repeated in surveys two through nine. We report quantitative survey data using descriptive statistics and, where appropriate, we employ tests of significance to compare the results for the two conditions (*HMH Fuse* and control). Questions regarding minutes of instruction and interaction with study materials are used in mediator analyses. The free-response portions of the surveys were minimally coded. Our analyses of the survey data fall into the following categories.

- Teacher Background
- Implementation Conditions
- Implementation Fidelity and Extent of Program Implementation

- Comparison of Classroom Implementation between *HMH Fuse* and Control Groups
- Teacher Satisfaction with *HMH Fuse*

Teacher Background

Because literature correlates teaching experience and content knowledge with teacher quality (Amrein-Beardsley, 2006; The Center for Public Education, 2005), this study collects teacher background data to provide a context for understanding the study results.

We collected the following teacher background data.

- Education level completed
- Credentials and certification
- Years of teaching experience

During the initial training for the research study, teachers received a Participant Information Packet which provided general information about the research study, data collection activities, and participant responsibilities. The packet also included a teacher consent form which was completed and collected at the training. In addition, teachers filled out a Teacher Background Form, providing researchers with information about their teaching history and contact information. Analysts used these data to perform balance checks.

Implementation Conditions

It is critical to interpret both the implementation and the outcome data with an understanding of the context in which implementation took place. Researchers constructed survey items specifically to understand the conditions under which teachers implemented *HMH Fuse* and the control curriculum. We collected survey data on program training, availability of materials, and initial issues with technology and support. We were specifically interested in the extent to which technical issues precluded implementation in *HMH Fuse* classes. We also surveyed teachers about their initial impressions of, and level of comfort with, *HMH Fuse*.

Implementation Fidelity and Extent of Program Implementation

In this study, fidelity of implementation was very clearly defined: the *HMH Fuse* students were to use only the iPads and the control students were to use only the print version of the text plus the on-line videos. Therefore, researchers examine the extent to which the *HMH Fuse* teachers and students used only *HMH Fuse* as their core Algebra curriculum. The researchers were also aware of the types of technology problems that could limit the implementation of technology-based programs. We therefore believed that technological obstacles could result in *HMH Fuse* students using control materials in lieu of the device to which they were assigned. In this scenario, any potential impact might be watered down due to decreased use of the product.

Researchers also looked for instances of contamination, where students in the control group interacted with *HMH Fuse*. While only students in the *HMH Fuse* group were assigned iPads containing the application and teachers understood that control students should not interact with *HMH Fuse*, contamination is always a potential impediment to reliable study results. In the event of contamination, any potential impact from the introduction of the program might be less evident due to control students also receiving some of the benefit (or detriment).

In addition, so that readers understand the types of activities that lead to the study results, researchers asked teachers which of the various features of *HMH Fuse* they used in their classrooms.

Comparison of Classroom Implementation between *HMH Fuse* and Control Groups

Researchers hypothesized that students using *HMH Fuse* would be more engaged, and therefore would spend more time learning algebra, leading to increased student achievement. Therefore, we investigated the number of minutes students spent learning algebra in each of the two conditions and whether the number of minutes of algebra instruction is associated with differences in student outcomes. Monthly surveys asked teachers to report the amount of time they spent using the program for active classroom instruction, in addition to the amount of time students spent working independently with the program during class time. Identical time questions were asked across eight surveys to gain an understanding of variation at different times during the school year, as well as to construct a stable average to compare the assignment groups. In addition researchers compared the number of videos watched in class and teacher time spent planning for instruction for each assignment group.

Teacher Satisfaction with the *HMH Fuse* Program

Finally, since teacher satisfaction is an important factor in decisions regarding program adoption, we asked teachers about their satisfaction with the two algebra programs, as well as whether they would recommend each program to other algebra teachers. The final survey also asked teachers whether they would continue teaching with *HMH Fuse* over the control curriculum, if given the chance.

Student Repeater Survey

Students completed a short (seven item) survey which was administered with each of the 11 chapter tests. Through these surveys we asked students (in both assignment groups) whether they have used *HMH Fuse.* In addition, we asked students to report the amount of time they spent doing algebra work during that chapter. Again, the program developers hypothesized that the *HMH Fuse* students would be more engaged, do more math work, and consequently achieve better mathematics scores.

One of the program's primary means of engaging students is the use of videos. Students in both conditions had access to the same videos, but control students had to log in online to view the videos, while *HMH Fuse* students had the videos readily accessible within the application. Therefore, through these surveys we asked both assignment groups to report the number of algebra videos they watched over the course of the unit and the reasons for watching videos.

Student Device Log Data

HMH provided researchers with log data from the student devices. Such data supplies researchers with information regarding the number of times individual students use distinct features of the application. For this research study, we explore and report on the following features which were selected by the program developer: answer checking on homework pages, homework help walkthrough, inline example references, quiz usage and results, video usage, Math in Motion usage, and adding notes through the note-taking tool. In addition, researchers report on the total usage or overall usage for participating students, as determined by the number of times a student clicks on anything within the application. We also explore whether there is a relationship between these usage variables and student performance on the three outcome measures.

Since the device's data log tracks student use of the application even while the student is outside of the classroom, researchers collected parental consent prior to accessing this data.

Formation of the Experimental Groups

This section describes the study sample that we will use to assess the impact of *HMH Fuse*. The sample of primary interest consists of the participating sections that were randomly assigned to *HMH Fuse* or control. This constitutes the baseline sample. The sample that is analyzed for a given outcome may be modified somewhat from baseline, through attrition or loss of units at different points during the experiment, for a variety of reasons.

Baseline Sample

Ideally, when assignment is randomized into the two conditions, the groups should look the same in terms of important characteristics, such as demographic composition, prior achievement, and other section characteristics. In addition, because we randomized sections within blocks (teachers), we can expect somewhat better balance than we would have if we hadn't randomized in this way. However, by chance (and because sections are not identical) the groups are never exactly balanced and may differ on important characteristics likely to affect the outcome.

Therefore, in this section we inspect the distribution of background characteristics for sections and students, looking in particular at whether these characteristics are balanced between the *HMH Fuse* and control groups.

In Table 8 we compare the composition of the control and *HMH Fuse* groups at the point we received the rosters (baseline sample). For each of the characteristics of this sample, we conducted a statistical test[3] to determine the likelihood of obtaining a chance imbalance as large as or larger than the one observed. While the randomization assures us that any imbalance was a result of chance, and is not an indication of selection bias, it is useful to examine the actual groups as formed at baseline to see whether the amount of imbalance is something we would expect to see less than 5% of the time. We see that balance is achieved on the observed characteristics.

TABLE 8. CHARACTERISTICS OF STUDY SAMPLE ROSTERS RECEIVED (BASELINE SAMPLE)

Student characteristics	Control	*HMH Fuse*	Less than 5% chance of seeing this much imbalance
Male	342 (51.51%)	163 (48.80%)	No
Grade 8	562 (84.64%)	278 (83.23%)	No
Disability	14 (2.11%)	5 (1.5%)	No
English speaker	605 (91.25%)	288 (86.23%)	No
Asian	179 (27.50%)	88 (26.59%)	
White	152 (23.35%)	44 (13.29%)	
Black	53 (8.14%)	33 (9.97%)	No
Mixed	21 (3.23%)	6 (1.81%)	
Indian	2 (0.31%)	0	
Hispanic	244 (37.48%)	160 (48.34%)	
Mean pre-test score	0.00	-0.03	No

Note. The following information is missing. English Speaker: 1 student; Ethnicity: 16 students; pretest: 42 students

[3] We used a *t* test that adjusted for clustering of students in sections. The criterion for significance was set at <.05.

Analytical Samples

Since some teachers or sections or students may be lost during the experiment, the analytical sample is the set of units actually available for statistical analysis for each of the outcomes. The loss of units randomized – in this case sections – during the experiment may cause the difference between conditions on the outcome to reflect imbalance on background characteristics, instead of differences caused by being exposed to *HMH Fuse.*

If the rate of overall attrition is large, even if there is no difference between conditions in the rate of attrition, then a loss of cases may induce bias in the result, if those who leave the program group are different from those who leave the control group. Therefore we adjust for this difference in the analysis. For example, we would want to adjust for the effect of the pretest if sections that attrite from the control group on average have lower achievement than sections that attrite from the program group.

If the rate of differential attrition is substantial, even if those who leave the two conditions are not fundamentally different, then the difference in the rate of attrition can induce bias in the result. Therefore we adjust for the characteristics that may end up being imbalanced between conditions as a result of the loss of cases. For example, we would want to adjust for the effect of the pretest if a larger proportion of low-performers leave the program group compared to the control group.

To assess the potential for bias, we assess whether the levels of overall and differential attrition at the level of randomization are large enough to be likely to induce bias by What Works Clearinghouse (WWC) standards (What Works Clearinghouse, 2008).

Attrition can also occur below the level of randomization, for example, at the student level. We are mostly concerned with attrition at the student level if there is substantial attrition at that level and if we have reason to believe that this attrition happens for different reasons between the two conditions. This is the case here for the student attitude questionnaire, therefore, we examine whether student characteristics are balanced between conditions in the analytical sample for this outcome.

Number of Units in the Sample and Attrition

Table 9 shows changes in the samples from the point at which the classes were randomized to the point at which the posttests (California Standard Test, End of Course Assessment and Student Attitude Questionnaire) were received. It is important to note that data collection processes were different and the amount of attrition was different for each measure.

TABLE 9. NUMBERS OF UNITS IN THE EXPERIMENTAL GROUPS AND ATTRITION OVER TIME

	Control				*HMH Fuse*			
Event	No. of schools	No. of teachers	No. of sections	No. of students	No. of schools	No. of teachers	No. of sections	No. of students
Randomization	6	11	23	n/a	6	11	11	n/a
(Loss prior to rosters)	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Fall rosters received	6	11	23	664	6	11	11	334
California Standards Test (CST) Analytical sample								

TABLE 9. NUMBERS OF UNITS IN THE EXPERIMENTAL GROUPS AND ATTRITION OVER TIME

	Control				*HMH Fuse*			
Event	No. of schools	No. of teachers	No. of sections	No. of students	No. of schools	No. of teachers	No. of sections	No. of students
(Loss due to lack of posttest)	(0)	(0)	(0)	(39)	(0)	(0)	(0)	(16)
Final count of units with CST posttest	**6**	**11**	**23**	**625**	**6**	**11**	**11**	**318**
End of Course Assessment Analytical sample								
(Loss due to lack of posttest)	(1)	(2)	(7)	(266)	(1)	(3)	(3)	(113)
Final count of units with End of Course Assessment	**5**	**9**	**16**	**398**	**5**	**8**	**8**	**221**
Student Attitude Questionnaire Analytical sample								
(Loss due to lack of consent)	(0)	(0)	(0)	(127)	(0)	(0)	(0)	(11)
With consent	6	11	23	537	6	11	11	323
(Loss due to lack of posttest)	(0)	(0)	(0)	(31)	(0)	(0)	(0)	(24)
Final count of units with Student Attitude Questionnaire posttest	**6**	**11**	**23**	**506**	**6**	**11**	**11**	**299**

California Standards Test (CST)

We started with 34 randomized sections (the randomized unit) and no sections were lost to attrition. However, we exclude the records for students without posttests from the analytical sample.

We started with 998 students with records for the CST. We lost 55 out of 998 students due lack of a posttest, resulting in an analytical sample of 943 students. We do not have a posttest for 16 out of 334 *HMH Fuse* students (4.8%) and 39 out of 664 control students (5.9%). There is a 1.1% rate of differential attrition at the student level. This is not a statistically significant difference between conditions in the proportion of students lost ($p = .50$).

End of Course Assessment

We exclude the records for students without posttests from the analytical sample for that outcome. For ten of the 34 randomized sections, teachers did not submit the End of Course Assessment. Two teachers failed to administer the assessment to any of their students while one additional teacher failed to submit posttests for the section assigned to *HMH Fuse*. This attrition (29.4%) includes three out of 11 *HMH Fuse* sections (27.3%) and seven out of 23 control sections (30.4%). The result is a 3.1% rate of differential attrition at the section level, which is not a statistically significant difference.

These rates of attrition and differential attrition at the section level are small enough that they result in a low level of potential bias by WWC standards (What Works Clearinghouse, 2008).

In terms of student records, the missing posttests result in an overall loss of 379 out of 998 student records, resulting in an analytical sample of 619 student records. This includes the data for 113 out of 334 *HMH Fuse* students (33.8%) and 266 out of 664 control students (40.1%). The 6.3% rate of differential attrition at the student level is not statistically significant (p = .52).

We noted that most students are without posttests due to two teachers not providing data for their students in both conditions. Therefore, bias due to non-participation is driven by teacher-level factors. Because the design blocks on teacher, we effectively lose two blocks. Because randomization is conducted within blocks, the statistical equivalence resulting from randomization within the remaining blocks is maintained. (The exception is the one teacher for whom we have data in the control sections but not the *HMH Fuse* section).

Student Attitude Questionnaire

Parental consent was required in order to use individual student data for the Student Attitude Questionnaire. No sections were lost to attrition.

We started with 998 students in 34 randomized sections. There was a loss of data for 193 out of 998 students, resulting in an analytical sample of data from 805 students. The sample includes 158 out of 664 control students (23.80%) and 35 out of 334 *HMH Fuse* students (10.48%)]. This 13.32% difference in the rate of attrition at the student level (p = .01) is statistically significant.

Much of the attrition was due to lack of parental consent obtained for students in the control condition. When considering attrition due to lack of consents, we lost data for 127 out of 664 students among the control group and only 11 out of 334 students assigned to *HMH Fuse*. The high return rate among *HMH Fuse* students resulted from district-hosted meetings for parents of students assigned to the *HMH Fuse* group. *HMH Fuse* parents signed documents accepting responsibility for the iPads in addition to consent forms for the research study. Parents of control students attended no such meetings; rather, consent forms were sent home from school, through the students, and collected by the classroom teachers. Therefore, fewer consent forms were received for students in the control group than for the *HMH Fuse* group.

Although no sections were lost to attrition, we follow Deke (2007) in the assessment that when classes are randomized but students attrite, then implications of attrition for bias depend on the exact nature of attrition. In the case of this outcome measure, the inclusion of students follows different mechanisms in the two conditions, since parents' opportunity to give consent to students' responding to the questionnaire in the *HMH Fuse* group was increased through the availability of the district-hosted meetings. As a result, different kinds of students may be excluded from analysis in the two conditions. To assess the potential for bias due to differential attrition of students, we examine whether there is a difference between conditions in the background characteristics of students in the analytical sample. Measures of these characteristics will be included in the analytic model for the impact analysis whether or not we establish equivalence between conditions.

Characteristics of the Initial Sample

For the Student Attitudes Questionnaire, we observe a 13.32% difference in the rate of attrition at the student level (p = .01). This is a statistically significant difference between conditions in the proportion of students initially randomized for whom we do not have posttests, which in our case coincides with the analytical sample

Due to the number of students without a posttest, and the circumstances leading to different consent rates in the two conditions, we examine the equivalence on background characteristics for the analytical sample. The results show balance on all factors. (We do not include a test of balance for the pretest because we used the residualized version of the pretest, and it is balanced by construction.)

TABLE 10. CHARACTERISTICS OF STUDY SAMPLE (ANALYTICAL)

	Control	*HMH Fuse*	Total	Less than 5% chance of seeing this much imbalance
Student characteristics				
Male	249 (49.21%)	142 (47.49%)	391 (48.57%)	No
Grade 8	427 (84.39%)	253 (84.62%)	680 (84.47%)	No
Disability	10 (1.98%)	7 (2.34%)	17 (2.11%)	No
English speaker	464 (91.70%)	254 (84.95%)	718 (89.19%)	No
Asian	153 (30.85%)	81 (27.27%)	234 (29.51%)	
White	120 (24.19%)	39 (13.13%)	159 (20.05%)	
Black	41 (8.27%)	29 (9.76%)	70 (8.83%)	No
Mixed	13 (2.62%)	5 (1.68%)	18 (2.27%)	
Indian	1 (0.20%)	0	1 (0.13%)	
Hispanic	168 (33.87%)	143 (48.15%)	311 (39.22%)	

REPORTING ON THE IMPACT OF *HMH FUSE*

Setting Up the Statistical Equation[4]

We put our data for students, teachers and classes into a system of statistical equations that allow us to obtain estimates of the effects of interest. The primary relationship of interest is the causal effect of the program on achievement as measured by the CST. We use SAS PROC MIXED and PROC

[4] The term "statistical equation" refers to a probabilistic model where the individual outcomes are on the left-hand side of the equation and terms for systematic and random effects are on the right-hand side of the equation. The goal of estimation is to obtain estimates for the effects on the right-hand side. Each estimate has a level of uncertainty that is expressed in terms of a standard error or *p* value. The estimate of main interest is for the program effect. In this experiment, we model program as a fixed effect. With randomized control trials, the modeling equation for which we are estimating effects takes on a relatively simple form. Each observed outcome is expressed as a linear combination of a variable indicating assignment status (program or control), one or more covariates that are used to increase the precision of the intervention effect, and usually a series of fixed or random effects, which are increments in the outcome that are specific to units. As a result of randomization, on average each covariate is distributed in the same way for both the program and control groups. For moderator analyses, we expand these basic models by including a term that multiplies the variable that indicates assignment status by the moderator variable. The coefficient for this interaction term is the moderator effect of interest.

GLIMMIX (SAS Institute Inc., 2006) as the primary software tools for these computations. The output of the analysis process consists of estimates of effects, as well as *p* values that tell us how much confidence we should have that the estimate are different from zero.

Program Impact

The primary question for the experiment was whether, following the intervention, students in *HMH Fuse* classrooms had higher scores on the algebra CST, End of Course Assessment, and Student Attitude Questionnaire than students in control classrooms. To answer this question, we analyzed outcomes for the randomized groups. The randomization resulted in two groups that at the outset are statistically equivalent. One receives *HMH Fuse* and the other one does not. As a result, the average difference between the randomized groups on the posttest is an accurate measure of the program effect plus random error. We can increase the accuracy of our effect estimates by accounting for the effects of covariates in the analysis. Therefore, our statistical equations included the following covariates modeled at the student level: the pretest and English learner status. We also had to account for the fact that students are clustered by sections. We expect outcomes for students who are in the same section to be dependent as a result of shared experiences. We had to add this dependency to our equation in order to prevent artificially high confidence levels about the results. To do this, we modeled a section-level random effect as we describe further in the upcoming section, titled *Fixed and Random Effects*.[5]

Handling Missing Data

To control for potential bias in the effect estimate arising from the covariates having missing values, we used a dummy variable method. With this approach, for each of the covariates that is included in the model, a dummy variable was created. This variable was assigned a value of one if the value of the variable was missing for any student, and zero otherwise. The missing values from the original variable were replaced with zero. The dummy method yields effect estimates with less bias than the tolerance threshold set by the What Works Clearinghouse with levels of attrition such as those observed here (this finding is obtained through a simulation study described in Puma, Olsen, Bell, & Price, 2009). Specifically, the method fares no worse and, in some cases, performs better when compared to other standard approaches, including case deletion and non-stochastic and several stochastic regression imputation methods.

When student achievement outcomes (posttests) were missing, we used listwise deletion and simply dropped the observation from the analysis. This approach to handling missing data is one of several recommended by Puma et al. (2009). In their simulation work, they found that this method produced impact estimates with bias that was smaller than 0.05 standard deviations of the outcome measure (they considered bias in both the estimated impact and its associated standard error).

---

[5] Our analytic models contain several covariates including measures of background characteristics, dummy variables to indicate missing values for the covariates, and dummy variables to indicate teachers. The reason for including these variables in the model is to increase the precision of the impact estimate or as a strategy for addressing missing values. In order to keep focus on the main results, we do not present estimates of the effects that correspond to these variables in the main body of the report (see the Appendix for the results for the full model.) Use of the dummy variable methods for addressing missing values for the covariates involves setting missing values to a constant (zero) which does not allow for a straight-forward interpretation of the effects of the covariates.

Covariates and Moderators at the Student and Teacher Level

In addition to the variable indicating whether a section is assigned to *HMH Fuse* or the control condition, we include in the statistical equation covariates that we expect to make a difference in the outcomes. For example, as was described previously, we add the pretest score into our statistical equations in order to increase precision. Some of the covariates are also used to model moderator effects. We consider whether there is a difference in the effect of the intervention for different levels of the covariate. For example, we consider whether the program is more effective for higher-performing students than for lower-performing students. We estimate this *difference* (between subgroups) *in the difference* (between the program and control groups) in posttest performance by including an interaction term in the statistical equation. This term multiplies together the variable that indicates whether the student is in the intervention group and the covariate. The coefficient for this term is a measure of the moderating effect of the covariate on the effect of the program. We call covariates that are included in such analyses potential moderators because they may moderate—either increase or decrease—the effect of the program on student outcomes.

Teacher Level Outcomes and Potential Mediators

We are also interested in measurable characteristics of teacher behavior, or beliefs, and student activity that can be measured during the experiment. Unlike the moderators, these are not pre-existing characteristics such as pretest score or English learner status. These factors are called potential mediators: "potential" because they are hypothesized, and "mediators" because they are outcomes that fall between the assignment mechanism and the final outcome (usually student achievement).

The objective of a mediation analysis is to examine whether an impact of the program on student achievement happens through an initial impact on an intermediate variable. If an impact is demonstrated on the intermediate variable, and we can also establish an association between the intermediate variable and student achievement independent of the effect of the program, then the intermediate variable may be a mediator of the impact on achievement.[6] Because we are not randomly assigning cases to levels of the mediator variable, we leave open the possibility that the mediating variables we are examining are proxies for hidden variables that are the true mediators of the process. That is, we cannot be sure of the causal status of the mediator.

We assess mediation whether or not there is an overall impact on student achievement because the mediating path that we are investigating may be one of several, and their effects may cancel when combined, leading to zero overall effect. Therefore, lack of an overall impact does not rule out mediation along the path of interest. On the other hand, if there is no impact on the posited mediator of interest, then we do not consider that mediating path further.

---

[6] Technically, the estimate of a given mediated effect is the product of the effect of treatment on the mediator, times the effect of the mediator on the final response variable, normally student achievement, holding constant the treatment effect (Krull & MacKinnon, 2001) In a mediation model with a single mediator, this is equivalent to (or for multilevel models, approximate to) the difference between (1) the effect of treatment on the final outcome before adjusting for the effect of the mediator, and (2) the effect of treatment on the final outcome after adjusting for the effect of the mediator (Krull & MacKinnon, 2001).

Fixed and Random Effects

The covariates in our equations measure either (1) fixed characteristics that take on a finite set of values (e.g., there are only two levels of gender) or (2) a set of characteristics that is assumed to have a distribution over a population and where we treat the values that we measure as though they were a random sample from that larger population. The former are called fixed effects; the latter, random effects. Random effects add uncertainty to our estimates because they account for sampling variation, or the changes we would observe in the outcomes if we re-sampled units from the same population. Fixed effects produce less uncertainty but also limit the extent to which we can generalize our results.

We usually treat the effects of units that were randomized as random, so that in the statistical equations, our estimates reflect the degree of uncertainty that comes if we were to draw a different sample of such units from the same population.[7] This allows us to argue for the generalizability of our findings from a sampling perspective. Treating the effects of units that were randomized as fixed forces us to use other arguments if our goal is to generalize.

Using random or fixed effects for participating units serves a second function: it allows us to more accurately represent the dependencies among cases that are clustered together, especially for the clusters randomly assigned to conditions. All the cases that belong to a cluster share an increment in the outcome—either positive or negative—that expresses the dependencies among them. An appropriate measure of uncertainty in our estimate of the program's effectiveness takes into consideration the relative levels of variation *within* and *between* the clusters randomized. All of our statistical equations include a student-level error term and a randomization-level error term. The variation in these terms reflect the differences we see (1) among students within clusters, and (2) across randomized clusters, that are not accounted for by all the other effects in our statistical equation.

The choice of terms for each statistical equation is not rigid but depends on the context and the importance of the factors for the question being addressed. The tables reporting the estimates resulting from the computation will provide an explanation of these choices in table notes where necessary for technical review.

Exploratory Investigations

Finally, to better understand unexpected results, in some cases we use other demographics, teacher characteristics, and supplementary observational data in exploratory investigations to generate additional hypotheses about which factors interact with the program. These results are considered exploratory, because they often follow inspection of the results of analyses that are planned at the design stage of the experiment. Their primary goal is to inform future studies.

Reporting the Results

[7] Although we seldom randomly sample cases from a broader population, and in some situations we use the entire population of cases that is available, we believe that it is still correct to estimate sampling variation (i.e., model random effects). It is entirely conceivable that some part or the whole set of participants at a level end up being replaced by another group (for whatever reason) and it's fair to ask how much change in outcomes we can expect from this substitution.

When we run the computations on the data, we produce several results: among them are effect sizes, the estimates for fixed effects, and *p* values.

Effect sizes

We translate the difference between program and control groups into a standardized effect size by dividing the average group difference by a measure of the variability in the outcome. This measure of variability is also called the standard deviation and can be thought of as the average distance of all the individual scores from the average score (more precisely, it is the square root of the average of squared distances). Dividing the difference by the standard deviation gives us a measure of the impact in units of standard deviation, rather than units of the scale used by the particular test. This standardized effect size allows us to compare the results we find with results from other studies that use different measurement scales. In studies involving student achievement, effect sizes as small as 0.1 (one tenth of a standard deviation) are sometimes found to be important educationally. We also report the effect size where we divide the average difference, adjusted for the effects of pretest score and other covariates, by the standard deviation. This is called the 'adjusted effect size'. This adjustment will often provide a more precise estimate of the impact.

Estimates

We provide estimates to approximate the actual effect size. Any experiment is limited to the small sample of students, teachers, and schools that represent a larger population in a real world (or hypothetical) setting. Essentially we are estimating the population value. When we report an estimate in a table, the value refers to the change in outcome for a one-unit increase in the associated variable. For example, since we code participation in the control group as 0, and participation in the program group as 1, the estimate is essentially the average difference in the outcome that we expect in going from the control to the program group while holding other variables constant.

p values

The *p* value is very important, because it gives us a gauge of how confident we can be that the result we are seeing is not due simply to chance. Specifically, it tells us what the probability is that we would obtain a result with a magnitude as large as—or larger than—the magnitude of the one observed when in fact there is no effect. Roughly speaking, it tells us the risk of concluding that the intervention has had an effect when in fact it hasn't. This mistake is also known as a false-positive conclusion. Thus a *p* value of .1 gives us a 10% probability of drawing a false-positive conclusion if in fact there is no impact of the program. This is not to be confused with a common misconception that *p* values tell us the probability of our result being true.

We can also think of the *p* value as the level of confidence, or the level of belief we have that the outcome we observe is not simply due to chance. While ultimately depending on the risk tolerance of the user of the research, we suggest the following guidelines for interpreting *p* values.

1. We have a high level of confidence when $p \leq .05$. (This is the level of confidence conventionally referred to as statistical significance.)
2. We have some confidence when $.05 < p \leq .15$.
3. We have limited confidence when $.15 < p \leq .20$.
4. We have no confidence when $p > .20$.

In reporting results with *p* values higher than conventional statistical significance, our goal is to inform the local decision makers with useful information and provide other researchers with data points that can be synthesized into more general evidence.

Results

Implementation Results

In this section we provide a description of math instruction among the control and *HMH Fuse* groups to inform the interpretation of student outcomes. Data for this section were obtained largely through the nine online teacher surveys. Additional data were obtained through the student surveys and program usage logs. Implementation results are reported in the following categories.

- Conditions for Implementation
- Implementation Fidelity and Extent of Program Implementation
- Comparison of Classroom Implementation between *HMH Fuse* and Control Groups
- Teacher Satisfaction with the *HMH Fuse* Program

Conditions for Implementation

Here we provide a description of the conditions under which implementation in *HMH Fuse* and control classrooms took place. Specifically, we present data on training and materials, teachers' initial impressions and comfort level with *HMH Fuse,* and impediments to implementation.

Training and Materials

All participating teacher received training for *HMH Fuse* by September 10, 2010. Since use of the print version of the program was considered *business as usual*, no additional training was offered for use of the print version of the text.

In response to the September survey, teachers reported that iPads had been distributed to all students in their classes, with the exception of one new student. In the same survey, nine of the eleven teachers responded that they had started using the iPad application for algebra instruction in their *HMH Fuse* class; one teacher had not yet begun using the application; and one teacher indicated that students were using their iPads at home for support with homework and review, but that they had not used them in class for direct instruction. However, all 11 teachers responded that students were taking home their iPads.

Initial Teacher Impressions

In October 2010, after the initial training, teachers were asked about their initial impressions of *HMH Fuse*. Teachers had generally high expectations for the device, with only one of the 11 teachers (9%) doubtful of the new program, as depicted in Figure 1.

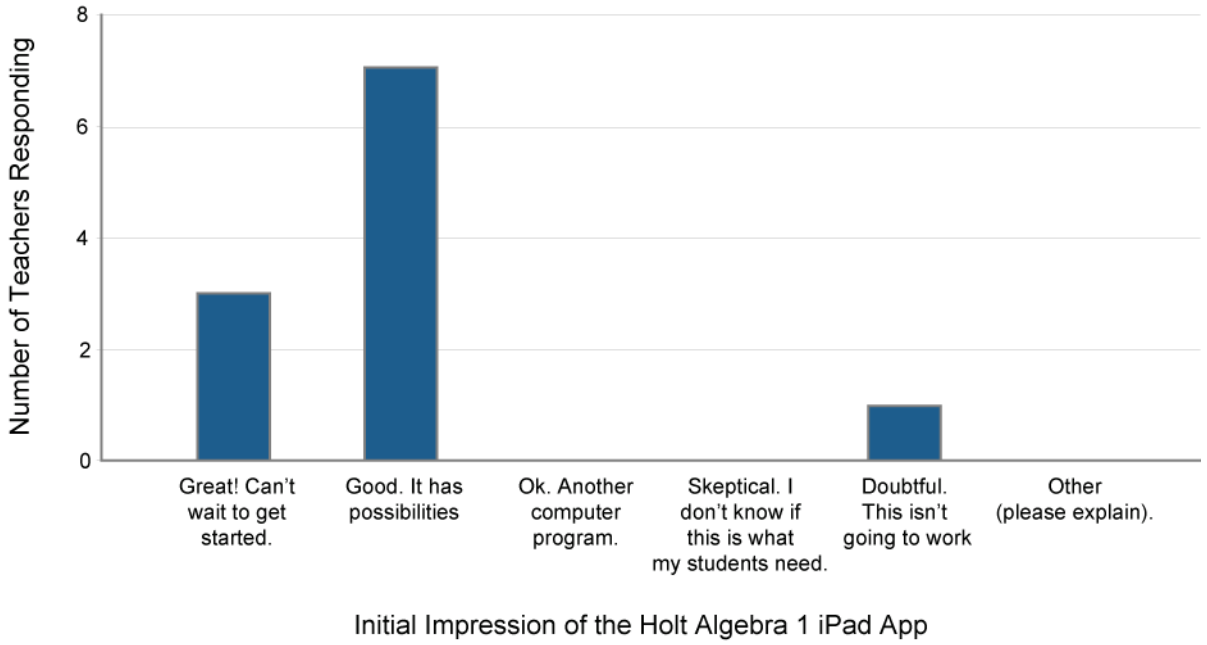


FIGURE 1. TEACHERS' INITIAL IMPRESSION OF *HMH FUSE*

In the same survey, teachers were asked about their level of comfort with using *HMH Fuse* as an instructional tool with their students. The majority of the teachers (82%) responded that they were somewhat or very comfortable with using *HMH Fuse,* while one was neutral and one was not very comfortable.

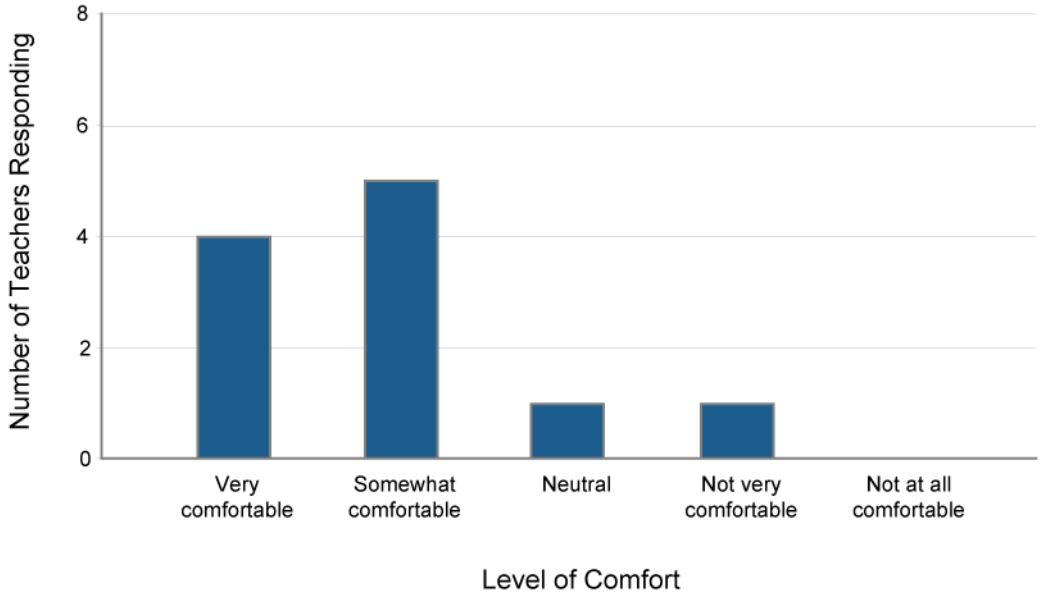


FIGURE 2. COMFORT LEVEL USING *HMH FUSE* AS AN INSTRUCTIONAL TOOL WITH STUDENTS

Impediments to Implementation

Researchers also surveyed teachers regarding any difficulties with technology. In response to the October survey, nine teachers (82%) reported that they had had technical issues with their iPads since deployment. They reported a variety of specific issues, with no overall trends. Teacher-reported problems included the following.

- Problems with internet connectivity (2 teachers)
- iPads crashing (2 teachers)
- Clicker app doesn't work (2 teachers)

Individual teachers also reported instances of a screen jiggling, a student changing and forgetting the password, an iPad that wouldn't turn on, an iPad that wouldn't charge, disabled iPads, and not being able to locate or open the application.

Support

Of the nine teachers who reported technical issues, eight (89%) reported contacting someone for support with these technical problems. These contacts ranged from individuals at Houghton Mifflin Harcourt and EduSoft (the testing company) to technical support from Apple and district contacts. Of those eight teachers, two (25%) reported that the issue had been resolved to their satisfaction; three were still in the process of reconciling their issue; two reported that their issues had been addressed, but not to their satisfaction, as the process took too long; and one teacher reported never having received a response to their inquiry.

Summary of the Conditions for Implementation

All participating teachers were trained in *HMH Fuse* at the beginning of the school year. Implementation of *HMH Fuse* began at the start of the school year in all four districts and all teachers confirmed that they had the needed materials by the time of the first survey. The majority of the teachers seemed to be generally optimistic and comfortable with using *HMH Fuse*. However, teachers did report issues with technology at the start of the year, although those issues varied widely. By the time of the second survey, all but one of the teachers had their issue resolved or were in the process of having their issue resolved, although two were not satisfied with the length of time required for resolution.

Implementation Fidelity and Extent of Program Implementation

In this section, we describe the extent to which the *HMH Fuse was* implemented.

Implementation Fidelity

Researchers informed teachers during the one day training that students in classes assigned to use the iPad were to use *HMH Fuse* as their sole algebra curriculum, while those in the control classes were to use only the print edition and online videos as their core curriculum. This was the only instruction given regarding implementation fidelity.

In order to gauge any instances of cross-over between the two assignment groups, researchers asked on three separate surveys whether students assigned to use the print version of the text had any interaction with *HMH Fuse*, and whether students in the *HMH Fuse* group had used the print version of the text.

Teachers reported in all three of the surveys that their program students had used the control materials - six (60%) of the teachers in October decreasing to three in May, as displayed in Table 11. In October, 5 of 6 teachers (83%) noted that the print text was used as a substitute when a student forgot his or her iPad or was having technical problems. One teacher used an older version of the textbook for homework problems, with students in both the *HMH Fuse* and control groups, because they believe the new text does not have good enough homework problems. In May, one teacher reported using the text as backup when the iPads weren't working, one teacher used the text for homework, and one teacher used the text to get answer keys before the Teacher's Edition was available for the iPad.

TABLE 11. HAVE YOU OR YOUR STUDENTS WHO ARE IN THE ALGEBRA 1 IPAD APP CLASS USED THE PRINT TEXTBOOK IN ANY CAPACITY SINCE RECEIVING THE IPAD APP?

	Yes	No
October (n = 10)	6 (60%)	4 (40%)
May (n = 11)	3 (27%)	8 (73%)

Table 12 reflects teacher responses regarding control students using the iPad app. In October two (18%) of the teachers indicated that there had been some contamination. This number decreased to one teacher in January, and at the end of the school year, teachers reported that there was no contamination of the control group. Teachers explained that some students share videos before school or have friends who are in the iPad classes. In the student repeater surveys, 3.7% of control students reported working on their algebra program on an iPad.

While many of the features of *HMH Fuse*, such as the videos, were highlighted in the training, teachers received no guidelines as to how often or for how much time they should use various components of the curriculum. No recommendations were made regarding how many videos they should watch or how many quizzes they should implement. Figure 3 displays teachers' responses regarding a variety of possible instructional activities with *HMH Fuse*. At the end of the school year, seven of eleven teachers (64%) had used the iPad screen with a projector, three of eleven (27%) had used the clicker app, and 100% had students work independently on the devices. Among the responses from teachers who had indicated 'Other Activities,' were the note-taking feature, videos, and the graphing calculator.

TABLE 12. HAVE THE STUDENTS WHO WERE ASSIGNED TO USE THE PRINT VERSION OF THE TEXT HAD ANY INTERACTION WITH THE HOLT ALGEBRA 1 IPAD APP IN YOUR CLASSROOM?

	Yes	No
October (n = 11)	2 (18%)	9 (82%)
May (n = 11)	0 (0%)	11 (100%)

Summary of Implementation Fidelity and Extent of Program Implementation

While no teachers indicated that they ever used the iPad or *HMH Fuse* in their control classrooms, some did note that students might, on occasion, share their iPad or the *Fuse* program with students in control classes, which student survey responses confirmed. In *HMH Fuse* classes, teachers occasionally used the control text, namely when technical problems prevented a particular student from being able to use *HMH Fuse*. It appears that any contamination was limited. Teachers used the iPads to varying degrees and in a variety of manners, however, they all allowed students to work independently on the iPad.

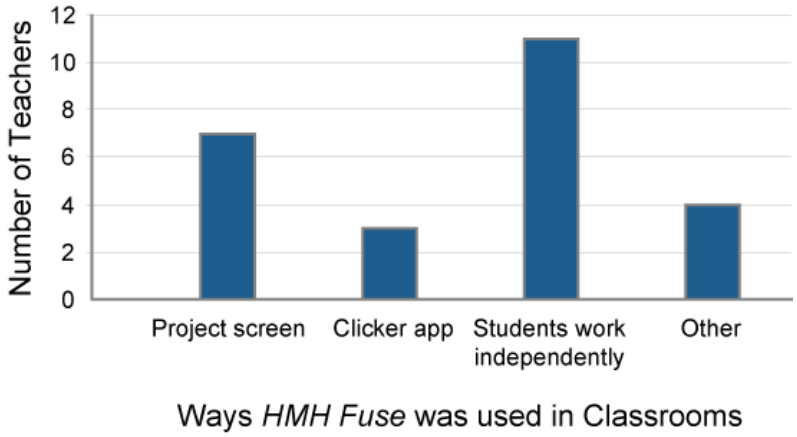


FIGURE 3. WAYS *HMH FUSE* WAS USED IN CLASSROOMS

Comparison of Classroom Implementation between *HMH Fuse* and Control Classes

Here we describe survey data comparing algebra instruction and learning in *HMH Fuse* and control classrooms. In each of Surveys 2 through 9 we posed questions regarding the amount of time spent teaching, learning, and watching videos during specified weeks. Table 13 shows teacher responses regarding the number of minutes they spent using the textbook and *HMH Fuse* for active classroom instruction (projecting or demonstrating in some other manner). On average,[8] teachers reported spending 70 minutes per week actively instructing with the print edition of the text compared to 22.5 minutes with *HMH Fuse*. With a *p* value smaller than .05, we have a high level of confidence that this result is not due to chance.

TABLE 13. WEEKLY MINUTES OF ACTIVE INSTRUCTION

	Median Weekly Minutes
Control (n =11)	70
***HMH Fuse* (n = 11)**	22.5
***p* value**	.03

Note. We used the Wilcoxon signed rank sum test. A paired t-test corroborated that the difference is statistically significant (p = .01).

We also posed repeated questions regarding the number of minutes students spent using the assigned curriculum during class time. Similar to the responses regarding teacher instructional time, teachers reported that during class time control students used the print version of the program materials (following along with teacher instruction, working independently, working in groups, etc.)

[8] We use the Wilcoxon Signed Rank Sum Test, because with small samples we have limited information about the distributions of the underlying distributions. Since this is a test of median differences, we display the median outcomes.

about 83 minutes per week as the median value,[9] as compared to program students who used *HMH Fuse* for only 42 minutes per week. However, with a *p* value of .41, we have no confidence that this difference is not due to chance.

TABLE 14. WEEKLY MINUTES OF STUDENT USE

	Median Weekly Minutes
Control (n =11)	82.5
***HMH Fuse* (n = 11)**	41.88
***p* value**	.41

Note. We used the Wilcoxon signed rank sum test. A paired t-test corroborated that the difference is not statistically significant (*p* =.27).

Over the course of the school year, twenty percent of students in control classes and 6% of students in *HMH Fuse* classes reported spending zero minutes working on algebra outside of class (Table 15, below.) With $p < .01$, we have a high level of confidence that this result is not due to chance. See Appendix B for details of analysis.

TABLE 15. MINUTES SPENT ON ALGEBRA OUTSIDE OF CLASS

	Zero Minutes	More than Zero Minutes
Control	113 (19.96%)	453 (80.04%)
HMH Fuse	17 (5.54%)	290 (94.46%)
Total	130 (14.89%)	743 (85.11%)

As displayed in Table 16, teachers reported that control classes did not watch any videos, while *HMH Fuse* classes only watched a video every fourth week, on average. However, due to the large *p* value, we have no confidence that the difference is not due to chance.

TABLE 16. NUMBER OF VIDEOS WATCHED IN CLASS

	Median Number of Videos
Control (n =11)	0
***HMH Fuse* (n = 11)**	.25
***p* value**	.31

Note. Wilcoxon signed rank sum test. Result corroborated by paired t-test.

[9] We use the Wilcoxon Signed Rank Sum Test, because with small samples we have limited information about the distributions of the underlying distributions. Since this is a test of median differences, we display the median outcomes.

Over the course of the school year, fewer students in control classes reported watching 8 or more videos each unit than students in *HMH Fuse* classes (Table 17, below). Videos may have been watched either during or outside of class time. With $p < .01$, we have a high level of confidence that this result is not due to chance. See Appendix B for details of analysis.

TABLE 17. FREQUENCIES OF MEDIAN RESPONSES TO NUMBER OF ALGEBRA VIDEOS WATCHED (COLLAPSED INTO TWO CATEGORIES)

	0-7	8 or more
Control	538 (95.73%)	24 (4.27%)
HMH Fuse	252 (82.89%)	52 (17.11%)
Total	790 (91.22%)	76 (8.78%)

In response to the student survey, students provided their primary and secondary reasons for watching algebra videos for each chapter. Table 18 reports the primary reason students reported for watching videos, on average across the surveys. The majority of students across both groups chose homework help as their main reason for watching videos.

TABLE 18. #1 REASON FOR WATCHING ALGEBRA VIDEOS

	Curiosity	Teacher Assigned/Suggested	To Help with Homework	Review for Test/Quiz	iPad Automatically Displayed
Control	16.4%	22.9%	40.4%	18.7%	1.7%
HMH Fuse	13.0%	16.5%	50.7%	18.0%	1.8%

Note. Average sample size of 537.

Table 19 reports the secondary reason students reported for watching algebra videos, on average. The majority of students chose test/quiz review as their secondary reason for watching videos.

TABLE 19. #2 REASON FOR WATCHING ALGEBRA VIDEOS

	Curiosity	Teacher Assigned/Suggested	To Help with Homework	Review for Test/Quiz	iPad Automatically Displayed
Control	17.9%	16.2%	26.0%	38.1%	1.9%
HMH Fuse	15.9%	11.2%	28.9%	41.5%	2.6%

Note. Due to the rounding of decimals, percentages may not add up to 100%. Average sample size of 534.

Table 20 displays usage data for seven program features in addition to an overall report of usage. The features that students used the most include quizzes, answer checking, videos, and inline example references, while students used homework help walkthrough the least.

TABLE 20. LOG USAGE DATA

Feature	Mean Number of Times Used Per Student	Standard Deviation	Standard Error	Min	Max
Total Usage (Overall counts)	3879.03	3432.00	194.92	90	30742
Math in Motion	13.76	15.13	0.86	0	94
Note-Taking	15.60	36.06	2.05	0	546
Quizzes	49.01	60.65	3.44	0	360
Inline Example References	28.57	56.14	3.19	0	477
Answer Checking (Homework)	41.86	44.97	2.55	0	249
Homework Help Walkthrough	8.38	12.13	0.69	0	101
Videos	33.51	39.35	2.24	0	273

Note. Based on device log data for 311 students.

On the March survey, researchers asked teachers about the content that they covered. Nine of the ten teachers (90%) reported that they taught the same lessons to students in both conditions. Sixty percent (6 out of 10) of the teachers reported that they had skipped chapters, as follows.[10]

- Three teachers had skipped Chapter 1
- Teachers who skipped Chapter One noted that it was a school site decision and that the chapter was review/ too basic.
- Three teachers had skipped Chapter 4.
- One said it was not tested on the CST. Two noted that their school/ math department had decided to teach only parts of Chapter 4

TABLE 21. HOURS PLANNING FOR ALGEBRA INSTRUCTION

	Weekly Planning Hours
Control (n = 11)	2.5
***HMH Fuse* (n = 11)**	1.5
***p* value**	.25

Note. We applied the Wilcoxon signed rank sum test. A paired t-test gave p =.10.

In response to survey questions about administration of the chapter tests, three of ten teachers (30%) reported that for at least one of the chapters that they had taught, they had failed to administer the

[10] Teachers were only asked about the chapters for which they had already received materials, which by March included Chapters1 through 9, but not 10 or 11.

chapter tests. We also asked teachers if they still administered the chapter tests when they had skipped a chapter. Not one of the ten responding teachers reported that they had administered a chapter test for a chapter that they had not taught.

To better understand any differences in the amount of time required to prepare for instruction in *HMH Fuse* and control classes, in Survey 5 we asked teachers to report the number of hours they spent each week planning for algebra instruction. As depicted in Table 21, teachers reported spending 2.5 hours each week, on average planning for their control classes as compared to 1.5 hours planning for *HMH fuse* classes. As noted in the table, two statistical tests gave different results but not inconsistent with a conclusion that this difference was a matter of chance.

Summary of Classroom Implementation of *HMH Fuse* and Control Groups

Researchers found that participating teachers spent significantly more time instructing with the assigned algebra program in control classes than in *HMH Fuse* classes. However, we did not find a difference between groups in the amount of time students spent with their algebra programs in class nor the numbers of videos watched in class. Very few videos were watched in class across both groups. We also did not find any difference in the amount of time planning for class instruction between *HMH Fuse* and control classes. We did confirm that teachers skip some chapters but do not administer the chapter test for the chapters skipped.

Teacher Satisfaction with *HMH Fuse*

Surveys deployed at the end of the study asked teachers to rate their overall satisfaction with *HMH Fuse* and their control curriculum, and whether they would recommend *HMH Fu*se to other teachers of algebra. In addition, researchers questioned teachers, in an open-ended format, about what they found particularly useful and difficult about each of the programs. Finally, researchers revisited the level of teacher comfort with the *HMH Fuse* technology, comparing responses to those obtained at the beginning of the school year, as reported earlier in this report in Figure 2.

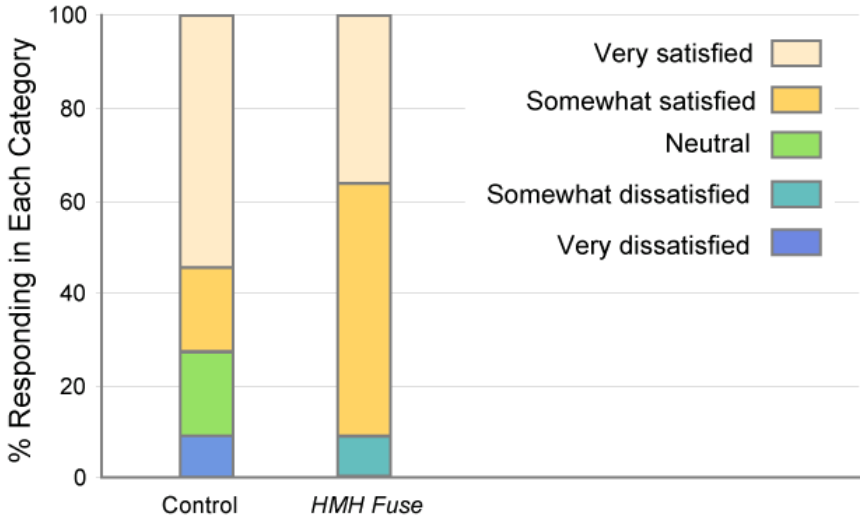


FIGURE 4. TEACHER SATISFACTION WITH *HMH FUSE*

As depicted in Figure 4, ten of 11 teachers (91%) reported that they were very satisfied or somewhat satisfied with *HMH Fuse*, while eight of the 11 teachers (73%) reported the same level of satisfaction with the control program. The one teacher who reported being somewhat dissatisfied with *HMH*

Fuse reported being very dissatisfied with the control curriculum. We have no confidence that the difference in satisfaction between the *HMH Fuse* and control curriculum is not due to chance.[11]

Table 22 shows that 10 of 11 (91%) teachers would recommend the print edition of the Holt Algebra 1 text and nine (82%) would recommend *HMH Fuse* to other algebra teachers. No teachers responded that they would not recommend either program. We have no confidence that the difference between programs with regard to teacher recommendation is not due to chance.[12]

TABLE 22. WOULD YOU RECOMMEND THIS PROGRAM TO OTHER ALGEBRA TEACHERS?

n = 11	Yes	No	I don't know
Control	10 (90.9%)	0 (0.0%)	1 (9.1%)
HMH Fuse	9 (81.8%)	0 (0.0%)	2 (18.2%)

In response to a teacher survey question regarding whether, given the opportunity, they would choose to use *HMH Fuse* to teach algebra instead of the print version of the Holt algebra text, 9 of 11(82%) said they would, with the remaining two teachers (18%) choosing I don't know.

TABLE 23. IF YOU HAD THE OPTION, WOULD YOU CHOOSE TO TEACH ALGEBRA USING THE ALGEBRA 1 IPAD APP INSTEAD OF THE PRINT VERSION OF THE HOLT ALGEBRA TEXT?

	Yes	No	I don't know
n = 11	9 (81.2%)	0 (0.0%)	2 (18.2%)

Note. Due to the rounding of decimals, not all percentages add up to 100%.

Table 24 lists some of the features that teachers describe as most useful and difficult with *HMH fuse* and with the print version of the text. We received fewer responses to the question about the print version, and there were fewer common themes among responses.

[11] A non-parametric test (Wilcoxon signed rank sum test) of a difference between conditions in satisfaction yielded a *p* value of 1.

[12] A non-parametric test (Fisher's exact test) of a difference between conditions in the distribution of counts across categories yielded a *p* value of 1.

TABLE 24. FEATURES THAT TEACHERS FOUND MOST USEFUL AND MOST DIFFICULT

Most useful features	Most difficult features
HMH Fuse	
• Videos (8 teachers) • Graphing calculator and linear/quadratic explorer (3 teachers) • Practice problems, View-in-Motion, Tutorials (6 teachers) • Note-taking feature (2 teachers) • Clicker app (2 teachers) • Compact and portable (2 teachers) • Quizzes and tests (2 teachers)	• No zoom feature (2 teachers) • Clicker app not working (2 teachers) • Finding extra practice at end of book (2 teacher) • No answers for study guide, homework, or review guides (2 teachers) • Accessing teacher edition, getting answers to questions in real time, can't freeze screen to point at something, screen not easily visible if projecting
Control	
• Bank of questions at end of book (2 teachers) • Supplementary materials (2 teachers) • Easy to use, well thought out, engaging, appropriate for all levels of student learning. • Study guide sections, online videos, practice problems, lots of homework problems.	• Not aligned to California standards or district pacing (3 teachers) • Chapters/materials that weren't necessary, not enough problems and not enough range from easy to hard, heavy

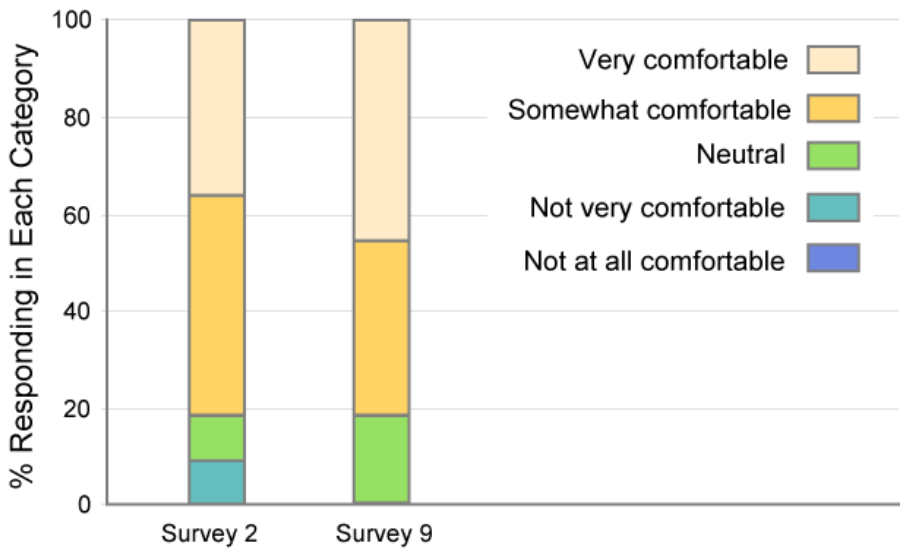


FIGURE 5. TEACHER COMFORT WITH *HMH FUSE*

Figure 5 compares teachers' responses regarding their level of comfort with *HMH Fuse* in the beginning and at the end of the school year. Out of the 11 respondents for each survey, nine teachers chose either very or somewhat comfortable in both Surveys 2 and 9, with one teacher moving from somewhat to very

comfortable. Survey 9 had no teachers reporting to be either not very or not at all comfortable. Due to the high *p* value, we have no confidence that any change over the school year is not due to chance variation.[13]

Summary of Teacher Satisfaction

Teachers indicate satisfaction with both programs (*HMH Fuse* and control) and would recommend both programs to other teachers of algebra. However, the majority of teachers would continue to use *HMH Fu*se over the control curriculum if they were given the opportunity.

STUDENT-LEVEL IMPACT RESULTS

Overview

The primary goal of our experiment was to understand the impact of *HMH Fuse* on student Algebra achievement. Here we examine the program's impact in three ways.

Program impact on students: We examine the average program effect for each outcome scale. First we address the impact on CST, the primary outcome measure, and End of Course achievement. Next we address whether being in an *HMH Fuse* or control class makes a difference on students' attitudes toward mathematics.

Moderation of the impact: For the two primary outcome scales, CST and End of Course, we examine whether the impact of the program varies depending on levels of potential moderating characteristics. Moderators are conditions or characteristics that are measured before the start of the program and that are associated with differences in the impact of the program. We always begin by examining whether the impact of the program differs depending on the students' pretest scores—do pretest scores moderate the impact?

Mediation of the impact: We examine whether the program has an impact on classroom practices or student outcomes that potentially mediate the impact on student achievement. If there is an impact on the intermediate variables, we examine whether it accounts for differences between the *HMH Fuse* and control groups in average student achievement.

Program Impact on Students

California Standards Test (CST)

In this section we address the impact of *HMH Fuse* on performance on the Algebra 1 CST.[14] Table 25 provides a summary of the samples used in the analysis and the results for the comparison of the

[13] A non-parametric test (Wilcoxon signed rank sum test) of a difference between conditions in change in level of comfort yielded a *p* value of 1.

[14] We include a series of covariates to improve precision (thus, an ANCOVA analysis) and model random section effects to reflect the cluster randomized design. The covariates included the pretest. We used the 6th grade scores as pretests for the 7th graders in the study and the 7th grade pretest scores for the 8th graders in the study. The CST is not a vertically scaled test; therefore, we rescaled the pretests to allow us to combine results from both grades in the analysis. Twelve dummy variables were added to reflect the blocking scheme. Dummy variables were also used as part of the approach to address missing values for the covariates.

scale scores for students in *HMH Fuse* and control groups.[15] The Unadjusted row includes the raw means and standard deviations, as well as counts for students, sections, and schools for the analytical sample. The last two columns provide the effect size, that is, the size of the difference between the means for *HMH Fuse* and control groups in standard deviation units and percentile ranking. Also provided is the *p* value, indicating the probability of arriving at a difference with a magnitude as large as—or larger than—the magnitude of the one observed when there truly is no difference. The Adjusted row is based on the same sample of students. The mean difference—and therefore the effect size—is regression-adjusted, which means that the effects of chance differences between conditions on the covariates are factored out. This adjustment also increases the precision of the program effect estimate by accounting for variation in the outcome variable.

TABLE 25. EFFECT SIZES FOR THE CALIFORNIA STANDARDS TEST

	Condition	Means [c]	Standard deviations	No. of students	No. of sections	No. of teachers	Effect size	*p* value	Percentile standing
Unadjusted effect size[a]	**Control**	362.42	64.85	625	23	11	-0.004	.98	0%
	HMH Fuse	362.15	61.18	318	11	11			
Adjusted effect size[b]	**Control**	362.42	As above				0.04	.52	2%
	HMH Fuse	364.98							

[a] The unadjusted effect size is Hedges' g adjusted for clustering of students in sections (Hedges, 2006).

[b] The adjusted effect size was computed by dividing the regression-adjusted effect estimate by the standard deviation of the posttest scores for the control group. Between-grade differences in the control posttest were factored out of the standard deviation in the denominator of the effect size. The *p* value corresponds to the significance test for the effect of *HMH Fuse* in the regression model. The *HMH Fuse* mean was obtained by adding the regression-adjusted estimate of the average one-year effect of *HMH Fuse* to the unadjusted control mean.

[c] Modeling separate teacher effects leads to estimates of control-group performance which are specific to teachers. For purposes of display, to set the performance estimate for the control group, we compute the overall average performance for the sample of control cases used to calculate the adjusted effect size. The estimated *HMH Fuse* effect, which is constrained to be constant for each teacher-block (i.e., it is modeled as fixed) is added to this estimate to show the relative advantage or disadvantage to being in the *HMH Fuse* group.

The adjusted analysis shows no impact of *HMH Fuse* on student performance on the Algebra 1 CST.[16]

End of Course Assessment

In this section we address the impacts of *HMH Fuse* on the End of Course Assessment. Table 26 provides a summary of the samples used in the analysis and the results for the comparison of the scale scores for students in *HMH Fuse* and control groups. The Unadjusted row and the Adjusted row exhibit the same kind of information as was described in the previous section addressing the CST outcome.

[15] The full set of effect estimates for the analysis is given in Appendix A.

[16] We conducted a series of sensitivity analyses to determine the robustness of our result from the benchmark model to small changes in specification of the analytic model. We corroborated the results from the benchmark model in each case

TABLE 26. EFFECT SIZES FOR THE END OF THE COURSE ASSESSMENT

	Condition	Means [c]	Standard deviations	No. of students	No. of sections	No. of teachers	Effect size	*p* value	Percentile standing
Unadjusted effect size[a]	**Control**	19.73	6.99	398	16	9	-0.33	.30	-12.99%
	HMH Fuse	17.52	6.03	221	8	8			
Adjusted effect size[b]	**Control**	19.73	As above				0.02	.90	.64%
	HMH Fuse	19.84							

[a] The unadjusted effect size is Hedges' g adjusted for clustering of students in sections (Hedges, 2006).

[b] The adjusted effect size was computed by dividing the regression-adjusted effect estimate by the standard deviation of the posttest scores for the control group. Between-grade differences in the posttest were factored out of the standard deviation in the denominator of the effect size. The p value corresponds to the significance test for the effect of HMH Fuse in the regression model. The program mean was obtained by adding the regression-adjusted estimate of the average one-year effect of HMH Fuse to the unadjusted control mean.

[c] Modeling separate teacher effects leads to estimates of control-group performance which are specific to teachers. For purposes of display, to set the performance estimate for the control group, we compute the overall average performance for the sample of control cases used to calculate the adjusted effect size. The estimated *HMH Fuse* effect, which is constrained to be constant for each teacher-block (i.e., it is modeled as fixed) is added to this estimate to show the relative advantage or disadvantage to being in the *HMH Fuse* group.

The adjusted analysis shows no impact of *HMH Fuse* on student performance on the End of Course Assessment.[17]

Student Attitude Questionnaire

For this report, we analyzed the impact on the Motivated Strategies for Learning scale overall, as well as on each of the five subscales. We also analyzed impact on the usable items from the Self-Confidence subscale of the Attitudes toward Mathematics scale. See Data Sources and Collection section for details.

For each subscale, a response of 1 indicates that a student felt the statement was not at all like me and a score of 7 indicates that a student felt the statement was very much like me. Scores on scales and subscales were coded so that a high score indicates a positive identification with the scale and a low score indicates a negative identification with the scale. A high score indicates a positive student outlook.

Motivated Strategies for Learning Questionnaire

We calculated Cronbach's alpha (internal consistency) for each of the five subscales. As displayed in Table 27, the internal consistencies of the subscales are within the range generally considered acceptable to good.

[17] We conducted a series of sensitivity analyses to determine the robustness of our result from the benchmark model to small changes in specification of the analytic model. We corroborated the results from the benchmark model in each case

TABLE 27. INTERNAL CONSISTENCY FOR MOTIVATED STRATEGIES FOR LEARNING QUESTIONNAIRE SUBSCALES

Subscales	Coefficient alpha[18]
Cognitive Strategy Use subscale	0.83
Intrinsic Value subscale	0.84
Self-Efficacy subscale	0.89
Self-Regulation subscale	0.68
Test Anxiety subscale	0.80

Table 28 displays the impacts of *HMH Fuse* on the combined score of The Motivated Strategies for Learning Questionnaire. The Unadjusted row and the Adjusted row exhibit the same kinds of values and statistics as were described in the previous section addressing the CST outcome.

TABLE 28. EFFECT SIZES FOR MOTIVATED STRATEGIES FOR LEARNING QUESTIONNAIRE

	Condition	Means[c], [d]	Standard deviations	No. of students	No. of sections	No. of teachers	Effect size	*p* value
Unadjusted effect size[a]	**Control**	4.81	0.85	506	23	11	0.05	.71
	HMH Fuse	4.85	0.84	299	11	11		
Adjusted effect size[b]	**Control**	4.81	As above				0.15	.07
	HMH Fuse	4.94						

[a] The unadjusted effect size is Hedges' g adjusted for clustering of students in sections (Hedges, 2006).

[b] The adjusted effect size was computed by dividing the regression-adjusted effect estimate by the standard deviation of the posttest scores for the control group . The *p* value corresponds to the significance test for the effect of *HMH Fuse* in the regression model. The *HMH Fuse* mean was obtained by adding the regression-adjusted estimate of the average one-year effect of *HMH Fuse* to the unadjusted control mean.

[c] Modeling separate teacher effects leads to estimates of control-group performance which are specific to teachers. For purposes of display, to set the performance estimate for the control group, we compute the overall average performance for the sample of control cases used to calculate the adjusted effect size. The estimated *HMH Fuse* effect, which is constrained to be constant for each teacher-block (i.e., it is modeled as fixed) is added to this estimate to show the relative advantage or disadvantage to being in the *HMH Fuse* group.

[d] In the 2nd interim report we reported sums of the mean subscale scores. Here we report the average of the mean subscale scores, therefore the values are not the same.

[18] Cronbach's Alpha from the literature on the Motivated Strategies for Learning Questionnaire subscales are: Cognitive Strategy Use (alpha = .75), Intrinsic value (alpha = .87), Self-Efficacy (alpha = .89), Self-Regulation (alpha = .74), and Test anxiety (alpha = .75).

The adjusted analysis shows a positive impact of *HMH Fuse* on student attitudes toward math. The overall effect size (in standard deviation units) is 0.15. The low *p* value for the effect (.07) gives some confidence that the actual difference is different from zero. [19]

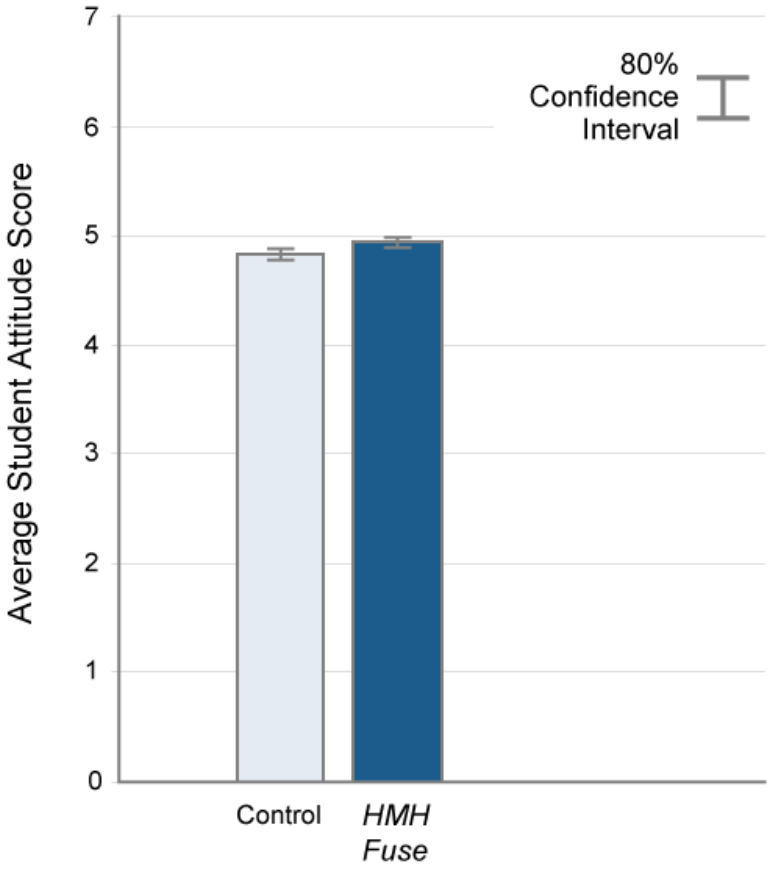


FIGURE 6. EFFECT SIZES FOR MOTIVATED STRATEGIES FOR LEARNING QUESTIONNAIRE

Figure 6 shows estimated performance on the posttest for the two groups. We added 80% confidence intervals to the tops of the bars in the figure. The lack of overlap in these intervals further indicates that the difference we observe is not simply due to chance.

Questionnaire Subscales

Table 29 displays the results for the impacts on each of the subscales. These being exploratory, we did not conduct sensitivity checks.

[19] We conducted a series of sensitivity analyses to determine the robustness of our result from the benchmark model to small changes in specification of the analytic model. The *p* values for the impact ranged between <.01 and .16. We conclude that as an initial exploration, an effect of *HMH Fuse* on this scale cannot be discounted; however, the observed impact is not robust to different specifications of the analytic model.

TABLE 29. EFFECT SIZES FOR MOTIVATED STRATEGIES FOR LEARNING QUESTIONNAIRE SUBSCALES

Subscale	Condition	Means[a]	Effect estimate (scale score units)	Effect size	*p* value
Cognitive Strategy Use	Control	4.73	0.08	0.08	.21
	HMH Fuse	4.81			
Intrinsic Value	Control	5.25	0.20	0.18	< .01
	HMH Fuse	5.45			
Self Efficacy	Control	5.19	0.14	0.12	.12
	HMH Fuse	5.33			
Self Regulation	Control	4.69	0.08	0.08	.33
	HMH Fuse	4.77			
Test Anxiety	Control	4.19	0.15	0.09	.22
	HMH Fuse	4.34			

[a] Modeling separate teacher effects leads to estimates of control-group performance which are specific to teachers. For purposes of display, to set the performance estimate for the control group, we compute the overall average performance for the sample of control cases used to calculate the adjusted effect size. The estimated *HMH Fuse* effect, which is constrained to be constant for each teacher-block (i.e., it is modeled as fixed) is added to this estimate to show the relative advantage or disadvantage to being in the *HMH Fuse* group.

We provide this table but caution against drawing firm conclusions from the results given the lack of robustness of the results for the combined scale. Future studies may wish to pay attention to the characteristics that are captured in the Intrinsic Value scale and to why this subscale appears to be especially sensitive to *HMH Fuse.*

Attitudes toward Mathematics Inventory

We examined the impact of *HMH Fuse* on the Self-Confidence component of the Attitudes toward Mathematics Inventory. As we described in an earlier section, our analysis was limited to two of the three items within this subscale because of an error in the construction of the questionnaire.

TABLE 30. INTERNAL CONSISTENCY FOR SELF-CONFIDENCE SUBSCALE (BASED ON THE TWO AVAILABLE ITEMS)

Subscale	Coefficient alpha
Self-Confidence	0.52

Here we report the coefficient alpha for the portion of the scale that is available for analysis, as well as the results of the impact analysis. Because responses to only part of the scale are available, we consider this to be exploratory and, therefore, do not conduct sensitivity checks.

The Cronbach's Alpha coefficient, displayed in Table 30, demonstrates poor internal consistency.

Table 31. displays the results for the impact of *HMH Fuse* on the Self-Confidence subscale. With an effect size of .04 and a high *p* value, there is no observed impact of *HMH Fuse* on this scale.

TABLE 31. EFFECT SIZES FOR SELF-CONFIDENCE SUBSCALE (BASED ON THE TWO AVAILABLE ITEMS)

Subscale	Condition	Means [a]	Effect estimate (scale score units)	Effect size	*p* value
Self-confidence	Control	4.89	0.06	0.04	.51
	HMH Fuse	4.95			

[a] Modeling separate teacher effects leads to estimates of control-group performance which are specific to teachers. For purposes of display, to set the performance estimate for the control group, we compute the overall average performance for the sample of control cases used to calculate the adjusted effect size. The estimated *HMH Fuse* effect, which is constrained to be constant for each teacher-block (i.e., it is modeled as fixed) is added to this estimate to show the relative advantage or disadvantage to being in the *HMH Fuse* group.

Moderation of the Impact

Next, we report our analysis of the moderating effects of pretest performance and English speaker status. We are interested primarily in the interaction between the moderating variable and program status, that is, whether the impact of *HMH Fuse* varies across levels of the moderating variable.

Including Pretests as a Moderator

CST Pretest and CST Achievement

We first show whether the impact of *HMH Fuse* varies for students at different levels of prior achievement on the CST. The 'Fixed Effects' in Table 32 provide the estimates of primary interest, including an estimate of the change in the impact of *HMH Fuse* for a 1-unit increase on the CST pretest. At the bottom of the table we give results for technical review—these often consist of what are called random effects estimates. Random effects are added to the statistical equation to account for dependencies in observed scores that happen because students come from the same sections.

TABLE 32. MODERATING EFFECT OF THE CST PRETEST ON THE IMPACT OF *HMH FUSE* ON CST ACHIEVEMENT

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Intercept: Outcome for the control student with an average pretest in the reference section with zero values for the covariates.	343.88	7.74	22	44.41	< .01
Change in outcome for the control student for each unit-increase on the pretest	36.49	2.41	867	15.14	< .01
Effect of *HMH Fuse* for a student with an average pretest	2.32	3.71	22	0.62	.54
Change in the effect of *HMH Fuse* for each unit-increase on the pretest	-0.04	3.43	867	-0.01	.99
Random effects	**Estimate**	**Standard error**		***z* value**	***p* value**
Section mean achievement	98.34	52.81		1.86	.03
Within-section variation	1901.59	91.31		20.83	<.01

[a] We do not display the fixed effect estimates for teachers or covariates used to improve precision. The intercept value refers to a specific teacher (arbitrarily chosen by the estimation routine) and for cases with zero values for the covariates. The results in the rest of the table do not depend on the teacher that the intercept refers to.

Note. The prior score was centered on its mean value.

The moderating effect of the CST pretest score on the impact of *HMH Fuse*, that is, whether the intervention was differentially effective for students at different points along the pretest scale, is shown in the fourth row. The coefficient, -0.04 is a very small difference in the impact associated with each one-unit increase on the pretest. The *p* value of .99 indicates that we can have no confidence that the true differential impact is different from zero. The impact of *HMH Fuse* was not different depending on the student's pretest scores on the CST.

Holt McDougal Algebra 1 Pretest and End of Course Achievement

Here we analyze whether the impact of *HMH Fuse* varies for students at different levels of prior achievement on the End of Course Assessment. The 'Fixed Effects' in Table 33 provide the estimates of primary interest, including an estimate of the change in the impact of *HMH Fuse* for a 1-unit increase on the Holt McDougal Algebra 1 pretest.

The moderating effect of the pretest score on the impact of *HMH Fuse*, that is, whether the intervention was differentially effective for students at different points along the pretest scale is shown in the fourth row. The coefficient, .003, is a very small difference in the impact associated with each one-unit increase on the pretest. The *p* value of .97 indicates that we can have no confidence that the true differential impact is different from zero. The impact of *HMH Fuse* was not different depending on the student's pretest score.

TABLE 33. MODERATING EFFECT OF HOLT MCDOUGAL ALGEBRA 1 PRETEST ON THE IMPACT OF *HMH FUSE* ON END OF COURSE ACHIEVEMENT

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Intercept: Outcome for the control student with an average pretest in the reference section with zero values for the covariates.	18.00	1.24	14	14.51	< .01
Change in outcome for the control student for each unit-increase on the pretest	0.31	0.09	531	3.43	< .01
Effect of *HMH Fuse* for a student with an average pretest	0.46	0.83	14	0.56	.58
Change in the effect of *HMH Fuse* for each unit-increase on the pretest	0.003	0.10	531	0.03	.97
Random effects	**Estimate**	**Standard error**		***z* value**	***p* value**
Section mean achievement	3.79	1.80		2.10	.02
Within-section variation	21.37	1.31		16.31	< .01

[a] We do not display the fixed effect estimates for teachers or covariates used to improve precision. The intercept value refers to a specific teacher (arbitrarily chosen by the estimation routine) and for cases with zero values for the covariates. The results in the rest of the table do not depend on the teacher that the intercept refers to.

Note. The prior score was centered on its mean value.

Student Attitude Questionnaire Pretest and Attitudes toward Math

Here we show whether the impact of *HMH Fuse* on student attitudes varies for students with different pretest scores on the math attitudes survey. The 'Fixed Effects' in Table 34 provide the estimates of primary interest, including an estimate of the change in the impact of *HMH Fuse* for a 1-unit increase on the pretest.

TABLE 34. THE MODERATING EFFECT OF THE STUDENT ATTITUDE QUESTIONNAIRE PRETEST ON THE IMPACT OF *HMH FUSE* ON ATTITUDE TOWARD MATH

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Intercept: Outcome for the control student with an average residualized pretest score in the reference section with zero values for the covariates.	4.34	0.16	21	27.14	< .01
Change in outcome for the control student for each unit-increase on the residualized pretest score	0.66	0.06	660	11.81	< .01
Effect of *HMH Fuse* for a student with an average pretest	0.11	0.08	21	1.39	.18
Change in the effect of *HMH Fuse* for each unit-increase on the pretest	-.08	0.09	660	-.91	.36
Random effects	**Estimate**	**Standard error**		***z* value**	***p* value**
Section mean achievement	0.03	0.02		1.76	.04
Within-section variation	0.43	0.02		18.18	< .01

[a] We do not display the fixed effect estimates for teachers or covariates used to improve precision. The intercept value refers to a specific teacher (arbitrarily chosen by the estimation routine) and for cases with zero values for the covariates. The results in the rest of the table do not depend on the teacher that the intercept refers to.

Note. The prior score was centered on its mean value.

The moderating effect of the pretest score on the impact of *HMH Fuse*, that is, whether the intervention was differentially effective for students at different points along the pretest scale, is shown in the fourth row. The coefficient, -.08, is a very small difference in the impact associated with each one-unit increase on the pretest. The *p* value of .36 indicates that we can have no confidence that the true differential impact is different from zero. The impact of *HMH Fuse* on a student's attitude toward math was not different depending on the student's attitude at the start of the study.

Including ELL Status as a Moderator

English Proficiency and CST Achievement.

We were also interested in the moderating effect of student English proficiency. In other words, we were interested in whether *HMH Fuse* was differentially effective for English proficient students and for English learners.

The 'Fixed Effects' in Table 35 provide the estimates of primary interest, including an estimate of the difference between English proficient and non-English proficient in the impact of *HMH Fuse* on CST achievement.

TABLE 35. THE MODERATING EFFECT OF ENGLISH PROFICIENCY ON THE IMPACT OF *HMH FUSE* ON CST ACHIEVEMENT

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Intercept: Outcome for the Non-English proficient control with an average pretest in the reference section with zero values for the covariates.	352.04	9.40	22	37.45	< .01
Change in outcome for each unit-increase on the pretest	35.49	2.24	896	15.86	< .01
Control group difference (English proficient minus not proficient) in the outcome.	-4.65	8.65	22	-0.54	.60
Effect of *HMH Fuse* for Non-English proficient student	1.99	8.64	896	0.23	.82
Average difference (English proficient minus not proficient) in the effect of *HMH Fuse*	8.14	9.33	896	0.87	.38
Random effects	**Estimate**	**Standard error**		***z* value**	***p* value**
Section mean achievement	116.84	59.60		1.96	.03
Within-section variation	2031.90	96.00		21.16	<.01

[a] We do not display the fixed effect estimates for teachers or covariates used to improve precision. The intercept value refers to a specific teacher (arbitrarily chosen by the estimation routine) and for cases with zero values for the covariates. The results in the rest of the table do not depend on which teacher the intercept refers to.

Note. The prior score was centered on its mean value.

The estimate of whether *HMH Fuse* was differentially effective for English proficient and English non-proficient students is shown in the fifth row. The coefficient is 8.14. The *p* value of .38 indicates that we can have no confidence that the true differential impact is different from zero. The impact of *HMH Fuse* on a student's CST score was not different depending on the student's English proficiency status.

English Proficiency and End of Course Achievement.

Here we show whether the impact of *HMH Fuse* on the End of Course Assessment varies for students with different English proficiency status. The 'Fixed Effects' in Table 36 provide the estimates of primary interest, including an estimate of the difference between English proficient and non-English proficient in the impact of *HMH Fuse*.

TABLE 36. THE MODERATING EFFECT OF ENGLISH PROFICIENCY ON THE IMPACT OF *HMH FUSE* ON END OF COURSE ACHIEVEMENT

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Intercept: Outcome for the Non-English proficient control with an average pretest in the reference section with zero values for the covariates.	19.54	1.14	14	17.09	< .01
Change in outcome for each unit-increase on the pretest	0.28	0.07	583	4.27	< .01
Control group difference (English proficient minus English non-proficient) in the outcome (n)	0.26	0.78	583	0.33	.75
Effect of *HMH Fuse* for English non-proficient student	-1.31	0.69	14	-1.89	.08
Average difference (English proficient minus English non-proficient) in the effect of *HMH Fuse*	1.69	0.99	583	1.71	.09
Random effects	**Estimate**	**Standard error**		***z* value**	***p* value**
Section mean achievement	3.98	1.85		2.15	.02
Within-section variation	21.67	1.27		17.08	< .01

[a] We do not display the fixed effect estimates for teachers or covariates used to improve precision. The intercept value refers to a specific teacher (arbitrarily chosen by the estimation routine) and for cases with zero values for the covariates. The results in the rest of the table do not depend on which teacher the intercept refers to.

Note. The prior score was centered on its mean value.

The estimate of the moderating effect of English proficiency status on the impact of *HMH Fuse*, that is, whether the intervention was differentially effective for English proficient and non-proficient students is shown in the fifth row. The coefficient is 1.69. The *p* value of .09 indicates that we can have some confidence that the true differential impact is different from zero.

Figure 7 provides a graphical representation of the differential effect estimate. We have some confidence of a negative effect of *HMH Fuse* on the End of Course Assessment for English non-proficient students ($p = .08$), and no confidence that there is an impact of *HMH Fuse* on the End of Course Assessment for English proficient students ($p = .69$).

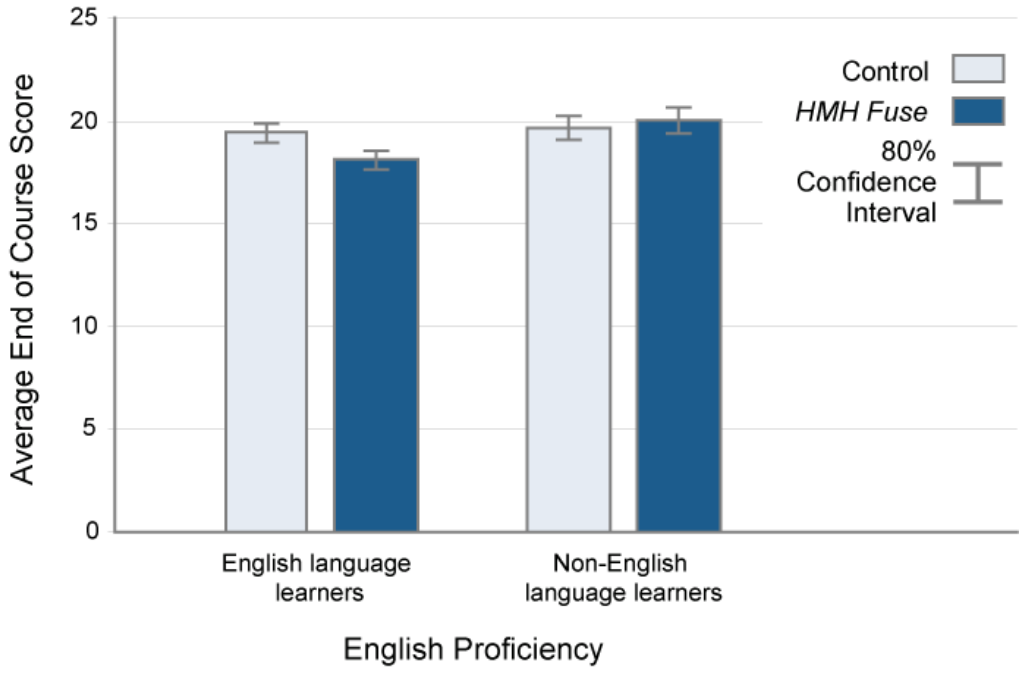


FIGURE 7. THE MODERATING EFFECT OF ENGLISH PROFICIENCY STATUS ON THE IMPACT OF *HMH FUSE* ON END OF COURSE ACHIEVEMENT

English Proficiency and the Student Attitude Questionnaire

Here we show whether the impact of *HMH Fuse* on student attitudes varies for students with different English proficiency status. The 'Fixed Effects' in Table 37 provide the estimates of primary interest, including an estimate of the difference between English proficient and non-proficient in the impact of *HMH Fuse*.

TABLE 37. THE MODERATING EFFECT OF ENGLISH PROFICIENCY ON THE IMPACT OF *HMH FUSE* ON THE STUDENT ATTITUDE QUESTIONNAIRE

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Intercept: Outcome for the English non-proficient control with an average pretest residual in the reference section with zero values for the covariates.	4.56	0.19	22	23.39	<.01
Change in outcome for each unit-increase on the residualized pretest	0.62	0.04	758	14.30	<.01
Control group difference (English proficient minus English non-proficient) in the outcome	-0.02	0.14	758	-0.12	.90
Effect of *HMH Fuse* for English non-proficient student	-0.14	0.19	22	-0.73	.47

TABLE 37. THE MODERATING EFFECT OF ENGLISH PROFICIENCY ON THE IMPACT OF *HMH FUSE* ON THE STUDENT ATTITUDE QUESTIONNAIRE

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Average difference (English proficient minus English non-proficient) in the effect of *HMH Fuse*	0.31	0.18	758	1.69	.09
Random effects	**Estimate**	**Standard error**		***z* value**	***p* value**
Section mean achievement	0.02	0.01		1.75	.04
Within-section variation	0.46	0.02		19.50	<.01

[a] We do not display the fixed effect estimates for teachers or covariates used to improve precision. The intercept value refers to a specific teacher (arbitrarily chosen by the estimation routine) and for cases with zero values for the covariates. The results in the rest of the table do not depend on which teacher the intercept refers to.

Note. The prior score was centered on its mean value.

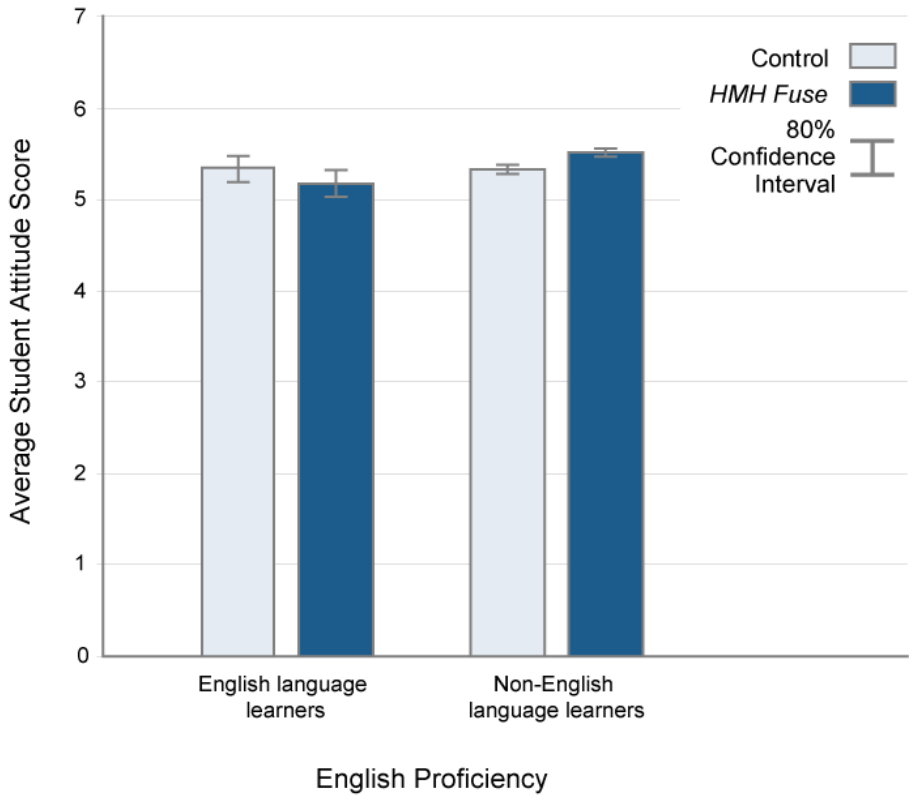


FIGURE 8. IMPACT OF *HMH FUSE* ON THE STUDENT ATTITUDE QUESTIONNAIRE FOR ENGLISH PROFICIENCY STATUS

The estimate of the moderating effect of English proficiency status on the impact of *HMH Fuse*, that is, whether the intervention was differentially effective for English proficient and non-proficient students is shown in the fifth row. The coefficient is .31. The *p* value of .09 indicates that we can have some confidence that the true differential impact is different from zero.

Figure 8 provides a graphical representation of the differential effect estimate. We have no confidence of an impact of *HMH Fuse* on the Attitude Questionnaire for English non-proficient students ($p = .47$), and a high-level of confidence that there is a positive impact of *HMH Fuse* on the Student Attitude Questionnaire for English Proficient students ($p = .01$).

Mediation of the Impact of *HMH Fuse*

Mediation is considered to occur when an impact of the program on student achievement happens through a prior impact on an intermediate variable. If an impact is demonstrated on the intermediate variable, and we can also establish an association between the intermediate variable and student achievement independent of the effect of the program, then the intermediate variable may be a mediator of the impact on achievement.

We assess mediation whether or not there is an overall impact on student achievement because the mediating path that we are investigating may be one of several, and their effects may cancel when combined, leading to zero overall effect. Therefore, lack of an overall impact does not rule out mediation along the path of interest. On the other hand, if there is no impact on the posited mediator of interest, then we do not consider that mediating path further.

In this section we examine the mediating role of three variables: (1) time spent studying algebra outside of the classroom, (2) number of algebra videos watched and (3) student attitude as measured by the questionnaire. See Appendix B.

We also examine whether there was an association between each posited mediator and student performance on the CST. Although this step is not a component of the mediation analysis proper, it gives descriptive information about whether there is a relationship between the intermediate process variable and student achievement. This is a purely exploratory outcome that we think may be of interest to the developer.

Time Spent on Algebra Outside of Class

We look first at the impact of *HMH Fuse* on time spent studying algebra outside of the classroom. Table 38 shows the result. The *p* value of .10 indicates gives some confidence that there is an impact.

TABLE 38. IMPACT OF *HMH FUSE* ON TIME SPENT ON ALGEBRA OUTSIDE OF CLASS

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Outcome for the control student	148.52	14.31	21	10.38	<.01
Effect of *HMH Fuse*	18.39	10.69	21	1.72	.10
Random effects	**Estimate**	**Standard error**		***z* value**	***p* value**
Residual variation	790.64	244.00		3.24	<.01

[a] We do not display the fixed effect estimates for teachers or covariates used to improve precision. The intercept value refers to a specific teacher (arbitrarily chosen by the estimation routine) and for cases with zero values for the covariates. The results in the rest of the table do not depend on which teacher the intercept refers to.

Note. The prior score was centered on its mean value.

With a positive impact on the proposed mediating variable, we used the software RMediation (Tofighi & MacKinnon, 2011) to obtain an estimated mediated effect of 1.511. The 80% confidence interval for the mediated effect is (-0.042, 3.528). Because the confidence interval spans zero we have no confidence that the impact on the section-level average of time spent studying algebra outside the class mediates an effect of *HMH Fuse* on student achievement.

As an additional exploratory step we examined whether there was an association between the mediator—the section-average time spent on the algebra program outside of class—and student performance on the CST. We expect a 0.08 point gain on the CST for each unit increase in the section-level average of time spent on the algebra program outside the class. With $p = .18$, we have limited confidence in this result.

Number of Algebra Videos Watched

We look first at the impact of *HMH Fuse* on the number of algebra videos watched outside of the classroom. Table 39 shows the distribution of students' responses. There is a difference between conditions in the proportion of students who report watching 0-7 videos versus 8 or more videos. We have a high level of confidence in this result ($p < .01$), based on a multilevel logistic regression equation (with a section random effect).

TABLE 39. FREQUENCIES OF MEDIAN RESPONSES TO THE NUMBER OF ALGEBRA VIDEOS WATCHED PER WEEK

	0-7	8 or more	Total
Control	538 (95.73%)	24 (4.27%)	562
HMH Fuse	252 (82.89%)	52 (17.11%)	304
Total	790 (91.22%)	76 (8.78%)	866

Analysis of a mediating effect requires a non-zero impact on the mediator and, with the software that we are using, RMediation (Tofighi & MacKinnon, 2011), that the mediator variable is normally distributed. The latter condition does not apply; therefore we did not run a formal mediation analysis.

As an additional exploratory step we investigated the association between the posited mediator—the median range of number of videos watched (0-7 or above)—and student performance. We have no confidence that the median level of the number of videos watched per week has a significant association with student performance on the CST ($p = .61$).

Attitudes toward Math

Our examination of the Motivated Strategies for Learning scale gave us some confidence that *HMH Fuse* had a positive impact on student attitudes toward math (p value = .07).

We estimated a mediated effect of 2.30 scale units. The 80% confidence interval for this effect is (0.93, 3.73). Based on the 80% confidence interval we concluded that the intervention program has a positive indirect effect on student achievement through the Motivated Strategies for Learning scale.[20]

As an additional exploratory step we examined whether student attitude toward math is associated with student achievement; in other words, whether attitude toward math is predictive of CST achievement. We observe a positive relationship between the Motivated Strategies for Learning scale and student achievement. Students who report higher motivation levels outperform students who report lower motivation levels by an average of 15.91 scale score points. The low p value (< .01) gives us strong confidence in this result.

[20] A positive indirect impact and a zero overall impact suggest that there may also exist a negative indirect impact occurring through an unidentified mediator.

ADDITIONAL RESULTS

Results for Each of the End of Chapter Tests

Table 40 looks at the difference in average mean scores between the control and *HMH Fuse* groups. This table is provided for potential formative value but no research conclusions should be drawn from the results.[21]

TABLE 40. IMPACT ON THE END OF CHAPTER TEST

Chapter	Condition	No. of sections	Estimate	*p* value
Chapter 1	Control	15	-0.27	.45
	HMH Fuse	9		
Chapter 2	Control	24	0.18	.64
	HMH Fuse	11		
Chapter 3	Control	23	0.04	.86
	HMH Fuse	11		
Chapter 4	Control	12	-0.17	.59
	HMH Fuse	5		
Chapter 5	Control	23	-0.10	.67
	HMH Fuse	11		
Chapter 6	Control	23	0.38	.23
	HMH Fuse	11		
Chapter 7	Control	23	-0.16	.70
	HMH Fuse	11		
Chapter 8	Control	22	0.24	.64
	HMH Fuse	11		
Chapter 9	Control	23	-0.28	.44
	HMH Fuse	11		
Chapter 10	Control	21	-0.14	.53
	HMH Fuse	9		
Chapter 11	Control	9	0.08	.86
	HMH Fuse	5		

We do not see a difference between conditions in average performance for any of the unit tests.

[21] The results were corroborated using a non-parametric test (Wilcoxon Rank-Sum Test). The results were robust to this alternative analysis—there were no differences in conclusions concerning statistical significance of the results.

Associations between Application Log Data and Student Outcomes

Table 41 looks at the association of usage of various application features (as determined by number of clicks) and student achievement on the three outcome measures. Because these data come from the iPad application, this information is only available for *HMH Fuse* students. The results should be considered for formative purposes only and no research conclusions should be drawn from the results.

TABLE 41. ASSOCIATIONS BETWEEN STUDENT ACTIVITY AND STUDENT ACHIEVEMENT

Feature	CST		End of Course Assessment		Motivated Strategies for Learning Questionnaire	
	Estimate	*p* value	Estimate	*p* value	Estimate	*p* value
Total Usage (Overall counts)	7.66	.21	0.20	.77	0.10	.28
Math in Motion	0.46	.92	-0.42	.39	0.03	.52
Note-taking	3.81	.08	0.02	.96	0.03	.44
Quizzes	8.56	< .01	0.11	.79	0.09	.07
Inline Example References	-2.61	.18	-0.23	.58	-0.01	.76
Answer Checking (Homework)	3.52	.18	-0.06	.85	0.04	.19
Homework Help Walkthrough	-1.22	.77	-0.67	.19	0.07	.21
Videos	3.23	.32	0.35	.22	0.04	.46

Note. All the counts are log-transformed. Each estimate is the coefficient in a regression of the achievement score against the log-counts for the given activity, with adjustment for clustering of outcomes in sections.

Although several of the associations achieve statistical significance, we do not adjust the results for multiple comparisons; therefore, given the number of outcomes, we expect a proportion of outcomes to reach significance by chance. The significant results should be corroborated through follow-up analyses.

Results for the Riverside Subgroup

Before our work on the average impact of *HMH Fuse* for the sample overall was completed, one of the districts—Riverside Unified—made public the results for the two teachers (and nine sections of Algebra 1) that had participated. Interestingly, their sample constitutes a very small RCT since randomization was blocked within teacher and the Riverside personnel who examined their own local data had no information on the results from the other districts. While not using appropriate statistical adjustments, they did report substantial differences between the data from students who

had the iPads and those using the conventional textbook. The results were later written up in a case study report by HMH (Houghton Mifflin Harcourt, n.d.), which described the school as follows:

The school "has been a school eager to employ new technology in the classroom–[it] had been an early adopter of a student laptop program, and was among the first schools to install interactive whiteboards. Naturally, the iPad *HMH Fuse* study was a perfect fit for the school ... The school's past experience with new technology had taught [its] teachers and leadership that the right implementation and support was vital to a successful experience." (p. 5)

We subsequently investigated whether *HMH Fuse* has a stronger impact for Riverside compared to the rest of the sample. We note that the hypothesis that the impact for Riverside will be greater than for the remainder of the sample was generated in the course of the study; therefore, though it was not identified ahead of time, it also was not proposed after seeing the results.[22] Indeed, from the point of view of Riverside, which initiated the examination, it was a test of *their* central hypothesis. Though the result is not strictly post hoc, the small sample admits plausible rival explanations for any observed differences. These would have to be ruled out with further experimentation.[23] With these caveats we analyze the results for Riverside.

Sample and Distribution of Background Characteristics

We limited the sample to students in Riverside with non-missing CST posttests. Students in grades 7 and 8 took the same test – Algebra 1 California Standards Test. The sample is broken down as follows.

---

[22] By the standards of the Institute of Education Sciences, these findings are considered exploratory. Its guidance has been stated as follows:

> The purpose of the *exploratory* analysis is to examine relationships within the data to identify outcomes or subgroups for which impacts may exist. The goal of the exploratory analysis is to identify hypotheses that could be subject to more rigorous future examination, but cannot be examined in the present study because they were not identified ahead of time or statistical power was deemed insufficient. Results from post hoc analyses are not automatically invalid, but, irrespective of plausibility or statistical significance, they should be regarded as preliminary and unreliable unless they can be rigorously tested and replicated in future studies. (Schochet, 2008, p. 4).

We note, however, that this was initiated by Riverside as a direct investigation of the impact on the local sample available to them.

[23] With small samples it is easy to get imbalance between conditions on observed or unobserved factors that influence performance. Also, with only two teachers in the analysis, characteristics unique to those teachers may be interacting with treatment to produce the effects – they may be exceptional in their use of the program. However, the result at Riverside is experimental, with randomization of sections conducted within teachers within the site; therefore, if the two teachers are exceptional, their effects on performance, independently of the effects of the usage of *HMH Fuse* with the sections assigned to the intervention, will be distributed evenly between their treatment and control sections. To corroborate the results reported here, and be more conclusive about results for Riverside, we recommend a replication trial for a wider sample of teachers from Riverside (selected randomly or purposively).

TABLE 42. SAMPLE SIZES IN RIVERSIDE

	Control	*HMH Fuse*	Total
N (Teachers)	2	2	2 [a]
N (Sections)	7	2	9
N (Students)	197	64	261

[a] Sections are randomized to conditions within teachers; therefore the teachers are represented in both conditions.

We tested whether characteristics of students were distributed differently between conditions for the Riverside sample. Results in Table 43 indicate balance.

TABLE 43. CHARACTERISTICS OF STUDY SAMPLE (ANALYTICAL) FOR THE RIVERSIDE SUBEXPERIMENT

	Control	*HMH Fuse*	Total	Less than 5% chance of seeing this much imbalance
Student characteristics				
Male	86 (43.65%)	31 (48.44%)	117 (44.83%)	No
Grade 8	152 (77.16%)	50 (78.13%)	202（77.39%）	No
Disability	4 (2.03%)	0 (0.00%)	4（1.53%）	No
English Speaker	195 (98.98%)	64 (100.00%)	259（99.23%）	No
Asian	16 (8.51%)	8 (12.70%)	24（9.56%）	
White	117 (62.23%)	31 (49.21%)	148（58.96%）	
Black	17 (9.04%)	8 (12.70%)	25（9.96%）	No
Mixed	0 (0.00%)	0 (0.00%)	0（0.00%）	
Indian	2 (1.06%)	0 (0.00%)	2（0.80%）	
Hispanic	36 (19.15%)	16 (25.40%)	52（20.72%）	

Note. Percentages are based on the number of students. The tests of equivalence are based on section-means of the student characteristics. Although there is no obvious large imbalance, and the statistical tests corroborate this, the small sample size gives limited sensitivity to detect possible differences between conditions in the distributions of the characteristics.

We conducted both t-tests and Wilcoxon tests to assess equivalence. The results from both types of tests were consistent. For the test of overall balance across categories of ethnicity, we ran Fisher's test.

Impact on the CST for the Riverside Sample

We start by showing the means and standard deviations for the CST posttests in each section, with posttest means listed from smallest to largest.

TABLE 44. RAW OUTCOMES BY SECTIONS FOR RIVERSIDE

	Teacher (1 or 2)	Condition (*HMH Fuse* or Control)	Posttest Mean (standard error)	Posttest standard deviation
Section 1	1	Control	330.05	48.25
Section 2	1	Control	347.88	61.79
Section 3	2	Control	354.61	56.56
Section 4	1	Control	356.97	77.64
Section 5	2	Control	376.70	57.40
Section 6	1	*HMH Fuse*	377.56	58.03
Section 7	2	Control	393.97	61.89
Section 8	2	*HMH Fuse*	400.32	55.56
Section 9	1	Control	417.59	56.22

Next, we applied the analytic model used to get the impact for the sample overall. The main results are displayed in Table 45. We observe a positive impact of *HMH Fuse* for the sample in Riverside. The impact is .23 effect size units, and we have strong confidence that the effect is not just a chance result.

TABLE 45. EFFECT SIZES FOR CST RIVERSIDE SAMPLE

	Condition	Means[c]	Standard deviations	No. of students	No. of sections	No. of teachers	Effect size	*p* value[d]	Percentile standing
Unadjusted effect size[a]	**Control**	368.21	66.35	197	7	2	0.37	0.141	14%
	HMH Fuse	392.16	56.17	64	2	2			
Adjusted effect size[b]	**Control**	368.21	As above				0.23	0.023	9%
	HMH Fuse	381.44							

[a] The unadjusted effect size is Hedges' g adjusted for clustering of students in sections (Hedges, 2006).

[b] The adjusted effect size was computed by dividing the regression-adjusted effect estimate by the standard deviation of the posttest scores for the control group. Between-grade differences in the posttest were factored out of the standard deviation in the denominator of the effect size. The p value corresponds to the significance test for the effect of *HMH Fuse* in the regression model. The treatment mean was obtained by adding the regression-adjusted estimate of the average one-year effect of *HMH Fuse* to the unadjusted control mean.

[c] Modeling separate teacher effects leads to estimates of control-group performance which are specific to teachers. For purposes of display, to set the performance estimate for the control group, we compute the overall average performance for the sample of control cases used to calculate the adjusted effect size. The estimated treatment effect, which is constrained to be constant for each teacher-block (i.e., it is modeled as fixed) is added to this estimate to show the relative advantage or disadvantage to being in the treatment group.

[d] The *p* value corresponds to a 1-tailed test of the statistical hypothesis that the impact at Riverside is greater than null. This reflects the intention to corroborate the positive findings reported by the district.

Sensitivity Checks

We conducted several sensitivity checks to examine whether the benchmark result holds up to small variations in the analytic approach[24]. All results that were based on alternative multilevel models were consistent with the main result. We used a model where pretest, grade and variables identifying teachers were the only covariates, with missing values for the covariates handled using the same method as the one for analyzing the average impact for the whole sample (p = .014), and with the same model, but where we removed cases with missing values for the covariates (p = .024). We also obtained an estimate of the impact for Riverside from the model used to assess the differential impact (see next section), (p = .011). In addition, we used the model for getting the benchmark result, but where we removed cases with missing values for any covariate (p=.031).

Differential Impact on the CST

In addition to considering whether there is an impact of *HMH Fuse* on the Riverside sample, we examined whether there is a difference between Riverside and the rest of the study sample in the impact. This is equivalent to a moderator analysis, with an indicator of membership in Riverside being the potential moderating characteristic. The sample used in the analysis is displayed in Table 46.

TABLE 46. COUNTS BY LEVEL OF THE MODERATOR

Condition	Moderator level	Number of sections	Number of teachers[a]	Number of students
HMH Fuse	Riverside	2	2	64
	Non-Riverside	9	9	254
Control	Riverside	7	2	197
	Non-Riverside	16	9	428

[a] Teachers have sections in both conditions and therefore are counted twice, once in each condition.

The results of the analysis are displayed in Table 47. We observe a differential impact of 12.23 scale score units (p = .077) giving us some confidence that the effect of *HMH Fuse* is not the same in Riverside as the rest of the study sample.

[24] Consistent with the benchmark analysis, the p values for the sensitivity analyses also correspond to one-tailed tests.

TABLE 47. MODERATING EFFECT OF MEMBERSHIP IN RIVERSIDE ON THE IMPACT OF *HMH FUSE* ON THE CST

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Intercept: Outcome for the non-Riverside control student with an average pretest and with zero values for the covariates.	349.82	7.24	22	48.35	<.001
Change in outcome for each one standard deviation increase on the pretest	35.40	2.22	896	15.98	<.001
Control group difference (Riverside minus non-Riverside) in the outcome.	9.49	8.94	22	1.06	.300
Effect of *HMH Fuse* for Non-Riverside student	-0.28	4.44	22	-0.06	.950
Average difference (Riverside minus non-Riverside) in the effect of *HMH Fuse*	12.23	6.58	22	1.86	.077
Random effects[b]	**Estimate**	**Standard error**		***z* value**	***p* value**
Section mean achievement	119.11	62.00		1.92	.027
Within-section variation	2030.54	95.90		21.17	<.001

We used this model to assess also the impact in Riverside, and separately for the sample outside Riverside. The impact in Riverside was 11.95 scale score units (p = .011) (we noted this result above as a sensitivity check for impact at Riverside), and the impact outside Riverside was -0.28 scale score units (p = .950). We have strong confidence of an impact at Riverside, and no confidence of an impact for the sample excluding Riverside.

Figure 9 represents the differential effect reported in Table 47. We notice that there is no difference between control and *HMH Fuse* for the non-Riverside districts but a distinct advantage for the *HMH Fuse* group in Riverside. As we note in the next section, the difference in pretest scores and other demographic differences do not explain the differential effectiveness since we showed that these factors do not moderate the impact of *HMH Fuse* on CST scores.

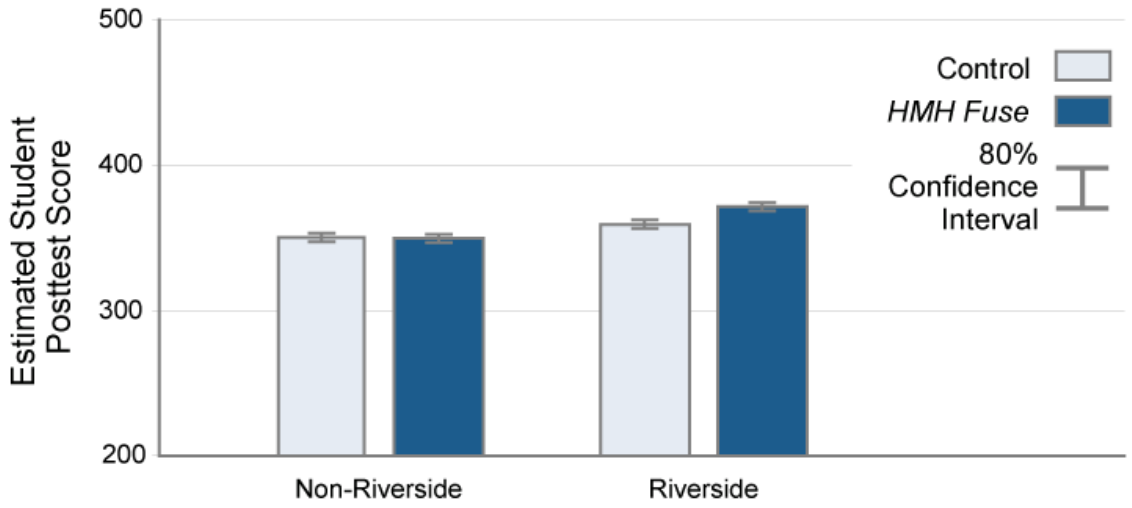


FIGURE 9. MODERATING EFFECT OF MEMBERSHIP IN RIVERSIDE ON THE IMPACT OF *HMH FUSE* ON THE CST

Potential Explanations for the Effect in Riverside

In this section we explore differences between Riverside and the other districts that may provide an explanation for the greater impact of *HMH Fuse*. These explorations are meant to be suggestive of areas for further investigation both in the current data as well as in new studies.

We compared the Riverside sample to the rest of the sample in terms their demographic characteristics and in terms of levels of implementation. For the latter, we focused on specific activities in the Riverside sections assigned to *HMH Fuse* compared to the other sections, examining: (1) overall number of clicks (2) the number of quiz data clicks (3) video usage (4) the amount of time teachers spent using the iPad for instruction and (5) the amount of time students spent using the iPad in the classroom. As explorations, we have not conducted statistical tests in all cases.

Differences in the Sample

Table 48 summarizes the tables in the earlier part of the report to emphasize the contrast of Riverside with the other districts.

TABLE 48. CHARACTERISTICS OF STUDY SAMPLE (ANALYTICAL)

	Non-Riverside	Riverside	Less than 5% chance of seeing this much imbalance
Student characteristics			
Male	357 (52.35%)	117 (44.83%)	No
Grade 8	596 (87.39%)	202 (77.39%)	No
Disability	11 (1.61%)	4 (1.53%)	No
English Speaker	589(86.36%)	259(99.23%)	Yes
Asian	237 (34.96%)	24 (9.56%)	
White	35 (5.16%)	148 (58.96%)	Yes
Black	56 (8.26%)	25 (9.96%)	

TABLE 48. CHARACTERISTICS OF STUDY SAMPLE (ANALYTICAL)

	Non-Riverside	Riverside	Less than 5% chance of seeing this much imbalance
Mixed	24 (3.54%)	0	
Indian	0	2 (0.80%)	
Hispanic	326 (48.08%)	52 (20.72%)	
Mean pre-test score	-0.15	0.44	Yes
Teacher characteristics			
Fewer than 4 years teaching experience	0	0	No

While we see differences in English proficiency (Riverside has very few English learners), ethnicity, and pretest score, we also note that none of these variables individually moderated the impact for the overall sample and therefore do not explain the different outcome for Riverside on the CST. (While it is not reported here, we ran a moderator analysis on ethnicity and found no effect.)

Overall Number of Clicks

We observed no difference between Riverside and non-Riverside in the overall number of clicks (p = .66). (Outcomes were log transformed because of excessive skew. The regression-adjusted difference in the log counts was .04.)[25]

Quiz Data Clicks

We observed a difference between Riverside and non-Riverside in the log count of the number of quiz data clicks (p < .01). Students in Riverside clicked more. The regression adjusted difference in log counts was .63.[26]

Video

We observed a difference between Riverside and the rest of the sample in the median number of videos viewed. For this outcome, students in Riverside were less likely to report a median number of videos watched of eight or more (p = .02).[27]

[25] The analysis adjusted for clustering of students in sections. We corroborated the result using the Wilcoxon test with non-transformed outcome data.

[26] The analysis adjusted for clustering of students in sections. We corroborated the result using the Wilcoxon test with non-transformed outcome data.

[27] We ran a parametric test where we adjusted for clustering of students in sections. We corroborated the results using Fisher's exact test.

TABLE 49. FREQUENCY OF VIDEO VIEWING BY *HMH FUSE* STUDENTS IN RIVERSIDE COMPARED TO *HMH FUSE* STUDENTS IN THE OTHER DISTRICTS

Frequency (Percent)	Median response across surveys		
	0-7	8 or more	Total
Non-Riverside	191 (79.58%)	49 (20.42%)	240
Riverside	61 (95.31%)	3 (4.69%)	64
Total	252	52	304

Amount of Time Teachers Spent Using iPad in Instruction

Table 50 shows a rank ordering of teachers by their reported average time in minutes per week spent using the *iPad* in instruction. We observe that the teachers in Riverside had the highest ranking.

TABLE 50. RANK ORDERING OF TIME IN MINUTES PER WEEK SPENT USING *HMH FUSE* IN INSTRUCTION

Teacher ID	District	Average time	Weeks with valid data
7	Non-Riverside	0	7
5	Non-Riverside	5.63	8
6	Non-Riverside	6.25	8
2	Non-Riverside	8.57	7
8	Non-Riverside	15.00	1
1	Non-Riverside	22.50	8
3	Non-Riverside	27.50	4
4	Non-Riverside	41.67	6
11	Non-Riverside	61.88	8
10	Riverside	73.75	8
9	Riverside	85.63	8

Amount of Time Students Spent Using iPad in Class.

Table 51 shows a rank ordering of teachers by their reported average time in minutes spent per week by their students using *HMH Fuse* in class. We observe that the teachers in Riverside were ranked second and third highest.

TABLE 51. RANK ORDERING OF MINUTES SPENT PER WEEK BY THEIR STUDENTS USING THE IPAD IN THE CLASS

Teacher ID	District	Average time	Weeks with valid data
2	Non-Riverside	8.57	7
7	Non-Riverside	18.57	7
1	Non-Riverside	30.00	8
5	Non-Riverside	32.50	8
8	Non-Riverside	40.00	1
6	Non-Riverside	41.88	8
4	Non-Riverside	50.00	6
3	Non-Riverside	67.50	4
10	Riverside	97.50	8
9	Riverside	104.38	8
11	Non-Riverside	167.50	8

Conclusion

Consistent with the observation that the Riverside school chosen to participate in the study was well experienced in implementation of new technologies, the usage as reported by the teachers was much greater than in most of the classes in other districts. While not entirely corroborated by the log data, greater amount of usage is a likely explanation for differential success. More detailed exploration of the data from this experiment can further refine this hypothesis and provide a strong basis for the next experimental test.

Discussion

Overview

The study, which took place in 7th and 8th grade algebra classrooms during the 2010-2011 school year, investigated whether *HMH Fuse* is effective at increasing algebra achievement and student attitudes toward math and whether impact varies for students with different characteristics. We also explored whether any impact on achievement is associated with certain mediating effects and whether the extent of use of *HMH Fuse* can be associated with any differences in algebra achievement. The study was conducted in four California school districts, and given the relatively small sample overall, our design aimed at establishing the average impact across the available units. An investigation of the differential impact between one district and the rest provided a strong indication that the average impact was misleading and has become the basis for the study's conclusion.

For this RCT, we randomly assigned one algebra period for each of the 11 participating teachers to the program condition, in which they use *HMH Fuse*. Each teacher's remaining algebra sections formed the control group assigned to use the regular text version of the Holt McDougal Algebra 1 2011 program. Our primary outcome measure for algebra achievement was the California Standards Test (CST). In addition, we used the End of Course Assessment. Student attitudes were measured by means of a Student Attitude Questionnaire consisting of two pre-existing measures. We also gathered implementation data via student and teacher surveys to inform outcome results.

Student Impact Results Averaged Across Districts

We found no impact of *HMH Fuse* on the primary measure of algebra achievement, the CST, on average across the four districts. We did find an impact on CST in one of the districts considered independently. We found no impact on the End of Course Assessment for the overall sample. There was an indication of a positive impact on student attitudes toward math as measured by the Student Attitude Questionnaire for the overall sample.

There was no moderating effect of pretest on any of the outcome measures. Specifically, the impact of *HMH Fuse* was not different depending on the student's pretest scores on the CST, End of Course Assessment, or Student Attitudes Questionnaire. However, for students who are not proficient in English we found a negative effect of *HMH Fuse* on the End of Course Assessment ($p = .08$), although not for the CST. We also found that although there was on average a positive effect on attitudes, English learners did not see any improvement on the Student Attitude Questionnaire as a result *HMH Fu*se.

We have some confidence in an impact of *HMH Fuse* on each of the three potential mediators: time spent on the algebra program outside the class, number of videos watched, and student attitude towards math. However, we found an indirect positive effect of *HMH Fuse* on student performance through only one mediator – student attitudes toward math.

Additional exploratory analyses provide descriptions of associations between iPad usage data and performance on the three main outcome measures.

Implementation Results Averaged Across Districts

Conditions for implementation were generally good across both groups; teachers received the necessary materials within the first few weeks of school, were mostly optimistic about *HMH Fuse,*

and reported feeling comfortable with the new program. However, many teachers reported technical difficulties with *HMH Fuse* that resulted in program students occasionally using the control curriculum, perhaps resulting in a watered-down program effect. Teachers and students also reported some instances of "contamination" in which those assigned to the control program interacted with *HMH* Fuse outside of class.

Teachers spent more time actively instructing with the control program than with *HMH Fuse*, although the difference between student use of either program in class was not significant. While both *HMH Fuse* and the control curriculum contain 300+ videos, which are deemed by the publisher to be a core component of the program, teachers reported watching very few videos in class. Student survey results and usage data from the iPads provide a rich description of how *HMH Fuse* was being used by students.

At the end of the school year, teachers reported high levels of satisfaction with both programs and would recommend *HMH Fuse* and the control curriculum to other algebra teachers. Nine of the eleven teachers would choose to continue teaching with *HMH Fuse* over the control curriculum. At the end of the study teachers reported feeling comfortable with *HMH Fuse,* although there was no change in their comfort levels since the beginning of the year.

Examination of District Differences

One of the school districts, Riverside Unified, initiated its own investigation of the data for the students that participated in both the *HMH Fuse* and the control classrooms. This work was conducted before Empirical Education had reported the overall results but found what appeared to be a strong impact (although an appropriate statistical test was not done). Because randomization was blocked by teacher, the evidence from that district alone constituted an RCT, albeit a very small one including only two teachers and nine sections of Algebra 1 randomized within teachers. We used the same statistical modeling approach to examine the subgroup impacts for the one district and for the other three. For the other three, and consistent with the overall results, there was no discernible difference between *HMH Fuse* and control. For Riverside, however, we found a substantial impact for which we can have strong confidence (p = .023). The adjusted effect size was 0.23, which is equivalent to a nine point increase in percentile standing. Since Riverside used percent proficient on the CST and our study used the CST scale score (as well as statistical adjustments), the results are not directly comparable. But our analysis does corroborate the results reported by Riverside for its participating teachers. It is also noteworthy that the teachers in that district reported more time instructing with *HMH Fuse* than reported by most of the other teachers in the study but that log data did not reflect more student usage.

Conclusion

After a one-year pilot implementation with *HMH Fuse,* we do not have evidence of a generalizable effect of the program on algebra achievement. We did find clear evidence that the effect was dependent on local conditions. For two teachers in one school—selected for the study on the basis of experience with technology innovations—there was an impact. Many characteristics of these teachers, their students, school, or district that can be put forward might explain the differential effectiveness of *HMH Fuse* in that district. The fact that the teachers reported using the application far more than other teachers is consistent with greater commitment to and experience with technology solutions. While we cannot generalize the results beyond these two teachers, the study is suggestive of approaches that may lead to success with applications such as *HMH Fuse*. It is notable

that there is a positive effect on student attitudes toward math, and students with positive attitudes toward math achieve higher scores on the CST.

At the end of the study, teachers handed in their iPads and discontinued implementation of *HMH Fuse*. Many questions remain to be answered including whether teachers may see an improvement in student outcomes in a second year of implementing *HMH Fuse*. The research indicates that under some conditions this new technology can have an impact. It remains to be seen whether these favorable conditions for implementation can be realized on a larger scale.

References

Amrein-Beardsley, A. L. (2006). Teacher research informing policy: An analysis of research on highly qualified teaching and NCLB. *Essays in Education*, 17

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

Bloom, H. S., Bos, J. M., & Lee, S. (1999). Using cluster random assignment to measure program impacts. *Evaluation Review, 23*, 445–469.

California Department of Education. (2010, March). *2009-2010 School Year Demographics*. Retrieved from http://dq.cde.ca.gov/dataquest/

California Department of Education. (2009). *Standardized Testing and Reporting (STAR) Program: Information for Parents.* Evanston, IL: Northwestern University. Retrieved from http://starsamplequestions.org/grades_9_11_math.pdf

Houghton Mifflin Harcourt (no date). HMH Fuse Algebra 1: Results of a yearlong Algebra pilot in Riverside, CA. Retrieved on March 5, 2012 from http://www.hmheducation.com/fuse/pdf/hmh-fuse-riverside-whitepaper.pdf

Krull, J. L., & MacKinnon, D. P. (2001). Multilevel modeling of individual and group level mediated effects. *Multivariate Behavioral Research, 36*, 249-277.

New America Foundation. (n.d.). *Federal Education Budget Project.* Retrieved from http://febp.newamerica.net/

Pintrich, P. R., & DeGroot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology, 82*(1), 33-40.

Puma, M. J., Olsen, R. B., Bell, S. H., & Price, C. (2009). *What to do when data are missing in group (cluster) randomized controlled trials.* (NCEE 2009-0049). Washington, DC: U.S. Department of Education.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods, Second Edition*. Newbury Park, CA: Sage.

Raudenbush, S. W., Spybrook, J., Liu, X., & Congdon, R. (2006). *Optimal design for longitudinal and multilevel research: Documentation for the "Optimal Design"* (Version 2.0) [Software]. Available from http://www.wtgrantfoundation.org/resources/overview/research_tools

SAS Institute. (2006). *SAS/STAT Software: Changes and Enhancements through Release 9.1.* Cary, NC: SAS Institute Inc.

Schochet, Peter Z. (2008). *Technical Methods Report: Guidelines for Multiple Testing in Impact Evaluations* (NCEE 2008-4018). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Shadish, W., Cook, T., & Campbell, D. (2002). *Experimental and Quasi-Experimental Designs For Generalized Causal Inference.* Boston, MA: Houghton Mifflin.

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics, 23,* 323-355.

State of California, Department of Finance. (2010, May). *January 2010 Cities and Counties Ranked by Size, Numeric, and Percent Change.* Sacramento, CA. Retrieved on January 21, 2011 from http://www.dof.ca.gov/research/demographic/reports/estimates/cities_ranked/2010/view.php

Tapia, M., & Marsh, G. E., II. (2002). *Confirmatory Factor Analysis of the Attitudes toward Mathematics Inventory*. Paper presented at the annual meeting of the Mid-South Educational Research Association in Chattanooga, TN on November 6-8, 2002.

The Center for Public Education. (2005, October 4). *Teacher quality and student achievement: Key lessons learned*. Retrieved December 1, 2008, from http://www.centerforpubliceducation.org/site/c.kjJXJ5MPIwE/b.1510943/k.9EF1/Teacher_quality_and_student_achievement_ Key_lessons_learned.htm

Tofighi, D. & MacKinnon, D. P. (2011). RMediation: An R package for mediation analysis confidence intervals. *Behavior Research Methods, 43,* 692-700.

What Works Clearinghouse. (2008*). Procedures and Standards Handbook (Version 2.1)*. Washington, DC: U.S. Department of Education.

Appendix A: Details of the Statistical Models

CST

TABLE A1. ESTIMATES OF FIXED EFFECTS FROM THE MULTILEVEL ANALYSIS OF THE IMPACT OF *HMH FUSE* ON CST

Fixed effects model	Estimate	Standard error	Degrees of freedom	*t* value	*p* value
Adjusted grand mean outcome on CST for controls	348.85	7.28	22	47.92	< .01
Effect of *HMH Fuse* intervention on performance on CST	2.56	3.88	22	0.66	.52
Effect associated with being in 7th grade (relative to 8th grade)	5.32	5.75	20	0.93	.37
Effect associated with being a student of teacher #1 (relative to teacher #11)	26.20	8.94	22	2.93	.01
Effect associated with being a student of teacher #2 (relative to teacher #11)	36.36	6.72	22	5.41	<.01
Effect associated with being a student of teacher #3 (relative to teacher #11)	-28.40	12.18	22	-2.33	.03
Effect associated with being a student of teacher #4 (relative to teacher #11)	0.02	14.89	22	0.00	1.00
Effect associated with being a student of teacher #5 (relative to teacher #11)	-10.59	5.60	22	-1.89	.07
Effect associated with being a student of teacher #6 (relative to teacher #11)	26.75	7.08	22	3.78	< .01
Effect associated with being a student of teacher #7 (relative to teacher #11)	-14.99	7.17	22	-2.09	.05
Effect associated with being a student of teacher #8 (relative to teacher #11)	-16.68	9.89	22	-1.69	.11
Effect associated with being a student of teacher #9 (relative to teacher #11)	-4.60	7.74	22	-0.59	.56
Effect associated with being a student of teacher #10 (relative to teacher #11)	12.73	8.35	22	1.52	.14
Effect associated with being male (relative to female)	-7.64	3.79	897	-2.02	.04
Effect associated with being a disabled student (relative to a nondisabled student)	-12.19	14.72	897	-0.83	.41
Effect associated with being designated as English proficient (relative to non-English proficient)	5.49	5.33	897	1.03	.30
Effect associated with being Asian (relative to White)	22.17	5.08	95	4.36	< .01

TABLE A1. ESTIMATES OF FIXED EFFECTS FROM THE MULTILEVEL ANALYSIS OF THE IMPACT OF *HMH FUSE* ON CST

Fixed effects model	Estimate	Standard error	Degrees of freedom	*t* value	*p* value
Effect associated with being Hispanic (relative to White)	2.37	4.09	95	0.58	.56
Effect associated with being Native Indian (relative to White)	44.27	62.64	95	0.71	.48
Effect associated with being designated of Mixed Ethnicity (relative to White)	14.20	11.02	95	1.29	.20
Effect associated with being Black (relative to White)	6.48	5.88	95	1.10	.27
Effect associated with dummy variable indicating missing value for ethnicity (relative to White)	27.76	26.04	95	1.07	.29
Effect associated with each unit increase on the pretest[a]	35.46	2.22	897	15.95	< .01
Effect associated with dummy variable indicating missing value for the pretest (relative to non-missing)	-12.67	16.75	897	-0.76	.45

[a] The dummy variable approach to handling missing data involves setting missing values for covariates to a constant. These effects are estimated with missing values set to zero; therefore, the effect estimates in this table should be interpreted accordingly.

TABLE A2. ESTIMATES OF RANDOM EFFECTS FROM THE BENCHMARK MULTILEVEL ANALYSIS OF THE IMPACT OF *HMH FUSE* ON CST

Random effects model	Estimate	Standard error	*z* value	*p* value
Variance component for sections	119.28	60.00	1.99	.02
Variance component for students within sections	2030.39	95.87	21.18	< .01

END OF COURSE

TABLE A3. ESTIMATES OF FIXED EFFECTS FROM THE MULTILEVEL ANALYSIS OF THE IMPACT OF *HMH FUSE* ON THE END OF COURSE ASSESSMENT

Fixed effects model	Estimate	Standard error	Degrees of freedom	*t* value	*p* value
Adjusted grand mean outcome on EOC for controls	18.91	1.11	14	16.99	< .01
Effect of *HMH Fuse* intervention on performance on EOC	0.11	0.84	14	0.13	.90
Effect associated with being in 7th grade (relative to 8th grade)	3.74	1.18	10	3.16	.01
Effect associated with being a student of teacher #1 (relative to teacher #11)	-4.77	1.41	14	-3.39	< .01
Effect associated with being a student of teacher #2 (relative to teacher #11)	-3.19	0.94	14	-3.40	< .01
Effect associated with being a student of teacher #3 (relative to teacher #11)	-3.20	2.01	14	-1.59	.13
Effect associated with being a student of teacher #4 (relative to teacher #11)	-2.11	1.93	14	-1.10	.29
Effect associated with being a student of teacher #5 (relative to teacher #11)	0.97	1.46	14	0.66	.52
Effect associated with being a student of teacher #6 (relative to teacher #11)	6.41	1.53	14	4.19	< .01
Effect associated with being a student of teacher #7 (relative to teacher #11)	-5.76	0.93	14	-6.18	< .01
Effect associated with being a student of teacher #8 (relative to teacher #11)	-9.50	1.39	14	-6.86	< .01
Effect associated with being male (relative to female)	-0.23	0.56	584	-0.41	.68
Effect associated with being a disabled student (relative to a nondisabled student)	-0.89	0.69	584	-1.29	.20
Effect associated with being designated as English proficient (relative to non-English Proficient)	0.98	0.56	584	1.76	.08
Effect associated with being Asian (relative to White)	1.62	0.91	57	1.78	.08
Effect associated with being Hispanic (relative to White)	-0.54	0.88	57	-0.61	.54
Effect associated with being designated of Mixed Ethnicity (relative to White)	-0.26	0.58	57	-0.45	.65

TABLE A3. ESTIMATES OF FIXED EFFECTS FROM THE MULTILEVEL ANALYSIS OF THE IMPACT OF *HMH FUSE* ON THE END OF COURSE ASSESSMENT

Fixed effects model	Estimate	Standard error	Degrees of freedom	*t* value	*p* value
Effect associated with being Black (relative to White)	-0.26	0.88	57	-0.29	.77
Effect associated with dummy variable indicating missing value for ethnicity (relative to White)	4.14	1.25	57	3.31	< .01
Effect associated with each unit increase on the pretest[a]	0.29	0.07	584	4.34	< .01
Effect associated with dummy variable indicating missing value for the pretest (relative to non-missing)	-1.10	0.80	584	-1.37	.17

[a] The dummy variable approach to handling missing data involves setting missing values for covariates to a constant. These effects are estimated with missing values set to zero; therefore, the effect estimates in this table should be interpreted accordingly.

TABLE A4. ESTIMATES OF RANDOM EFFECTS FROM THE BENCHMARK MULTILEVEL ANALYSIS OF THE IMPACT OF HMH FUSE ON END OF COURSE

Random effects model	Estimate	Standard error	*z* value	*p* value
Variance component for sections	3.96	1.84	2.15	.02
Variance component for students within sections	21.71	1.27	17.09	< .01

STUDENT QUESTIONNAIRE

TABLE A5. ESTIMATES OF FIXED EFFECTS FROM THE MULTILEVEL ANALYSIS OF THE IMPACT OF *HMH FUSE* ON THE STUDENT QUESTIONNAIRE

Fixed effects model	Estimate	Standard error	Degrees of freedom	*t* value	*p* value
Adjusted grand mean outcome on student questionnaire for controls	4.42	0.16	22	27.02	< .01
Effect of *HMH Fuse* intervention on outcome on student questionnaire	0.13	0.07	22	1.94	.07
Effect associated with being in 7th grade (relative to 8th grade)	0.11	0.09	20	1.15	.26
Effect associated with being a student of teacher #1 (relative to teacher #11)	0.19	0.14	22	1.39	.18
Effect associated with being a student of teacher #2 (relative to teacher #11)	0.07	0.22	22	0.34	.73
Effect associated with being a student of teacher #3 (relative to teacher #11)	-0.05	0.20	22	-0.24	.82
Effect associated with being a student of teacher #4 (relative to teacher #11)	0.27	0.16	22	1.68	.11
Effect associated with being a student of teacher #5 (relative to teacher #11)	-0.01	0.14	22	-0.10	.92
Effect associated with being a student of teacher #6 (relative to teacher #11)	0.24	0.12	22	2.03	.06
Effect associated with being a student of teacher #7 (relative to teacher #11)	0.07	0.12	22	0.57	.58
Effect associated with being a student of teacher #8 (relative to teacher #11)	-0.33	0.15	22	-2.21	.04
Effect associated with being a student of teacher #9 (relative to teacher #11)	0.26	0.12	22	2.16	.04
Effect associated with being a student of teacher #10 (relative to teacher #11)	0.53	0.15	22	3.47	< .01
Effect associated with being male (relative to female)	-0.02	0.06	759	-0.32	.75
Effect associated with being a disabled student (relative to a non-disabled student)	-0.19	0.15	759	-1.29	.20
Effect associated with being designated as English proficient (relative to non-English proficient)	0.13	0.10	759	1.30	.19
Effect associated with being Asian (relative to White)	0.22	0.09	83	2.57	.01

TABLE A5. ESTIMATES OF FIXED EFFECTS FROM THE MULTILEVEL ANALYSIS OF THE IMPACT OF *HMH FUSE* ON THE STUDENT QUESTIONNAIRE

Fixed effects model	Estimate	Standard error	Degrees of freedom	*t* value	*p* value
Effect associated with being Hispanic (relative to White)	-0.01	0.08	83	-0.10	.92
Effect associated with being Native Indian (relative to White)	-0.60	0.07	83	-9.24	< .01
Effect associated with being designated of Mixed Ethnicity (relative to White)	0.19	0.15	83	1.25	.22
Effect associated with being Black (relative to White)	0.11	0.10	83	1.10	.27
Effect associated with dummy variable indicating missing value for ethnicity (relative to White)	0.38	0.21	83	1.87	.07
Effect associated with each unit increase on the pretest[ab]	0.62	0.04	759	14.28	< .01
Effect associated with dummy variable indicating missing value for the pretest (relative to non-missing)	-0.05	0.10	759	-0.48	.63

[a] The dummy variable approach to handling missing data involves setting missing values for covariates to a constant. These effects are estimated with missing values set to zero; therefore, the effect estimates in this table should be interpreted accordingly.

[b] This is the residualized pretest. Since students' knowledge of their study condition had a potential biasing effect on the pretest questionnaire data, researchers conducted post hoc analytic corrections to factor out possible effects of assignment on the pretest scores. The timing of the administration of the pretest could not be helped. For instructional purposes, students were assigned to conditions prior to the start of the school year, and the questionnaire could not be administered until students were in class. Therefore, it was inevitable that students would already know their group assignment by the time of the administration of the questionnaire. It is plausible that students assigned to use *HMH Fuse* could show a positive change of attitude simply from the excitement of knowing that they would receive an iPad. In fact, the students in the *HMH Fuse* condition outperformed those in the control condition by a small but statistically significant margin on the pretest (p = .03). Post hoc analytic corrections do not assure the elimination of bias, though we felt that a different analysis may be preferable given the circumstances described. To do this, we regressed the pretest score on the indicator of assignment status and used the residuals from this regression, instead of the pretest, as a covariate in the analysis of the impact on the posttest.

TABLE A6. ESTIMATES OF RANDOM EFFECTS FROM THE BENCHMARK MULTILEVEL ANALYSIS OF THE IMPACT OF *HMH FUSE* ON STUDENT QUESTIONNAIRE

Random effects model	Estimate	Standard error	*z* value	*p* value
Variance component for sections	0.02	0.01	1.72	.04
Variance component for students within sections	0.46	0.02	19.51	< .01

TABLE A7. *HMH FUSE* SCORES BY CHAPTER

Assignment condition	Number of students	Number of questions possible	Average number of correct answers	Standard deviation	Standard error	Minimum	Maximum
Chapter 1							
Control	309	27	18.30	4.25	0.24	5	27
HMH Fuse	221	27	17.57	4.03	0.27	0	27
Total	**530**						
Chapter 2							
Control	515	25	16.37	4.58	0.20	4	25
HMH Fuse	308	25	16.26	4.29	0.24	4	25
Total	**823**						
Chapter 3							
Control	511	21	13.18	4.33	0.19	0	21
HMH Fuse	294	21	12.40	4.13	0.24	1	21
Total	**805**						
Chapter 4							
Control	325	16	11.49	2.57	0.14	0	16
HMH Fuse	156	16	10.73	3.01	0.24	0	16
Total	**481**						
Chapter 5							
Control	518	13	7.99	2.57	0.11	0	13
HMH Fuse	317	13	7.17	2.51	0.14	0	13
Total	**835**						
Chapter 6							
Control	522	13	8.96	2.34	0.10	0	13
HMH Fuse	314	13	8.48	2.46	0.14	0	13
Total	**836**						
Chapter 7							
Control	510	27	17.87	5.22	0.23	0	27
HMH Fuse	304	27	16.51	5.37	0.31	4	27
Total	**814**						
Chapter 8							
Control	478	23	14.81	4.35	0.20	2	23
HMH Fuse	300	23	13.61	4.60	0.27	2	23

TABLE A7. *HMH FUSE* SCORES BY CHAPTER

Assignment condition	Number of students	Number of questions possible	Average number of correct answers	Standard deviation	Standard error	Minimum	Maximum
Total	**778**						
Chapter 9							
Control	516	15	10.03	3.17	0.14	0	15
HMH Fuse	302	15	8.84	3.39	0.20	0	15
Total	**818**						
Chapter 10							
Control	453	20	10.38	3.79	0.18	0	19
HMH Fuse	234	20	9.05	3.72	0.24	1	18
Total	**687**						
Chapter 11							
Control	199	20	8.60	4.38	0.31	1	19
HMH Fuse	152	20	7.34	3.68	0.30	0	19
Total	**351**						

Note. Descriptive statistics reflect student-level data, as opposed to section-level data as reported in the mediator analysis.

Appendix B: Measures of Potential Mediators

Minutes Spent on Algebra Program Outside of Class

For the analysis of the impact of *HMH Fuse* on time spent on the algebra program outside of class, we considered two outcomes. First, because we were aware of a predominance of zero responses, especially among control students, we examined whether there was a difference between conditions in students responding zero, versus a value of more than zero, to this question. To receive a zero response for this analysis, a student would have to indicate zero minutes on each of the eleven surveys to which they responded, since we took the average of responses across surveys. The cross-tabulation of zero versus higher-than-zero responses is shown below. We have a high level of confidence that a higher proportion of *HMH Fuse* students than control students spent some time on algebra outside the classroom ($p < .01$), based on a multilevel logistic regression equation (with a section random effect).

TABLE B1. MINUTES SPENT ON ALGEBRA OUTSIDE OF CLASS

	None	Some	Total
Control	113 (19.96%)	453 (80.04%)	566
HMH Fuse	17 (5.54%)	290 (94.46%)	307
Total	130 (14.89%)	743 (85.11%)	873

We were also interested in whether there is an impact on the average number of minutes. Due to the predominance of zero responses, the outcomes were not normally distributed, and there was not a simple transformation that would produce a normally distributed response variable. We approached analysis in three ways.

1. We ran the analysis assuming a continuously distributed outcome and where we did not transform the outcome, with the assumption that with the larger sample of students (N = 873) violation of the normality assumption would not have a major effect on the statistical significance of the result. The regression-adjusted average difference was 17.9 additional minutes for the *HMH Fuse* group and we had some confidence of a real impact ($p = .10$).
2. We categorized the outcomes and analyzed them as ordinally distributed. The categories were zero minutes, more than zero but no more than 45 minutes per week, more than 45 minutes but no more than 90 minutes per week, more than 90 minutes but no more than 150 minutes per week, and more than 150 minutes per week. This approach did not work because the estimation process did not converge to yield a random effect estimate for sections, which is critical to judging the significance of the result.
3. We took the section means of responses across surveys and conducted a section-level analysis treating the outcome as continuous. (The section-average responses were close to normally distributed.) We found that the difference in average time spent was 18.4 minutes per week favoring the *HMH Fuse* group with a p value of .10.

Results (1) and (3) give convergent evidence of an impact. We have some confidence that the intervention has a positive effect on time spent using the algebra program outside the classroom.

With an impact on the posited mediating variable, we ran a mediation analysis using the results of (3). First we examined whether there was an association between the mediator – the section-average time spent on the algebra program outside the class – and student performance on the CST.. We expect a 0.08 point gain on the CST for each unit increase in the section-average of time spent on the algebra program outside the class. With $p = .18$, we have limited confidence in this result.

We used the software RMediation (Tofighi & MacKinnon, 2011) to obtain an estimated mediated effect of 1.511. The 80% confidence interval for the mediation effect is (-0.042, 3.528). Because the confidence interval spans zero we have no confidence that the section-level average of time on studying algebra outside the class mediates an effect of the program.

(Note: We weighted outcomes in the analyses involving minutes spent on the algebra program outside of class. We considered each survey occasion for each teacher to constitute a block, and we removed each block where there were no responses in either conditions [likely an indication that the teacher did not administer the survey on that occasion because the corresponding chapter was not taught] or where there was a complete imbalance between conditions in response – that is, where students responded only in one condition. This has the effect of weighting upward responses of students who answered more frequently, or, alternatively, weighting upward outcomes for teachers whose students, both program and control, responded on more survey occasions.)

Number of Algebra Videos Watched

The question gave five response options: (1) 0-7, (2) 8-15, (3) 16-23, (4) 24-31, and (5) 32 or more. Students were able to respond to this item on up to 11 survey occasions. To analyze this result we considered each student's median response over the 11 surveys. For example, if a student selected 0-7 on five surveys, 8-15 on one survey, and 16-23 on five surveys, her median response level would be 8-15. (Six of 866 students gave responses to an even number of surveys and picked two different middle responses. With very few such cases we included them in the higher category of the middle two.)

This categorization led to the following distribution of students' median responses.

TABLE B2. FREQUENCIES OF MEDIAN RESPONSES TO THE NUMBER OF ALGEBRA VIDEOS WATCHED PER WEEK

	0-7	8-15	16-23	24-31	32 or more	Total
Control	538 (95.73%)	19 (3.38%)	4 (0.71%)	0 (0.00%)	1 (0.18%)	562
HMH Fuse	252 (82.89%)	36 (11.84%)	12 (3.95%)	4 (1.32%)	0 (0.00%)	304
Total	790 (91.22%)	55 (6.35%)	16 (1.85%)	4 (0.46%)	1 (0.12%)	866 (100%)

Only five out of 562 control students and 16 out of 304 *HMH Fuse* students had responses in the top three categories. Therefore, we collapsed the top four categories. The distribution of median responses (0-7 versus higher) was different for the two conditions – a larger proportion of control than *HMH Fuse* cases selected the 0-7 category ($p < .01$). We used a multilevel logit model (section modeled as random) with blocking on teacher to estimate the impact.

Analysis of a mediating effect requires a non-zero impact on the mediator, and, with the analysis program that we are using (RMediation), that the mediator variable is normally distributed. The latter condition does not apply here; therefore we do not run a formal mediation analysis. However, we did investigate the association between the posited mediator—the median range of number of videos watched, 0-7 or above—and student performance. We have no confidence that the median level of the number of videos watched per week has a significant association with student performance on the CST ($p = .61$).

TABLE B3. FREQUENCIES OF MEDIAN RESPONSES TO NUMBER OF ALGEBRA VIDEOS WATCHED (COLLAPSED INTO TWO CATEGORIES)

	0-7	8 or more	Total
Control	538 (95.73%)	24 (4.27%)	562
HMH Fuse	252 (82.89%)	52 (17.11%)	304
Total	790 (91.22%)	76 (8.78%)	866

TABLE B4. FREQUENCY OF SELECTING 0-7 ALGEBRA VIDEOS

	Always 0-7	Sometimes 8 or more	Total
Control	438 (77.94%)	124 (22.06%)	562

HMH Fuse	161 (52.96%)	143 (47.04%)	304
Total	599 (69.17%)	267 (30.83%)	866

A response option of 0-7 did not allow us to assess how many students were consistently responding zero. However, as with the analysis of the impact on minutes spent doing algebra we were interested in the proportion of students consistently giving the lowest possible response (i.e., those who consistently selected 0-7 on all the surveys to which they responded). The difference in response rates for the two conditions is given in Table B4.

A larger proportion of control than *HMH Fuse* cases were consistently selecting the lowest response option
($p < .01$). We used a multilevel logit model (section modeled as random) to estimate the impact.