RESEARCHREPORT

AN EMPIRICAL EDUCATION REPORT

Assessing Impacts of Math in Focus, a 'Singapore Math' Program for American Schools

A REPORT OF FINDINGS FROM A RANDOMIZED CONTROL TRIAL

November 30, 2012

Andrew P. Jaciw
Whitney Hegseth
Boya Ma
Garrett Lai
Empirical Education Inc.

Empirical Education

Empirical Education Inc. · 425 Sherman Avenue, Ste 210, Palo Alto, CA 94306 · empiricaleducation.com · 650.328.1734

Acknowledgements

We are grateful to the people in Clark County 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.

Assessing Impacts of Math in Focus, a 'Singapore Math' Program for American Schools

A Report of Findings from a Randomized Control Trial

Reference for this report: Jaciw, A. P., Hegseth, W., Ma, B., & Lai, G. (2012, November). *Assessing Impacts of Math in Focus, a 'Singapore Math" Program for American Schools: Findings from an RCT.* (Empirical Education Rep. No. Empirical_MIF121130-FR1-Y1-O.1). Palo Alto, CA: Empirical Education Inc.

Table of Contents

Introduction

Houghton Mifflin Harcourt (HMH) contracted with Empirical Education Inc. to conduct a one-year randomized control trial (RCT) aimed at producing evidence of the effectiveness of *Math in Focus*™ (*MIF*) for third, fourth, and fifth grade students. We report here on the final results of this research that began in Clark County School District, Nevada, in August 2011.

The *Math in Focus* curriculum provides elementary math instruction based on the pedagogical approach used in Singapore, typified by a carefully sequenced and paced instructional style that focuses on fewer topics in greater depth at each grade level to ensure mastery. According to HMH, it is a "concrete to pictorial to abstract" (CPA) approach to instruction that is designed to support conceptual understanding. The instruction centers on problem solving using multiple models to help students visualize and understand math. HMH reports that the *MIF* curriculum is also closely aligned with the Common Core State Standards (CCSS).

"For over a decade, research studies of mathematics education in high-performing countries have pointed to the conclusion that the mathematics curriculum in the United States must become substantially more focused and coherent in order to improve mathematics achievement in this country. To deliver on the promise of common standards, the standards must address the problem of a curriculum that is 'a mile wide and an inch deep.' [The CCSS] are a substantial answer to that challenge" (CCSS Initiative, n.d.). HMH reports that *MIF* is closely aligned with the CCSS, which focuses more on in-depth learning than previous math standards did. The in-depth content provides for greater focus on math concepts and problem solving.

This difference between the CCSS-oriented *MIF* and the existing Nevada math standards and content lead us to specific expectations about where the major impacts will be seen. We expect *MIF* students to perform comparatively better than other students on achievement measures that emphasize depth, and comparatively worse than other students on achievement measures that emphasize breadth. Likewise, we expect *MIF* students to perform better on a test that measures complex problem solving skills, and not as well on a test that measures multiple procedural or computation skills. Because the Nevada state math standards have not fully shifted over to the CCSS, and because the state's assessment tests students on procedural skills as well as strategic thinking and problem solving skills, we do not expect a positive impact of *MIF* on student state test performance (State of Nevada Department of Education, n.d.).

We used three measures of math achievement: Stanford Achievement Test 10 (SAT 10), which has two sections: Problem Solving and Procedures, and Nevada's Criterion Referenced Test (CRT). Given the mapping of these tests to the characteristics of *MIF*, we address the following primary research question.

- Do students who belong to grade-level teams randomly assigned to be given the *MIF* curriculum and professional development perform differently on tests of math achievement than students in teams not randomly assigned to receive the *MIF* curriculum and training? More specifically we expect a positive impact on the

Problem Solving section of SAT 10. We expect a smaller or no impact on SAT 10 Procedures and the CRT.

We also address the following secondary questions.

- Is *MIF* differentially effective in its impact on student achievement depending on the minority status of the student?
- Does *MIF* lead to a decrease in the percentage of math standards covered, and if so, does this impact account for the effects of *MIF* on student math achievement?

In addition to addressing these questions, this study documents how *MIF* was implemented and reports on teacher satisfaction with the program.

For this experimental study, we worked with HMH to recruit 12 schools with grades 3, 4, and 5. For most schools, grades 4 and 5 were identified as one team to be randomized, and grade 3 formed the other team. A coin toss determined which randomized team would join the *MIF* group (the program group trained on *MIF*) and which would join the control group (the one receiving 'business as usual'). Technically, each school constituted a randomized block, with the two randomized teams (grades 4 and 5 in one team, and grade 3 in the other) forming a matched pair. For the schools that did not have a participating grade 3, we would randomize grades 4 and 5 into two different groups, one grade-level team would be randomized to the *MIF* group, and the other would be randomized to the control group. Altogether we randomized 22 grade-level teams, 12 of which were assigned to *MIF,* and 10 of which were assigned to control.

An RCT eliminates a variety of biases that could otherwise compromise the validity of the research. For example, it ensures that teachers in both groups were not selected on the basis of their interest in trying *MIF* and in their ability to take advantage of the new program. Random assignment to experimental conditions does not, however, assure that we can generalize the results beyond the district where the research was conducted. We designed our study to provide useful information that will support local decision making by taking into account the specifics of district characteristics and details of local implementation. The results are not applicable to school districts with practices and populations different from those in this experiment. This report provides a rich description of the conditions of the implementation to provide the reader with an understanding of the context for our findings.

Methods

Our experiment results in a comparison of outcomes for grade-level teams where *MIF* was in place and grade-level teams using the district's current variety of methods. The outcomes of interest are the student test scores in math on the SAT 10 Math Problem Solving and Procedures assessments as well as on Nevada's CRT.

This section details the methods used to assess, at a specific level of confidence, whether assignment to *MIF* results in an impact on outcomes, including achievement. We begin with a description and rationale for the experimental design and go on to describe the program, 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 *MIF*.

Experimental Design

There is always a level of uncertainty in our estimates of the effects of a program. The uncertainty can be understood in terms of the likelihood that we would obtain a different result if we selected a new study sample from a hypothetical pool of similar schools. It is important to recognize that the study results could change if we were to select a new sample. Technically, these random differences in results are attributed to sampling variation.

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. Technically, we are trying to limit the effect of sampling variation on our estimates.

The design of the experiment is based on our best estimation of the amount of variability in outcomes that is not attributable to the program, and we attempt to detect the stable signal (the effect) if it exists by dampening the random variation that obscures it to the extent possible.

Due to the challenges inherent in recruiting schools and the voluntary nature of any experimental study, the sample was largely one of convenience. The reader must be cautious in generalizing the results beyond the sample, taking into consideration the particular characteristics of the sample and other conditions of the study. Before beginning the experiment, we create a design or plan in which we establish the specific questions to be answered.

First, before seeing the results, we specify the research questions and identify the effects that we will analyze to address the questions. This includes average impacts, as well as differential and mediated effects of the program. 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 appear 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 large enough to be considered significant, but are attributable only to 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 an effect of a certain magnitude, or how big an effect we can detect given the sample sizes that are 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.

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 specified the best states for conducting the study and recommended several districts. In addition, Empirical Education called district contacts that had previously shown interest in participating in studies and also sent emails to all districts for which they had contact information and which contain elementary schools in the specified states. Although several districts showed some interest in participating in the study, in the end Clark County was the only district to agree to participate.

Once Clark County signed a district agreement, the district point of contact (POC) sent an email to all elementary school principals telling them to contact Empirical Education if they were interested in participating. Empirical invited interested principals to participate in one of two conference calls so they could learn about the study and ask questions. Researchers also met by telephone individually with principals who were not available at the time of the conference calls.

Once each principal had agreed to participate in the study, Empirical set up conference calls for their teachers to learn about the study and sent them participation packets containing teacher consent forms. Researchers urged teachers and principals to return signed teacher consent forms prior to randomization.

Randomization

We would like to determine whether *MIF* caused a difference in outcomes. To do so we have to isolate its effect from all the other factors influencing performance. Randomization ensures that, on average, characteristics other than the program that affect the outcome are evenly distributed between program and control groups. By evening out the effects of these factors between conditions, we arrive at an unbiased estimate of the program effect. Any remaining departures from the true values of the effects are due to chance differences between conditions and are not due to any systematic differences.

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 under teachers in the schools. 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 that can be implemented individually per teacher can use a teacher-level randomization.

For this experiment, we randomized intact grade-level teams which volunteered for participation to the *MIF* and control groups instead of randomizing teachers within grade-level teams. Randomizing whole teams allowed for collaboration within grades, which is regarded as an important component of *MIF*. Because groups, instead of students, were assigned to *MIF* or the control materials, this kind of experiment is often called a ‘group randomized trial.’

What Factors May Moderate the Impact of *MIF*?

The selected design allows us to measure the differential effectiveness of *MIF* across specific types or subgroups of students. The variables differentiating the students are ones that were measured before the experiment started, and that we had reason to believe would affect the magnitude of the effect of *MIF.* Technically, these are called potential moderators because they may moderate (increase or decrease) the impact of *MIF.* We measure the effect of the interaction between each potential moderator and the variable indicating assignment (i.e., to *MIF* or control); that is, we measure whether the effect of *MIF* changes across levels of each moderator.

For this study we compared the program’s impact by student minority status. We chose this moderator to assess whether the impact of *MIF* is different for groups traditionally underrepresented and underserved, such as minority students. We also examined the moderating effect of the pretest; that is, whether the impact of *MIF* varies depending on a student’s performance relative to average performance for the grade level.

What Factors May Mediate Between *MIF* and the Outcome?

We also identified variables that we believed would facilitate the effect of *MIF* on student outcomes. These are called ‘potential mediators’ because we hypothesize that they mediate the effects of *MIF* on student achievement. They are intermediate outcomes, measured in both conditions after the start of the experiment but prior to student posttests. That is, a mediator lies along the causal path between the point where we assign cases to the intervention or to the control group, and the point when we measure student performance after the intervention is over. *MIF* must have an impact on the mediator for that mediator to potentially facilitate impact on achievement. We usually think of a mediator as a factor in *how* the program has an impact. Based on the nature of the program, we identified process variables that were likely to facilitate the overall impact of the program. We first tested whether *MIF* caused a difference between the program and control group in these 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 of random assignment we are sure that a difference between conditions in the mediating process is an effect of *MIF*. However, because we don’t randomly assign cases to levels of the mediator, when we observe mediation, we cannot be sure whether the mediator is a cause of change in achievement or is a proxy for a mediating factor that is not identified.

In this experiment, we measured the percentage of Nevada state math standards covered throughout the school year as a proxy for whether teachers were taking a breadth or depth approach to math instruction. We chose this particular mediator because *MIF* takes a ‘depth-over-breadth’ approach to teaching and learning, meaning that *MIF* teachers may teach for deeper conceptual learning and cover fewer standards, which could lead to their students performing better than control students on the Problem Solving section of the SAT 10.

However, because *MIF* teachers may cover fewer standards, their students may perform worse than control students on the CRT. The goal is to explore the possibility that increased depth of coverage mediates a positive impact on problem solving while decreased breadth of coverage mediates a negative impact on a measure that tests a broader content domain (the CRT).

How Large a Sample Do We Need?

We conducted a power analysis to determine the number of grade-level teams that the experiment would need in order to say with specific levels of confidence that the program 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. As a hypothetical example, using percentile ranks as the measure of impact, we may predict that a program of this type can often move an average student 15 percentile points. As a practical matter for educators, however, an improvement as small as 10 percentile points may have value. The researcher may then set the smallest effect of interest to be 10 points or better. Thus, if the program makes less than a 10-point difference, the practical value will be no different from zero. It is necessary to decide in advance on this value as part of the power analysis because it determines the sample size. 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). Whatever the MDES for a study, it remains possible that effects exist that are smaller than the MDES but that we are unlikely to detect with the sample size available.

We designed this experiment to detect an effect of .25, measured in standardized effect size units. Given how recruitment went and with some attrition in the fall of 2011, based on the sample available for analysis where we hold constant the rest of the parameters from our original power analysis, the MDES increases to 0.28.[1]

How Much Variation is there between Grade-Level Teams?

When we randomize at the grade level but the outcome of interest is a test score of students associated with those grade-level teams, we pay special attention to the differences among

[1] Based on the results of this study, the achieved sample-based values of the MDES are .21, .11 and .18 for CRT, SAT 10 Problem Solving and SAT 10 Procedures, respectively. The team-level r-squared values are .85, .99 and .94, for CRT, SAT 10 Problem Solving and SAT 10 Procedures, respectively.

grade-level teams in student average scores. The greater the variation in the grade-level team averages of student scores, the more grade-level teams we need in the experiment to detect the impact of the program. This is because the extra variation among grade-level teams adds noise to our measurement which makes the effect of the program, the signal, harder to detect. A summary statistic that is important for the statistical power calculation is the intraclass correlation coefficient (ICC). Technically, it is the ratio of the variation in the grade-level team averages of students’ scores to the total variation in students’ scores. A larger ICC means between grade-level team differences in student posttest scores contribute more noise to our program effect estimate. A larger sample of grade-level teams is then needed to dampen the noise to acceptable levels. We select a value of the ICC before the beginning of the study.

It is possible that certain design strategies lower the ICC. For example, the process of randomizing teams within schools eliminates the variation across teams attributable to between-school differences, thereby effectively lowering the ICC. Because we do not have reliable estimates of the benefits of this strategy, we do not figure them into our power calculations; therefore, in the event that matching was successful, our power calculation can be considered conservative in its determination of the number of grade-level teams needed. (The ICC, like other parameters in the power calculation, reflects our best estimate of what these values are—largely based on compilations of results from other studies. It is not possible to get estimates of these parameters using data from the study at hand until after the study is over.)

For this experiment we assumed an intraclass correlation of .15.

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 likely to be associated with performance. These are called covariates because they are likely to co-vary with the outcome. By including covariates in the analysis we increase the precision of our effect estimates by accounting for some of the variation in the outcome; that is, by effectively dampening some of the noise so that the signal—the effect of *MIF*—becomes easier to detect. (Randomization assures that the covariates on average take the same value in both conditions; however, in any one trial they may be imbalanced by chance. Adjusting for the effects of this imbalance increases the precision of our estimate of the effect of *MIF*). Technically, a covariate-adjusted analysis is called an analysis of covariance (or ANCOVA). In our experiments, a student’s score on a pretest is almost always the covariate most closely associated with the outcome. Where possible we adjust for the effect of the pretest.

In the impact analyses in this report, we model other covariates in addition to the pretest. Because we do not have reliable information about how much additional variance these covariates account for, we do not figure their effects into the power analysis; therefore, in the event that covariates other than the pretest account for remaining variation in the posttest (i.e., after figuring in the effects of the pretest) our power calculation can be considered conservative in its determination of the number of grade-level teams needed.

In this experiment, we assumed a correlation between the pre- and posttests of .80. This is consistent with values observed in the Variance Almanac which provides a compendium of empirically-based values.

Are There Subgroups of Particular Interest?

Estimates of effects for subgroups have less precision than for the sample overall because subgroups constitute smaller samples. If a subgroup is a subsample of the units randomized (e.g., grade-level teams in certain schools) then this usually has more of an effect on precision than if the subgroup involves a subsample within units randomized (e.g., certain students within each team). In the current experiment, we examine whether the impact of *MIF* on the CRT and SAT 10 math subscales is moderated depending on a student's pretest performance and student's minority status.

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 this is in fact so (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 this is in fact so (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 the above factors into consideration, and with the number of grade-level teams that were available for this study, we estimated that the smallest standardized effect size that we can detect assuming conventional tolerances for drawing false-positive or false-negative conclusions is .28 standard deviation units. This is equivalent to an absolute difference of 11 percentile points for math 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 SAT 10 and found that student's score to be 11 percentile points higher (i.e., at the 61st percentile) or 11 percentile points lower (i.e., at the 39th percentile) than the median score for the control distribution. The sample size calculation was conducted using Optimal Design, a software program developed for this purpose (Raudenbush, Spybrook, Liu, & Congdon, 2006).

Twenty-two teams were randomized. Of these, 18 were available for analysis of impact on SAT 10 outcomes and 20 were available for the analysis of impact on CRT.

SITE DESCRIPTION

Clark County School District (CCSD) is located in Clark County, Nevada. Las Vegas is the county seat. The county's total population is 1,951,269 (U.S. Census Bureau, 2010). CCSD's operating budget was $2,145,000,000 in the 2010 - 2011 school year and the estimated per-pupil expenditure was $6,940, compared to the estimated national per student expenditure of $11,391 for 2010 - 2011 (Clark County School District Quick Facts, 2011). CCSD has a total enrollment of 307,059 students for 2010 - 2011 (NCES, n.d.). Table 1 provides information about the entire district including the schools that participated in the study.

TABLE 1. DEMOGRAPHICS OF CLARK COUNTY SCHOOL DISTRICT

Demographics	
Total schools	370
Total full-time equivalent teachers	15,472
Student-to-teacher ratio	20
Student population	307,059
English Language Learners	17%
White	32%
Black	12%
Hispanic	42%
Asian	7%
Pacific Islander	1%
American Indian/Native Alaskan	1%
Multi racial/No response	5%

Sources: http://nces.ed.gov/ccd/districtsearch/
http://www.nevadareportcard.com/profile/ethnicity.aspx?levelid=D&entityid=02&yearid=10-11

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

HOUGHTON MIFFLIN HARCOURT *MATH IN FOCUS*

The program consists of a Singapore math curriculum that relies on a CPA approach to teaching and learning math, as well as extensive training and professional development.

Training/Professional Development

Although the *MIF* curriculum consistently asks teachers and students to move through its Instructional Pathway[2] and use the CPA approach, trainers explained that the curriculum is intended to build teacher capacity by leaving many decisions up to the practitioner. Because

[2] The Instructional Pathway of *MIF* consists of the following sections: Teach/Learn, Guided Practice, Let's Practice, and Practice and Apply (student workbook).

teachers are expected to differentiate instruction and iterate between teaching and letting students solve problems on their own, professional development is integral to effective implementation of the curriculum.

HMH provided trainings to the Clark County *MIF* teachers in August, September, and November of 2011, as well as January and March of 2012.

The August training occurred the week before the start of instruction. It was offered at different times on three different days to accommodate different availability among teachers, with most sessions lasting 1.5 hours and one lasting 3 hours.[3] During these sessions, an HMH Account Manager and an HMH Associate Product Manager distributed some of the materials to the teachers and provided teachers with an introduction to the program and the Singapore philosophy and pedagogy.

The September and November trainings consisted of day-long, grade-level sessions, each conducted by one of two *Math in Focus* National Specialists with HMH. Both trainings followed a similar format. Teachers were first asked how they felt about the curriculum and, in November, how their instruction had changed since the previous session. The trainer then discussed the different elements of the Instructional Pathway. For each element, the trainer provided strategies for teaching and classroom management, as well as a rationale for the curriculum and pedagogy. Trainings ended with the trainer helping teachers plan how to teach difficult upcoming chapters, as well as modeling how to teach various problems using manipulatives and bar models.

The January and March trainings were catered very specifically to the Clark County *MIF* teachers’ needs. For the January training, a *Math in Focus* National Specialist with HMH met with a small group of teachers, all from the same grade level, to model a *MIF* lesson. After the planning the lesson as a group, the HMH Specialist modeled the lesson in one of the *MIF* classrooms. Each teacher was asked to watch a different group of students during the model lesson, and after it was over, teachers met with the Specialist to debrief. The March training consisted of day-long, grade-level sessions where teachers met with the same *Math in Focus* Specialist to 1) analyze student work in relation to Common Core math standards, 2) plan lessons so students develop the ability to visualize, make generalizations, and be exposed to a variety of math problems, and 3) plan out lessons and content to be covered before the CRT assessment and for the remainder of the year.

Math in Focus Materials

MIF is an elementary math curriculum that consists of both teacher and student materials. The program comes with the following print, manipulative, and digital components.

[3] The sessions were intended to take 3 hours but trainers usually finished in half the allotted time.

- Print Materials: Student Textbooks A and B, Student Workbooks A and B, Teacher's Edition A and B, Implementation Guide, Assessments, Enrichment A and B, Extra Practice A and B, Reteach A and B, School-to-Home Connections, Transition Guide, Place Value Mats
- Manipulatives: 30-Student Manipulative Kit (for classes that do not already have manipulatives), Core Manipulative Kit (for classes that already have a base of manipulatives)
- Digital Materials: Online Student Book eBook, Online Teacher's Edition eBook, Online Test Generator, Online Workbook Printable PDFs, Online Virtual Manipulatives, Online Transition Resource Map, Online Math Background Videos, Online Teacher Resource Blackline Masters, Interactive Whiteboard Lessons, Online Student Interactivities, Online Common Core Focus Lessons and Activities

Expectations of Implementation

HMH representatives described their expectations for *MIF* implementation during the teacher trainings as well as during a meeting with researchers on January 10, 2012. There was a strong emphasis on using the CPA approach and not skipping elements of the Instructional Pathway in the training sessions. One HMH trainer recognized that teachers would need a longer time to get through this curriculum than stated by the book because of the large gap in learning between previous curricula and standards and the *MIF* curriculum and Common Core State Standards. This HMH trainer, therefore, told teachers that they could take more time with the lessons than projected by the book, and go back to previous grade level materials to build a foundation for their students. Although the trainer said it was acceptable to supplement the curriculum with other materials and instructional strategies that align to the Singapore approach to build a foundation and number sense for their students, teachers were discouraged from doing too much of this.

In addition to the above, there were more general guidelines set by HMH representatives; they also listed three main stipulations of ideal implementation of the *MIF* curriculum during the meeting with researchers. Teachers should do all of the following;

Use *MIF* as their core math curriculum, which the representatives further defined as using *MIF* for at least 80% of the math instruction each day.

Follow the elements of the Instructional Pathway—specifically, the Teach/Learn, Guided Practice, and Let's Practice components, without skipping around too much.

Use the CPA approach to mathematical learning and problem solving in a way that was recommended by HMH representatives and was contingent on the grade level and the chapter being taught.

Control Materials

In response to Survey 1, control teachers reported the math materials they use in their classrooms. The curricula vary across and at times, within schools, with *Envisions* and

Investigations reported the most frequently. Following are the materials reported with the number of teachers reporting each item in parentheses (some teachers listed more than one).

- Envisions (27)
- Investigations (17)
- Scott Foresman (5)
- Pearson SuccessNet (2)
- Everyday Math (2)
- No set curriculum (1)

Schedule of Major Milestones

Table 2 lists the major milestones in this study and associated dates.

TABLE 2. RESEARCH MILESTONES

Date	Milestone
August 2011	Obtained district agreement
August 25, 30 & 31, 2011	Teacher introductory trainings
September 28, 29 & 30, 2011	Teacher grade-level trainings
September 2011	Initiation of monthly teacher surveys
October 2011	Collected district rosters, demographics, and CRT pretest data
October - December 2011	Participants administered SAT 10 student pretest
November 2011	Collected second wave of district rosters, demographics, and CRT pretest data
November 16, 17, & 18, 2011	Teacher grade-level trainings
January 17, 18, & 19, 2012	Teacher grade-level model lesson trainings
February 13 - 16, 2012	Classroom observations
March 6, 7, & 8, 2012	Teacher grade-level trainings
May 2012	Participants administered SAT 10 student posttest
May 15, 2012	Conducted focus groups with *MIF* teachers
July 2012	Collected additional posttest data from district

Data Sources and Collection

The data for this study were primarily provided by the school district and collected by Empirical Education. Clark County School District provided CRT assessment scores, student demographics, and class rosters. Participating teachers also administered SAT 10 assessments for the purposes of this study. Empirical Education collected implementation data by means of teacher background forms, training and classroom observations, a teacher focus group, and teacher and principal surveys. In addition, we have reviewed various program documents and materials.

Class Rosters and Demographic Data

We collected class rosters and student demographics from the Clark County School District in October and November 2011, and again in July 2012. These data are required for a series of analyses, including to check for balance on characteristics of the groups as formed through randomization, and to assess impact and differential impact, where it is necessary to identify membership of students in subgroups and randomized teams. Specifically, we asked the district to provide the following student data.

- Name and unique identifier
- Date of birth
- Grade
- Gender
- Ethnicity
- English proficiency status
- Disability status (whether or not student has a disability or is in special education, but not the specific condition)
- Socioeconomic status (as measured by eligibility for free or reduced-priced lunch)
- Classroom teacher name and unique identifier
- School name and unique identifier

All student and teacher data having individually identifying characteristics were stripped of such identifiers for analysis, and the data were stored using security procedures consistent with the provisions of the Family Educational Rights and Privacy Act (FERPA). This experiment falls within the protocol approved by Empirical Education's Institutional Review Board (IRB), Ethical and Independent Review Services. Under this protocol and following FERPA guidelines, student or parental permission was not necessary, nor was it was required by the school district.

Achievement Measures

We employed two outcome measures to determine whether *MIF* is effective at increasing math achievement of students in third, fourth, and fifth grade classes: the Stanford Achievement Test 10 and the Nevada Criterion-Referenced Test. Of these, the primary outcome measure is the SAT 10 Problem Solving subscale, because this measure tests for the in-depth conceptual knowledge that is prioritized by both the *MIF* program and the CCSS. The SAT 10 Problem Solving outcome is also the primary measure for which we expect *MIF* to have an impact.

The Stanford Achievement Test

Pearson produces the norm-referenced and standards-based Stanford Achievement Test, Tenth Edition—which researchers used as a pretest and posttest. According to Pearson, the

SAT 10 “measures content and processes adapted from the new National Council of Teachers of Mathematics Principles and Standards for School Mathematics (PSSM) and state standards including number sense and operations; patterns, relationships, and algebra; geometry and measurement; and data, statistics, and probability. Questions assess processes in communication and representation; estimation; mathematical connections; and reasoning and problem solving. Mathematics Problem Solving measures the skills and knowledge necessary to solve problems in mathematics. Mathematics Procedures measures the ability to apply the rules and methods of arithmetic to problems that require arithmetic solutions” (Pearson, n.d.). Researchers chose the SAT 10 because it is closely aligned with the Common Core Standards for math procedures and math problem solving.[4] Students receive a vertically equated, nationally normed scaled score between 392 and 801.

The Nevada Criterion-Referenced Test

The CRT is a standards-based assessment that functions as an indicator of student performance. The CRTs are developed by Nevada educators and content specialists and are then revised by WestEd (Nevada Proficiency Examination Program, 2012). The CRTs measure students' progress toward achieving Nevada's state-adopted academic content standards in reading, math, and science. Because the test is linked to Nevada’s standards, there is no national comparison. Students receive a scale score between 100 and 500. Based on this scale score, Nevada uses four performance levels to report student achievement on the CRTs: Exceeds Standard, Meets Standard, Approaches Standard, and Emergent/Developing. The CRTs are neither vertically aligned nor vertically scaled. Students take a math assessment in grades 3 through 8.

Testing Schedule

Teachers administered the SAT 10 pretest in the fall of the 2011 - 2012 school year, after randomization. Schools received the paper and pencil assessments for participating third grade teachers at the end of September/early October 2011. Teachers returned students’ completed scannable forms in the Federal Express envelopes provided to them by Empirical Education. The tests were sent to a scoring warehouse, where Pearson then translated the student test data into electronic form and transferred the data to Empirical Education. Teachers of fourth and fifth grade students administered the web-based assessment to their students in their school computer labs between October 11 and December 16, 2011.

The majority of the pretests (approximately 94%) were administered in October, although approximately 4% of the teachers administered the SAT 10 in November, and approximately 2% in December. All tests were untimed, though according to teacher survey data, teachers reported giving students anywhere between 40 and 250 minutes to complete the test.

[4] Information on Pearson’s alignment study of SAT 10 with Common Core Standards may be found at: http://www.pearsonassessments.com/hai/images/PDF/Stanford_10_Alignment_to_Common_Core_Standards.pdf

Teachers administered the SAT 10 posttest in May 2012. For the posttest, participating teachers of all three grades administered the web-based assessment to their students in their school computer labs. All tests were untimed, though the amount of time teachers gave students to complete the test varied within and between schools, and often depended on computer lab availability.

For the CRT, we use scores available from the previous year's spring testing (2010 - 2011 school year) as a pretest measure, and the scores from spring 2012 as the outcome measure. These data were collected directly from Clark County School District.

Program Implementation Measures

In addition to rosters, assessment, and demographic data, we also collected implementation data over the entire period of the experiment, beginning with the teacher trainings in August 2011 and ending with the academic calendar of the district in June 2012. Data collected through teacher background forms, training and classroom observations, teacher focus groups, and nine web-based teacher surveys (as well as two web-based principal surveys) are used to provide both descriptive and quantitative evidence of the implementation. Table 3 outlines the timeline of the major data collection phases.

TABLE 3. IMPLEMENTATION DATA COLLECTION SCHEDULE FOR THE *MIF* STUDY

Data collection elements	2011-2012 school year									
	Aug	Sept	Oct	Nov	Dec	Jan	Feb	Mar	April	May
Training observations	X	X		X		X		X		
Teacher surveys		X	X	X	X	X	X	X	X	X
Principal surveys					X					X
Classroom observations							X			
Focus group										X

Teacher Background Form

Prior to randomization and the initial training for the research study, teachers received a Participant Information Packet. This packet provided general information about the research study, data collection activities, and participant responsibilities, in addition to the teacher consent form. It also included a teacher background form for teachers to complete, providing researchers with information about their teaching history and contact information. We used this information to help describe the context of implementation and to assess balance between program and control on teacher background characteristics.

Teacher Training Observations

We observed the initial and all four subsequent teacher trainings in Clark County School District, and asked additional questions about each training through the teacher online surveys.

Classroom Observations

We conducted classrooms observations over a period of four days in February 2012. In general, classroom observation data was used to inform the description of the learning environment and instructional strategies employed by the teachers. These data helped further explain the data that we garnered from teacher survey responses.

We observed a purposive sample that contained fairly equal representation of teachers from the various schools (9 of the 11 participating schools), grade levels (3, 4, and 5), condition (*MIF* and control), and years of teaching experience (more or fewer than four years of teaching experience).

During these classroom visits, we observed the classroom context for instruction in both conditions. In the control classrooms, we documented the different curricula enacted across the classrooms and how the teachers carry out their normal math instruction, while in the *MIF* classrooms, we documented how teachers implement and how students use the *MIF* program.

Focus Groups

On May 15, 2012, we conducted one focus group with a stratified random sample of fourteen *MIF* teachers. The group consisted of teachers from various schools, grade levels, and with varied years of teaching experience. Teachers responded to questions that asked them to elaborate on their survey responses, characterize their classrooms, compare *MIF* to previous curricula they've used, and provide descriptions of their overall experience with *MIF* in their particular school contexts.

Teacher and Principal Surveys

Surveys were deployed to participating teachers from September 2011 through May 2012. Table 4 outlines the survey schedule and the response rates for the 35 control and 39 *MIF* teachers participating in the study. The response rates were extremely high, with an overall rate of 99.6% for all surveys combined. In addition to teacher surveys, two principal surveys were deployed in December 2011, and May 2012, per HMH's request.

TABLE 4. SURVEY PARTICIPATION RATES

		Response rates for active teachers		
Survey	Date	*MIF* group	Control group	Total
Survey 1	September 2011	100.0%	100.0%	100.0%
Survey 2	October 2011	100.0%	100.0%	100.0%
Survey 3	November 2011	100.0%	100.0%	100.0%
Survey 4	December 2011	100.0%	100.0%	100.0%
Survey 5	January 2012	100.0%	100.0%	100.0%
Survey 6	February 2012	97.4%	100.0%	98.7%
Survey 7	March 2012	97.4%	100.0%	98.7%
Survey 8	April 2012	100.0%	100.0%	100.0%
Survey 9	May 2012	97.4%	100.0%	98.7%
Total		99.1%	100.0%	99.6%

Note. Response rates were calculated for active teachers only. For those teachers who attrited from the study, we did not deploy surveys to them post-attrition, nor did we include them in the counts after they attrited.

We developed the survey questions to account for the various aspects of teacher and student actions associated with instruction and learning, including the extent of student exposure to material (opportunities to learn with the curriculum). For example, in order to characterize the average time teachers and students spend using manipulatives, we ask the same question across Surveys 2 through 9. These questions, together with other types of survey questions, allow us to draw inferences about the nature of math instruction in terms of specific practices in both control and *MIF* classrooms.

We report quantitative survey data using descriptive statistics and, where appropriate, we employ tests of significance to compare the results for the two conditions (*MIF* and control). Questions regarding percentage of math standards taught are used in mediator analyses.

Survey topics include, but are not limited to, the following.

- Teacher background
- Conditions for Math Instruction
- Extent of Program Implementation and Implementation Fidelity
- Comparison of Classroom Practices between *MIF* and Control Groups
- Teacher Satisfaction with *MIF*

Teacher Background

We collected the following teacher background data.

- Education level completed
- Credentials and certification

- Years of teaching experience

Conditions for Math Instruction

We constructed survey items specifically designed to provide the reader with an understanding of the conditions under which teachers implemented the *MIF* and control programs, which is critical to understanding the achievement results. We collected survey, observation, and focus group data on *MIF* teachers' experience with the *MIF* trainings, support, and materials. We also asked control teachers about the amount of math professional development they received, as well as the extent to which that math professional development prepared them to use their math programs in their classrooms. Finally, we asked *MIF* teachers the extent to which the trainings prepared them to use various aspects of the *MIF* program in their classrooms. Data from these questions will help HMH improve the way they train and deliver support to teachers who use the *MIF* program.

Extent of Program Implementation and Implementation Fidelity

We surveyed *MIF* teachers to see how many *MIF* chapters they covered with their students over the course of the school year. We also designed survey questions to gather data on the extent to which *MIF* teachers implemented the program with fidelity. According to HMH implementation experts, the main fidelity requirements for *MIF* teachers are that they: use *MIF* as their core math curriculum, follow the elements of the Instructional Pathway without skipping around too much, and employ the CPA approach to mathematical learning and problem solving. Therefore, we used survey questions to examine the extent to which the *MIF* teachers and students used only *MIF* as their core math curriculum. We also examined the extent to which teachers followed the elements of the Instructional Pathway and employed the CPA approach when teaching with *MIF.* If teachers were not implementing with fidelity, any potential impact of *MIF* might be watered down due to decreased use or misuse of the product.

Comparison of Classroom Practices between *MIF* and Control Groups

We also collected survey data to compare math teaching and learning in *MIF* and control classrooms. We collected data at multiple time points in order to compare the average classroom practices between the *MIF* and control groups. HMH stressed the importance of using manipulatives and transition materials with *MIF* students as they/their teachers transitioned into the *MIF* program. To determine whether *MIF* teachers were doing this significantly more than control teachers, surveys asked teachers in both conditions the number of minutes they spent in a given week using math manipulatives, and using transition materials. We also asked teachers in both conditions to report the amount of time they spend planning their math lessons, because *MIF* trainers emphasized that *MIF* took more planning and preparation than other, more scripted math programs. Finally, HMH representatives communicated the importance of both teacher collaboration and taking a depth over breadth approach to teaching mathematical concepts with *MIF*. We therefore surveyed both groups of teachers about whether they collaborated on math instruction with other math teachers, and whether they were covering all Nevada state math standards at the times designated by their district pacing calendar.

Teacher Satisfaction with *MIF*

Finally, since teacher satisfaction is an important factor in decisions regarding program adoption, we asked teachers about their initial and final impressions of *MIF,* as well as how well their (and their principals') personal beliefs aligned to the *MIF* approach that was presented during trainings. The final survey also asked *MIF* and control teachers whether they would choose to teach with *MIF*/their current math program, if given the option. Additionally, the May principal survey asked *MIF* principals if they would recommend *MIF* to other principals and their teachers.

Formation of the Experimental Groups

This section describes the study sample that we will use to assess the impact of *MIF*. We start with the baseline sample which consists of the participating grade-level teams that were randomly assigned to the *MIF* or control group and for which we have information. The sample for which outcomes are analyzed may be modified somewhat from baseline as a result of attrition or for other reasons that data become unavailable.

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, average prior achievement, and other section characteristics. In addition, because we randomized grade-level teams within blocks (schools), we can expect somewhat better balance than we would have if we hadn't randomized in this way. However, by chance (and because grade-level teams 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 teachers and students, looking in particular at whether these characteristics are balanced between the *MIF* and control groups.

In Table 5, we compare the composition of the control and *MIF* groups at the point we received the rosters (baseline sample). For each of the characteristics of this sample, we conducted a statistical test[5] to determine the probability 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 systematic differences between the groups that could lead to 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

[5] We used a t test that adjusted for clustering of students in randomized teams. The criterion for significance was set at ≤ .05.

than 5% of the time (the standard conventionally used to assess if an effect is statistically significant). We see that balance is achieved on the observed characteristics.

TABLE 5. CHARACTERISTICS OF THE STUDY SAMPLE BASED ON ROSTERS RECEIVED (BASELINE SAMPLE)

	Control	*MIF*	Less than 5% chance of seeing this much imbalance
Background characteristics			
Asian	66 (6%)	79 (7%)	No
Hispanic	474 (46%)	624 (52%)	
Indian	5 (<1%)	8 (1%)	
Mixed	61 (6%)	76 (6%)	
Black	108 (11%)	154 (13%)	
White	309 (30%)	268 (22%)	
Male	499 (49%)	609 (50%)	No
Receiving free or reduced lunch	572 (57%)	741 (63%)	No
Disabled students	133 (13%)	123 (10%)	No
Grade 3	231 (23%)	471 (39%)	No
Grade 4	312 (30%)	376 (31%)	
Grade 5	482 (47%)	363 (30%)	
Fewer than 4 years teaching experience	2 (5%)	8 (17%)	No
Fewer than 4 years elementary math teaching experience	3 (8%)	9 (20%)	No
CRT[a]	0.00	- 0.14	No
SAT 10 Problem Solving[a]	- 0.00	- 0.25	No
SAT 10 Procedures[a]	0.00	- 0.16	No

[a] z-transformed within grade using mean and standard deviation for controls

Analytical Samples

Since some grade-level teams, teachers, 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 grade-level teams—and their members 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 *MIF.*

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 grade-level teams that attrite from the control group on average have lower achievement than grade-level teams 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.

We report overall and differential attrition at the level of randomization and below. This allows calculation of potential for bias according to What Works Clearinghouse standards (WWC, 2008).

Sample Sizes, Attrition, and Equivalence Tests

Table 6 shows changes in the samples from the point at which the grade-level teams were randomized to the point at which the posttests (SAT 10 Problem Solving, SAT 10 Procedures, and CRT) were received. It is important to note that data collection processes were different and the amount of attrition was different for each measure.

TABLE 6. NUMBERS OF UNITS IN THE EXPERIMENTAL GROUPS AND ATTRITION OVER TIME

	Control				*MIF*			
Event	No. of schools	No. of teams	No. of teachers	No. of students	No. of schools	No. of teams	No. of teachers	No. of students
Randomization	10	10	41	n/a	12	12	52	n/a
(Loss prior to rosters)	(1)	(1)	(4)	n/a	(1)	(1)	(6)	n/a
Fall rosters received	9	9	37	1025	11	11	46	1210
SAT 10 Problem Solving Analytical sample								
(Loss due to lack of posttest)	0	0	(2)	(241)	(2)	(2)	(7)	(353)
Final count of units with SAT 10 Problem Solving[a]	9	9	35	784	9	9	39	857
SAT 10 Procedures Analytical sample								
(Loss due to lack of posttest)	0	0	(2)	(233)	(2)	(2)	(7)	(375)
Final count of units with SAT 10 Procedures[b]	9	9	35	792	9	9	39	835
CRT Analytical sample								
(Loss due to lack of posttest)	0	0	0	(84)	0	0	0	(84)
Final count of units with CRT posttest	9	9	37	941	11	11	46	1126

[a] Of the 241 control students without posttests, 57 were lost because of lack of responses from the two attrited teachers in that condition, and 184 were lost from teachers for whom we have responses for at least some other students; of the 353 *MIF* students without posttests, 168 were lost due to no outcomes from the two randomized teams, and 185 were lost from teachers for whom we have responses for at least some other students.

[b] Of the 233 control students without posttests, 57 were lost because of lack of responses from the two attrited teachers in that condition, and 176 were lost from teachers for whom we have responses for at least some other students; of the 375 *MIF* students without posttests, 168 were lost due to no outcomes from the two randomized teams, and 207 were lost from teachers for whom we have responses for at least some other students.

Note. In the above table, most schools are double counted because grade-level teams from both conditions are in most of the participating schools.

SAT 10 Problem Solving

Two teams, one from each condition, were lost after randomization but before student rosters were established. Two additional teams assigned to *MIF* did not provide any posttest scores. Since we do not include students without posttests in the analysis these teams are counted as part of team-level attrition.

We lost three out of 12 grade-level teams in the *MIF* group, and one out of 10 grade-level teams in the control group, resulting in overall attrition of 18.18% at the level of randomization. The rate of differential attrition at the level of randomization is the difference between the rate of attrition in the program group, which is 3/12 X 100% = 25%, and the rate of attrition in the control group, which is 1/10 X 100% = 10%. The difference is 15%.

In terms of student records, we did not obtain posttests for 594 of 2235 (26.58%) students on the initial rosters (i.e., 353 out of 1210 [29.17%] *MIF* students, and 241 out of 1025 [23.51%] control students). The rate of differential attrition at the student level is 5.66%. (Note that these rates include the loss of students resulting from not having any posttests for the four grade-level teams.) We used a randomized block design, with randomization of grade-level teams within schools. We lost one school in its entirety. Because randomization is conducted within blocks, this does not affect the statistical equivalence in the remaining intact blocks.

We examine the equivalence on background characteristics for the analytical sample.

TABLE 7. CHARACTERISTICS OF STUDY SAMPLE (ANALYTICAL SAMPLE FOR SAT 10 PROBLEM SOLVING)

	Control	*MIF*	Total	Less than 5% chance of seeing this much imbalance	Effect size
Background characteristics					
Asian	54 (7%)	64 (7%)	118 (7%)	No	0.05
Hispanic	355 (45%)	418 (49%)	773 (47%)		0.08
Indian[a]	3 (< 1%)	7 (1%)	10 (1%)		0.46
Mixed	48 (6%)	56 (7%)	104 (6%)		0.04
Black	76 (10%)	93 (11%)	169 (10%)		0.08
White	247 (32%)	219 (26%)	466 (28%)		- 0.18
Male	375 (48%)	430 (50%)	805 (49%)	No	0.06
Receiving free or reduced-price lunch	419 (55%)	479 (57%)	898 (56%)	No	0.07
Disabled students	96 (12%)	89 (10%)	185 (11%)	No	- 0.11
Grade 3	185 (24%)	330 (39%)	515 (31%)	No	0.43
Grade 4	231 (29%)	240 (28%)	471 (29%)		- 0.04
Grade 5	368 (47%)	287 (33%)	655 (40%)		- 0.34
Fewer than 4 years teaching	2	6	8	No	0.67

TABLE 7. CHARACTERISTICS OF STUDY SAMPLE (ANALYTICAL SAMPLE FOR SAT 10 PROBLEM SOLVING)

	Control	*MIF*	Total	Less than 5% chance of seeing this much imbalance	Effect size
experience[a]	(6%)	(15%)	(11%)		
Fewer than 4 years elementary math teaching experience[a]	3 (9%)	7 (18%)	10 (14%)	No	0.51
Recalibrated SAT 10 Problem Solving pretest[b]	0.04	- 0.17	- 0.07	No	- 0.20

[a] Given the low counts, there is a potential for bias in the results and they should be interpreted with caution.

[b] z-transformed within grade using mean and standard deviation for controls

Note. The effect size for the pretest is the mean difference between the *MIF* and control group in the pretest scores for students, expressed in units of the pooled within-group standard deviation of the pretest. The rest of the characteristics take on a value of zero or one, therefore we adopted the Cox index as the effect size measure. The Cox index is the natural logged odds ratio divided by the constant 1.65.

SAT 10 Procedures

The levels of attrition and differential attrition at the level of randomization are the same as they were for SAT 10 Problem Solving. In terms of student records, we did not obtain posttests for 608 of 2235 (27.20%) students on the initial rosters (i.e., 375 out of 1210 [31%] *MIF* students, and 233 out of 1025 [22.73%] control students). The rate of differential attrition at the student level is 8.27%. (Note that these rates include the loss of students resulting from not having any posttests for the four grade-level teams lost as a whole to attrition.)

We examine the equivalence on background characteristics for the analytical sample.

TABLE 8. CHARACTERISTICS OF STUDY SAMPLE (ANALYTICAL SAMPLE FOR SAT 10 PROCEDURES)

	Control	*MIF*	Total	Less than 5% chance of seeing this much imbalance	Effect size
Background characteristics					
Asian	54 (7%)	62 (7%)	116 (7%)	No	0.05
Hispanic	358 (45%)	398 (48%)	756 (46%)		0.06
Indian[a]	4 (1%)	6 (1%)	10 (1%)		0.21
Mixed	49 (6%)	56 (7%)	105 (6%)		0.05
Black	75 (9%)	94 (11%)	160 (10%)		0.12
White	251 (32%)	219 (26%)	470 (29%)		- 0.16
Male	379 (48%)	425 (51%)	804 (49%)	No	0.07
Receiving free or reduced-price lunch	424 (55%)	466 (57%)	890 (56%)	No	0.06
Disabled students	96 (12%)	90 (11%)	186 (11%)	No	- 0.08
Grade 3	184 (23%)	332 (40%)	516 (32%)	No	0.47
Grade 4	234 (30%)	239 (29%)	473 (29%)		- 0.03
Grade 5	374 (47%)	264 (32%)	638 (39%)		- 0.40
Fewer than 4 years teaching experience[a]	2 (6%)	6 (15%)	8 (11%)	No	0.67
Fewer than 4 years elementary math teaching experience[a]	3 (9%)	7 (18%)	10 (14%)	No	0.51
Recalibrated SAT 10 Procedures pretest[b]	0.04	- 0.08	- 0.02	No	- 0.11

[a] Given the low counts, there is a potential for bias in the results and they should be interpreted with caution.

[b] z-transformed within grade using mean and standard deviation for controls

Note. The effect size for the pretest is the mean difference between the *MIF* and control group in the pretest scores for students, expressed in units of the pooled within-group standard deviation of the pretest. The rest of the characteristics take on a value of zero or one, therefore we adopted the Cox index as the effect size measure. The Cox index is the natural logged odds ratio divided by the constant 1.65.

CRT

Two teams, one from each condition, were lost after randomization but before student rosters were established. Thus, we lost one of the 12 grade-level teams in the *MIF* group, and one of the 10 grade-level teams in the control group, resulting in overall attrition of 9.09% at the level of randomization. The rate of differential attrition at the level of randomization is the difference between the rate of attrition in the *MIF* group, which is 1/12 X 100% = 8.33%, and the rate of attrition in the control group, which is 1/10 X 100% = 10%. The difference is 1.67%.

In terms of student records, we did not obtain posttests for 168 of 2235 (7.52%) students (i.e., 84 out of 1025 [8.20%] *MIF* students, and 84 out of 1210 [6.94%] control students). The rate of differential attrition at the student level is 1.26%.

We examine the equivalence on background characteristics for the analytical sample.

TABLE 9. CHARACTERISTICS OF STUDY SAMPLE (ANALYTICAL SAMPLE FOR CRT)

	Control	*MIF*	Total	Less than 5% chance of seeing this much imbalance	Effect size
Background characteristics					
Asian	59 (6%)	74 (7%)	133 (6%)	No	0.03
Hispanic	445 (47%)	582 (52%)	1027 (50%)		0.11
Indian[a]	4 (< 1%)	8 (1%)	12 (1%)		0.31
Mixed	56 (6%)	73 (6%)	129 (6%)		0.05
Black	96 (10%)	137 (12%)	233 (11%)		0.12
White	280 (30%)	252 (22%)	532 (26%)		- 0.23
Male	456 (49%)	566 (50%)	1022 (49%)	No	0.04
Receiving free or reduced-price lunch	532 (58%)	691 (63%)	1223 (61%)	No	0.12
Disabled students	123 (13%)	120 (11%)	243 (12%)	No	- 0.14
Grade 3	213 (23%)	445 (40%)	658 (32%)	No	0.49
Grade 4	283 (30%)	348 (31%)	631 (31%)		0.02
Grade 5	445 (47%)	333 (30%)	778 (38%)		- 0.46
Fewer than 4 years teaching experience[a]	2 (5%)	8 (17%)	10 (12%)	No	0.79

TABLE 9. CHARACTERISTICS OF STUDY SAMPLE (ANALYTICAL SAMPLE FOR CRT)

	Control	*MIF*	Total	Less than 5% chance of seeing this much imbalance	Effect size
Fewer than 4 years elementary math teaching experience[a]	3 (8%)	9 (20%)	12 (14%)	No	0.61
Recalibrated CRT pretest[b]	- 0.00	- 0.12	- 0.07	No	- 0.11

[a] Given the low counts, there is a potential for bias in the results and they should be interpreted with caution.

[b] z-transformed within grade using mean and standard deviation for controls

Note. The effect size for the pretest is the mean difference between the *MIF* and control group in the pretest scores for students, expressed in units of the pooled within-group standard deviation of the pretest. The rest of the characteristics take on a value of zero or one, therefore we adopted the Cox index as the effect size measure. The Cox index is the natural logged odds ratio divided by the constant 1.65.

REPORTING ON THE IMPACT OF *MIF*

Setting Up the Statistical Equation[6]

We put our data for students, teachers, and grade levels into a system of statistical equations that allow us to obtain estimates of the effects of interest. The primary relationships of interest are the causal effects of the program on achievement as measured by the SAT 10 Problem Solving, SAT 10 Procedures, and CRT assessments. We use SAS PROC MIXED and PROC 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 estimates are different from zero.

Program Impact

The primary question for the experiment was whether, following the intervention, students in *MIF* classrooms had different scores than students in control classrooms on the SAT 10

[6] 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 represented 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 (*MIF* 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 (schools, teams or students). As a result of randomization, each covariate is distributed in the same way for both the *MIF* 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 moderating effect of interest.

Problem Solving subscale, the SAT 10 Procedures subscale, and the state test (CRT). To answer this question, we analyzed outcomes for the sample available at the end of the experiment; that is, for the cases who did not attrite. The randomization resulted in two groups that are statistically equivalent. One is assigned to *MIF* and the other to ‘business as usual.’ As a result, the average difference between the complete set of randomized groups on the posttest is an accurate measure of effect of being randomized to *MIF* or control 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, grade level, gender, disability status, social-economic status (in terms of whether or not students received free or reduced priced lunch), ethnicity and years of teaching experience at the teacher level. We also had to account for the fact that students are clustered by grade-level teams. We expect outcomes for students who are grouped together in grade-level teams to be dependent as a result of shared experiences. We had to represent this dependency in our equation in order to prevent artificially high confidence levels about the results. To do this, we modeled a team-level random effect as we describe further in the upcoming section, titled *Fixed and Random Effects*.[7] The impact analysis is conducted on the sample with posttests. As we saw in the section on attrition, not all teams randomized were analyzed.

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 a given 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 when levels of attrition are consistent with those observed in this study (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

[7] 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 schools and grades. 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 Appendix B for the results for the full model for the main analyses). 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.

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).

Covariates and Moderators at the Student Level

In addition to the variable indicating whether a team is assigned to *MIF* 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 *MIF* for different levels of the covariate. For example, we consider whether *MIF* is more effective for higher-performing students than for lower-performing students. We estimate this *difference* (between subgroups) *in the difference* (between *MIF* 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 *MIF* 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 teacher behaviors and student activities that can be measured during the experiment. Unlike the moderators, these are not pre-existing characteristics such as pretest scores or ethnicity. 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.[8] 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

[8] Technically, the estimate of a given mediated effect is the product of the effect of program on the mediator, times the effect of the mediator on the final response variable, normally student achievement, holding constant the program 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 program on the final outcome before adjusting for the effect of the mediator, and (2) the effect of program on the final outcome after adjusting for the effect of the mediator (Krull & MacKinnon, 2001).

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.

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 resampled 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.[9] 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 for the benchmark analyses 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.

[9] 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). If we consider our study sample to be drawn from a larger hypothetical population, the variation represents differences we would expect when resampling from that hypothetical population.

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

When we run the computations on the data, we produce several results: among them are effect sizes, the estimates of 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 *MIF* groups to inform the interpretation of student outcomes. We obtained data for this section through nine online teacher surveys, two online principal surveys, classroom and training observations, and a teacher focus group. We provide implementation results in the following categories.

- Conditions for Math Instruction
- Extent of Program Implementation and Implementation Fidelity
- Comparison of Classroom Practices between *MIF* and Control Groups
- Teacher Satisfaction with *MIF*

Conditions for Math Instruction

Here we provide a description of the conditions under which math instruction in *MIF* and control classrooms took place. Specifically, we present data on materials, the amount of professional development received by both groups of teachers, and the extent to which *MIF* trainings or math professional development prepared teachers to implement *MIF* or their current math program.

MIF Materials

HMH shipped materials to schools in August 2011, before the start of the school year. However, eight schools notified HMH or Empirical in early September that shipments were arriving at the wrong schools, at which point HMH made efforts to ship the correct materials to the correct schools. In response to the first survey, five teachers reported that they were still missing some *MIF* student workbooks and textbooks at the end of September. Additionally, one teacher reported in Survey 2 that she didn't have enough manipulatives at the end of October.

Professional Development

HMH provided trainings for all participating *MIF* teachers in August, September, and November of 2011, as well as in January and March of 2012. The trainings in September, November, and March were all six hours long, the January trainings were three hours long, and the August introductory trainings ranged from 1.5 to three hours. Training attendance rates were 90%, 92%, 97%, 92%, and 90%, respectively with all *MIF* teachers participating in at least one of the training sessions.

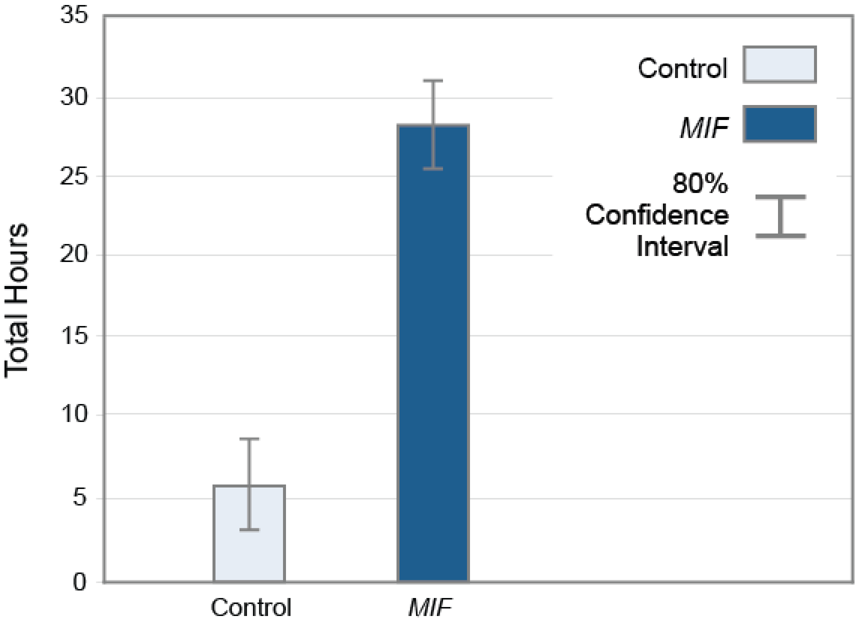

FIGURE 1. NUMBER OF HOURS SPENT ON PROFESSIONAL DEVELOPMENT

In Surveys 3, 6, and 8, we asked both *MIF* and control teachers the number of hours of math professional development (both *MIF* and other training) they had received. As displayed in Figure 1, *MIF* teachers reported receiving an average of 28 hours of professional development, as compared to an average of six hours among control teachers. With a *p* value smaller than .01, we have a high level of confidence that the observed difference in the amount of professional development can be attributed to *MIF*.

These same surveys asked about the extent the professional development prepared the respondents to use their math program in their classroom. One can see in Figure 2 that the majority of the control respondents (89% [*n* = 8] for Survey 3, 88% [*n* = 7] for Survey 6, and

86% [*n* = 6] for Survey 8) reported being moderately, more than moderately, or completely prepared to use their math program as a result of their professional development. In fact, between 13% and 14% of control respondents reported that they felt completely prepared to use their math program after their math professional development.

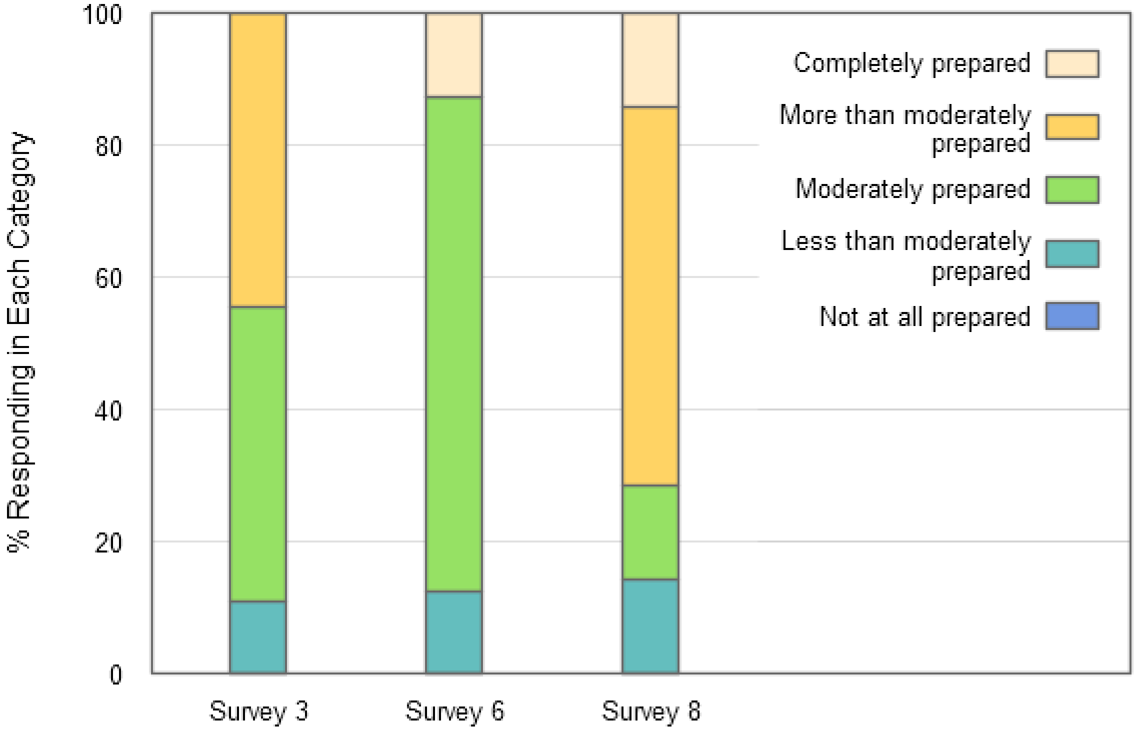

FIGURE 2. CONTROL TEACHER PREPARATION LEVELS FOR USING THEIR MATH PROGRAM

One recurring theme elicited by the teacher focus group was the importance of *MIF* trainings for implementing *MIF* successfully. An example of this theme was seen when participants were asked what type of teacher they would recommend *MIF* to. One participant first responded by saying “a teacher who is trained by HMH.” Four other participants went on to explain that any teacher could teach with *MIF*, permitting they have HMH’s support.

In Survey 1, after the August introductory training, we asked *MIF* teachers how well that training prepared them to use the *MIF* program in their classroom. In subsequent surveys, we asked *MIF* teachers more specifically about how well trainings prepared them to plan math lessons, use the CPA approach, use manipulatives during math class, facilitate activities that require students to use metacognitive reasoning, facilitate student collaboration, differentiate instruction, get students caught up to grade-level material, and facilitate teacher collaboration.

Among the 36 teachers responding to Survey 1, 50% (*n* = 18) reported feeling moderately prepared to teach with *MIF* after the August training, while 25% (*n* = 9) reported feeling less than moderately prepared, and 25% (*n* = 9) reported feeling not at all prepared to teach with *MIF*. No teachers reported feeling more than moderately or completely prepared by this initial training.

The following figures demonstrate that, apart from teachers’ general feelings of preparation after the August training, the majority of teachers reported feeling moderately prepared, more than moderately prepared, or completely prepared in response to each of the questions on every survey. For all but one of the questions (whether teachers felt prepared to facilitate activities that require students to use metacognitive reasoning), the proportion of teachers feeling moderately or more prepared was the lowest after the January training.

After the September training, all *MIF* teachers reported that their level of preparation was moderate or higher for planning math lessons, as can be seen in Figure 3. Although 20% (*n* = 7) of *MIF* teachers reported that they felt less than moderately or not at all prepared after the January training, and 3% (*n* = 1) and 6% (*n* = 2) of *MIF* teachers stated they felt less than moderately prepared after the November and March trainings, the majority of teachers continued to feel moderately or more prepared after every training.[10] In fact, between 12% and 27% of respondents reported that they felt completely prepared to plan math lessons after the trainings.

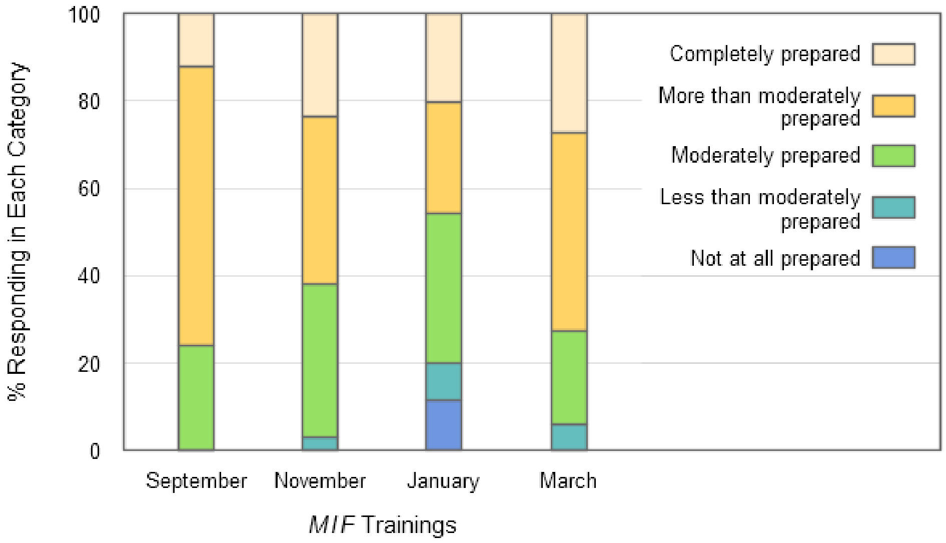

FIGURE 3. *MIF* TEACHER PREPARATION LEVELS FOR PLANNING *MIF* LESSONS

[10] Results presented in this report exclude data that fall outside of predetermined expected ranges. For questions that asked teachers to recall the extent to which specific activities occurred during the week the survey was sent to them, we excluded data where teachers responded two weeks or more after the survey was administered. Additionally, the numbers responding to specific questions may be different from the number of teachers in the study for the following reasons: 1) non-response to the survey 2) non-response to a given question 3) not applicable, for example, questions about perceptions about *MIF* training were only asked of teachers who stated that they had attended the training.

One can see in Figure 4 that, after the September and November trainings, 100% of the *MIF* teachers responding to the surveys reported being moderately, more than moderately, or completely prepared to use the CPA approach, which is a key component of *MIF*. Although 8% (*n* = 3) and 6% (*n* = 2) of *MIF* teachers reported feeling less than moderately prepared after the January and March trainings, respectively, and 8% (*n* = 3) of teachers reported feeling not at all prepared to use the CPA approach after the January training, the majority of teachers continued to feel moderately or more prepared after every training.[11] In fact, between 6% and 24% of respondents reported that they felt completely prepared to use the CPA approach after the trainings.

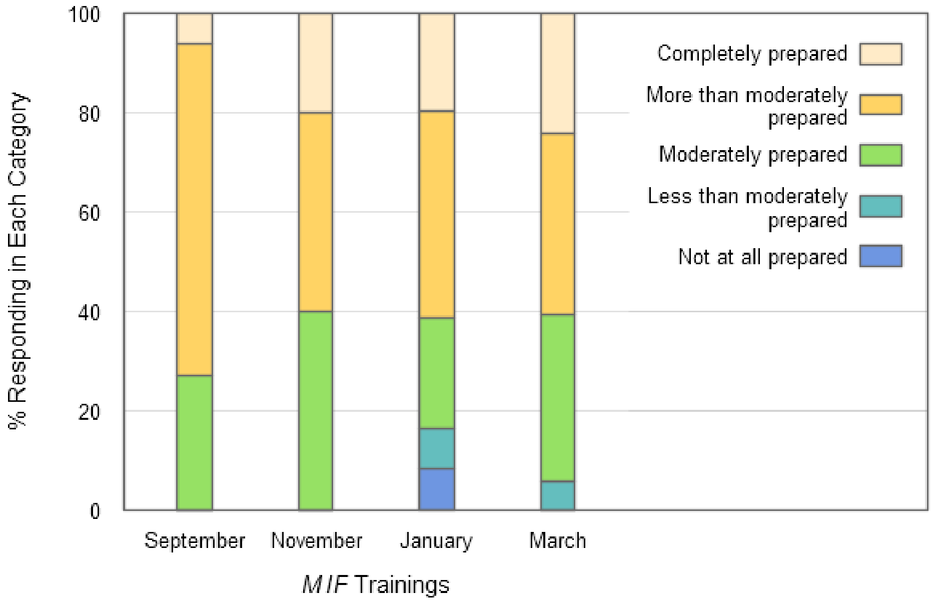

FIGURE 4. *MIF* TEACHER PREPARATION LEVELS FOR USING THE CPA APPROACH

Figure 5 demonstrates that the majority of *MIF* teachers (88% [*n* = 30] in September, 89% [*n* = 31] in November, 78% [*n* = 28] in January, and 91% [*n* = 30] in March) reported feeling moderately or more prepared to use manipulatives during math class after each of the four trainings. In fact, between 9% and 21% of respondents reported that they felt completely prepared to use manipulatives during math class after the trainings. However, between 8% and 9% of *MIF* teachers reported that they felt less than moderately prepared after all four trainings, and as many as 14% (*n* = 5) of *MIF* teachers stated they felt not at all prepared to use manipulatives after the January training.

[11] Due to the rounding of decimals, not all percentages add up to 100%.

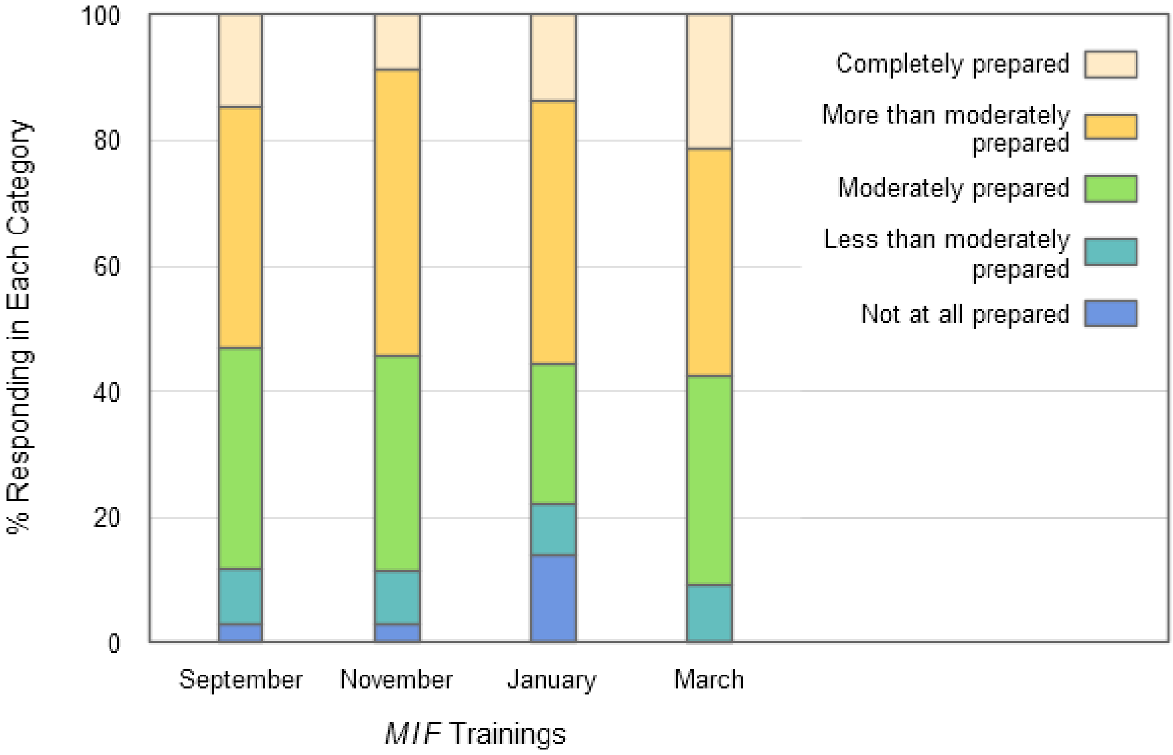

FIGURE 5. *MIF* TEACHER PREPARATION LEVELS FOR USING MANIPULATIVES DURING MATH CLASS

Figure 6 demonstrates that the majority of *MIF* teachers (94% [*n* = 32] in September, 89% [*n* = 31] in November, 92% [*n* = 33] in January, and 100% [*n* = 34] in March) reported feeling moderately or more prepared to facilitate activities that require their students to use metacognitive reasoning after each of the four trainings. Additionally, between 11% and 22% of respondents reported that they felt completely prepared after the trainings. However, between 3% and 11% of *MIF* teachers reported that they felt less than moderately prepared after the September, November, and January trainings, and 6% (*n* = 2) of *MIF* teachers stated they felt not at all prepared to facilitate activities that require their students to use metacognitive reasoning after the January training.

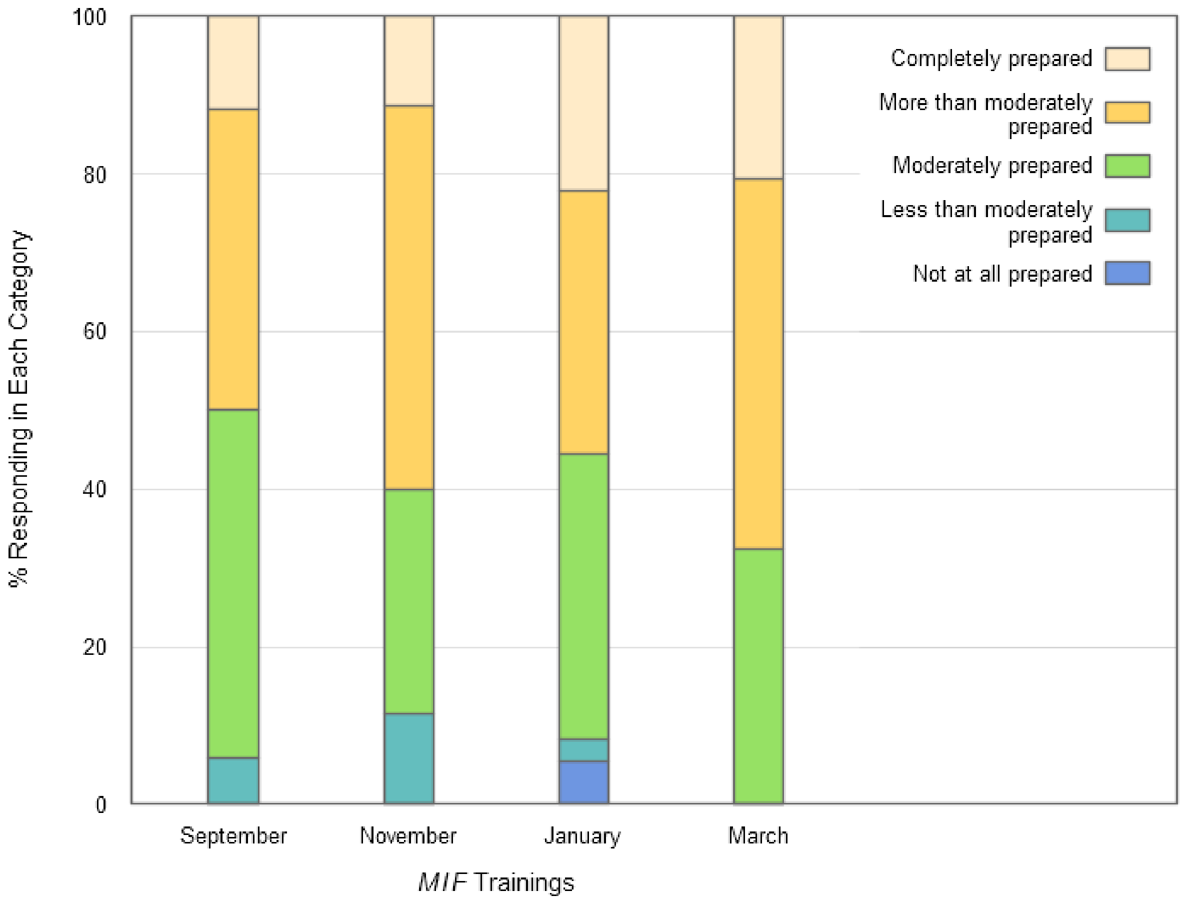

FIGURE 6. *MIF* TEACHER PREPARATION LEVELS FOR FACILITATING ACTIVITIES THAT REQUIRE STUDENTS TO USE METACOGNITIVE REASONING

Figure 7 demonstrates that the majority of *MIF* teachers (88% [*n* = 29] in September, 94% [*n* = 33] in November, 83% [*n* = 30] in January, and 97% [*n* = 33] in March) reported feeling moderately, more than moderately, or completely prepared to facilitate student collaboration after each of the four trainings. In fact, between 14% and 25% of respondents reported that they felt completely prepared after the trainings. However, between 3% and 12% of *MIF* teachers reported that they felt less than moderately prepared after all four trainings, and 6% (*n* = 2) of *MIF* teachers stated they felt not at all prepared to facilitate student collaboration after the January training.

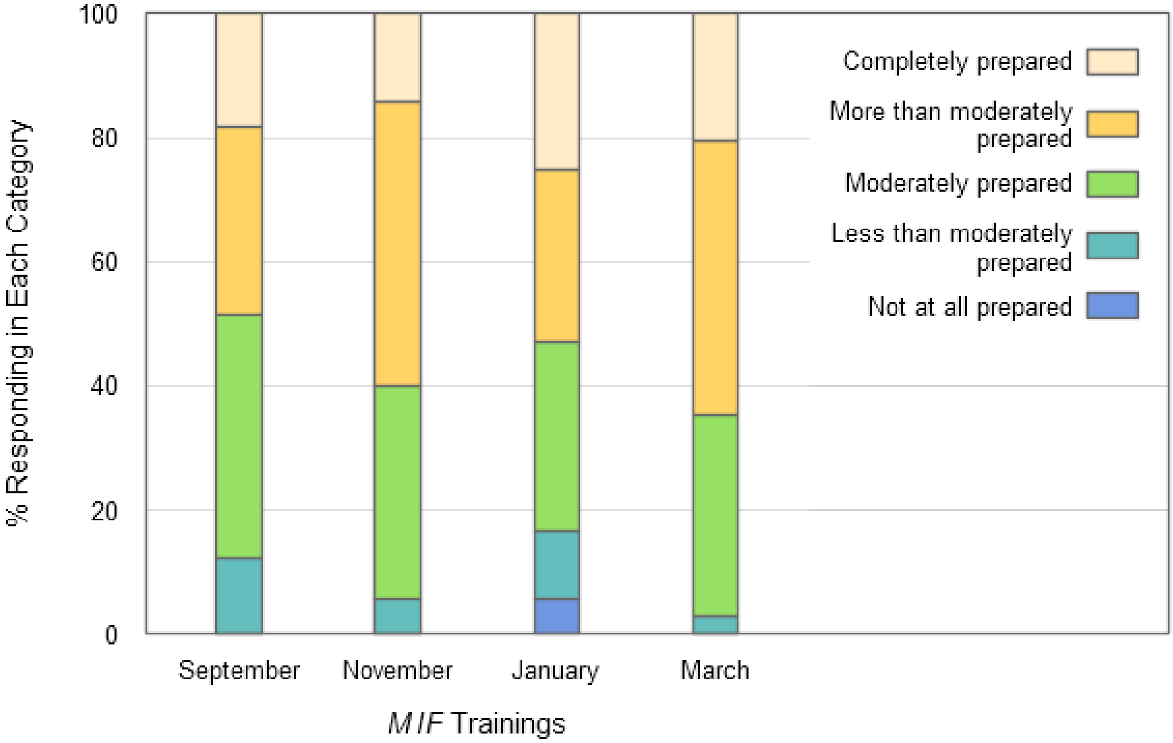

FIGURE 7. *MIF* TEACHER PREPARATION LEVELS FOR FACILITATING STUDENT COLLABORATION

Figure 8 demonstrates that the majority of *MIF* teachers (76% [*n* = 25] in September, 79% [*n* = 27] in November, 63% [*n* = 22] in January, and 88% [*n* = 30] in March) reported feeling moderately or more prepared to differentiate instruction after each of the four trainings. In fact, between 9% and 21% of respondents reported that they felt completely prepared after the trainings. However, between 12% and 26% of teachers reported feeling less than moderately prepared to differentiate instruction after the four trainings. Additionally, 3% (*n* = 1) and 11% (*n* = 4) of *MIF* teachers stated they felt not at all prepared to differentiate instruction after the November and January trainings, respectively.

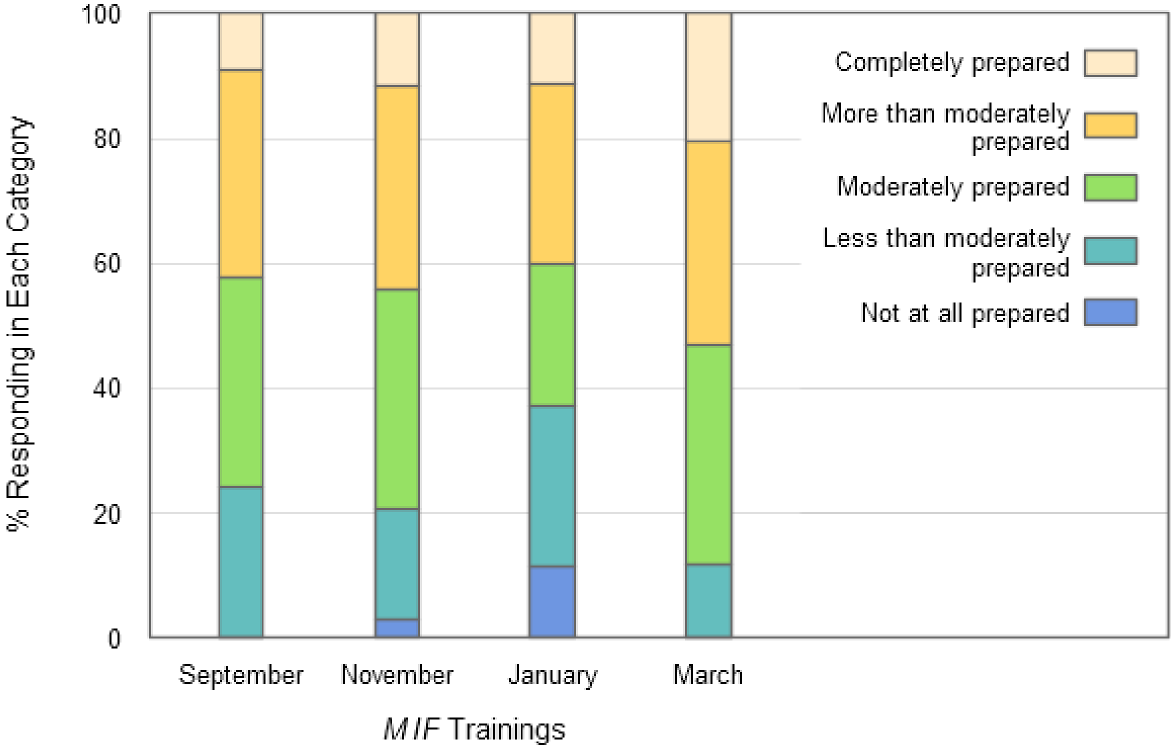

FIGURE 8. *MIF* TEACHER PREPARATION LEVELS FOR DIFFERENTIATING INSTRUCTION

Figure 9 demonstrates that the majority of *MIF* teachers (79% [*n* = 27] in September, 80% [*n* = 28] in November, 72% [*n* = 26] in January, and 91% [*n* = 31] in March) reported feeling moderately or more prepared to get students caught up to grade-level material after each of the four trainings. In fact, between 3% and 32% of respondents reported that they felt completely prepared. However, between 9% and 22% of teachers reported feeling less than moderately prepared to get students caught up to grade-level material after the four trainings. Additionally, 6% (*n* = 2) of *MIF* teachers stated they felt not at all prepared after both the November and January trainings.

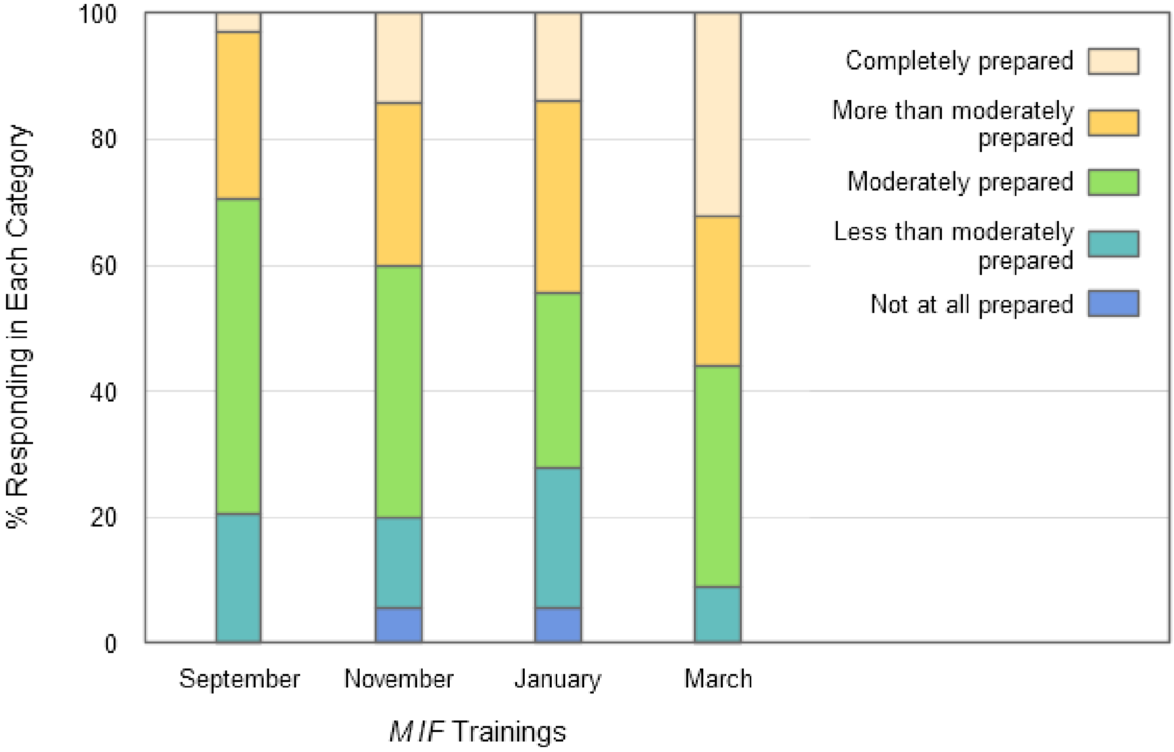

FIGURE 9. *MIF* TEACHER PREPARATION LEVELS FOR GETTING STUDENTS CAUGHT UP TO GRADE-LEVEL MATERIAL

Figure 10 illustrates that the majority of *MIF* teachers (94% [*n* = 32] in September, 94% [*n* = 33] in November, 92% [*n* = 33] in January, and 100% [*n* = 34] in March) reported feeling moderately or more prepared to facilitate teacher collaboration after each of the four trainings, as can be seen in Figure 10. There was an uptick in teachers' reports of feeling completely prepared to facilitate teacher collaboration as the school year progressed, with 9% of teachers (*n* = 3) reporting feeling completed prepared in September, then 14% (*n* = 5) in November, then 22% (*n* = 8) in January, then 38% (*n* = 13) in March. However, 6% of teachers (*n* = 2) reported feeling less than moderately prepared after the September and November trainings. Additionally, after the January training, 3% (*n* = 1) of teachers felt less than moderately prepared, and 6% (*n* = 2) of teachers stated they felt not at all prepared to facilitate teacher collaboration.

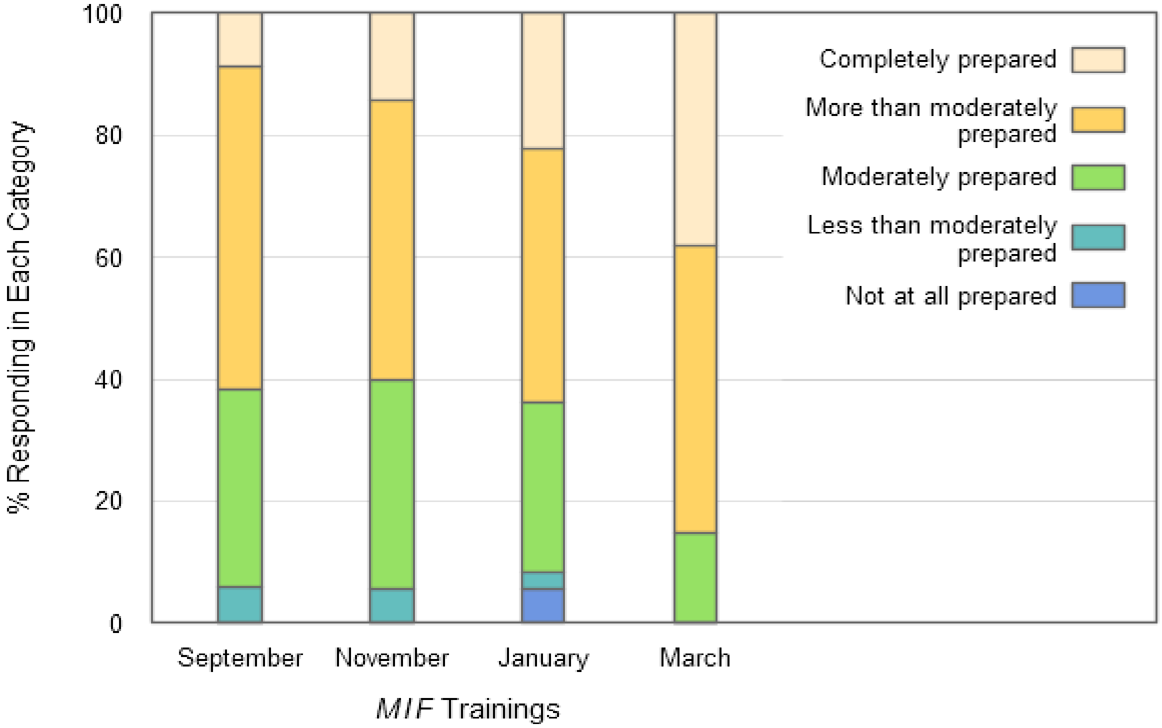

FIGURE 10. *MIF* TEACHER PREPARATION LEVELS FOR FACILITATING TEACHER COLLABORATION

Summary of the Conditions for Math Instruction

MIF teachers repeatedly stated in the various data collection activities throughout the year that the HMH trainings were necessary for implementing with *MIF*. Although most *MIF* teachers attended the introductory training in August, they reported in surveys, subsequent trainings, and the focus group that they considered themselves as being moderately, less than moderately, or not at all prepared for teaching with *MIF* up until the September training. Thus, most teachers did not feel adequately trained until the end of September. Similarly, some *MIF* teachers reported in surveys that they received materials as late as September or October, thereby delaying their ability to start implementing *MIF* in the way HMH intended. These data indicate that some teachers and students had less than a full school year’s worth of exposure to the *MIF* program and materials. However, teachers also reported feeling largely prepared to do many activities as a result of their *MIF* trainings, though they seemed to feel less prepared after the January training overall.[12] The January training was different from the other *MIF* trainings in that it was the only training where a *MIF* National Specialist with HMH met with a small group of teachers, all from the same grade level, to model a *MIF*

[12] Refer to Appendix A to see respondents’ preparation levels after each training—separated into the three different grade levels.

lesson. While control teachers received significantly less professional development than their *MIF* counterparts, they reported similarly mixed, yet positive views about the extent to which their professional development prepared them to use their math program in their classroom.

Extent of Program Implementation and Implementation Fidelity

In this section, we describe the extent of implementation of the *MIF* program, as well as implementation fidelity.

Extent of Implementation

In the September and November *MIF* trainings, an HMH trainer told teachers that no teacher is able to get through all *MIF* chapters during their first year teaching with it. The trainer explained that teaching and learning with *MIF* necessitated a 'mental shift,' and so it would initially take teachers longer to teach than the suggested pacing in the book.

In Survey 8, we asked *MIF* teachers to report the number of chapters they had taught. On average, third grade teachers taught 65% of the chapters in their *MIF* books, while fourth grade teachers taught 52% of their *MIF* book chapters, and fifth grade teachers taught 49% of their *MIF* book chapters.

Implementation Fidelity

Here we examine the extent to which teachers implemented *MIF* with fidelity. HMH representatives stipulated that for ideal implementation of the *MIF* program teachers should do all of the following.

1. Use *MIF* as their core math curriculum
2. Follow the elements of the Instructional Pathway without skipping around too much[13]
3. Use the CPA approach to mathematical learning and problem solving

Specifically, teachers implemented with fidelity if they met all of the following criteria.

- In response to the survey question asking the number of minutes they devote to *MIF* instruction each day, teachers reported using *MIF* at least 80% of their total instructional time for math.
- In response to the survey question regarding how frequently they incorporated each of the elements of the Instructional Pathway for any given chapter, teachers reported incorporating the 'Teach/Learn,' 'Guided Practice,' and 'Let's Practice' components of *MIF* frequently or always, and incorporating the 'Recall Prior Knowledge' and 'Pretest' components sometimes, frequently, or always (answer options were: never, seldom, sometimes, frequently, and always).

[13] The Instructional Pathway of *MIF* consists of the following sections: Teach/Learn, Guided Practice, Let's Practice, and Practice and Apply (student workbook).

- Teachers responded appropriately with never, seldom, sometimes, frequently, or always in response to the survey question asking how frequently they incorporated C, P, and A into their lessons.[14]

Only 21% of *MIF* teachers (*n* = 8) passed all three criteria that were set by HMH for fidelity of implementation. Following are the numbers that passed each separate criterion.

Using *MIF* as the Core Curriculum

As displayed in Table 10, thirty-eight percent of *MIF* teachers (*n* = 15) reported teaching *MIF* at least 80% of the time they devoted to math instruction in their classrooms. The average percentage of time across all nine surveys that *MIF* teachers devoted to teaching *MIF* was 73%.

Another recurring theme that emerged from training observations and the focus group was the pressure that teachers felt to supplement *MIF* with other programs. At least one to two teachers spoke of this pressure in the focus group as well as during each of the September, November, January, and March trainings, when they reported that they felt the need to use other math programs in addition to *MIF.* The primary reason for supplementing *MIF* with other programs was because *MIF* was not as aligned to the district pacing calendar or to the breadth of standards that the state test would be testing students on. Whenever a teacher brought up the need to supplement *MIF* as a result of this pressure caused by the discrepancy between *MIF* pacing and district pacing, a few other teachers nodded to affirm their agreement.

Following the Elements of the Instructional Pathway

Table 10 displays that 82% of teachers (*n* = 32) reported implementing with fidelity in terms of incorporating elements of the Instructional Pathway. When asked about how frequently they used the Instructional Pathway in Surveys 4 and 7, those 32 teachers reported teaching the ‘Teach/Learn,’ ‘Guided Practice,’ and ‘Let's Practice’ components of *MIF* frequently or always, and teaching the ‘Recall Prior Knowledge’ and ‘Pretest’ components sometimes, frequently, or always (answer options were: never, seldom, sometimes, frequently, and always).

[14] HMH gave Empirical Education a matrix of acceptable survey responses depending on a teacher’s grade level and the chapter they were teaching at the time of the survey.

TABLE 10. PERCENTAGE OF *MIF* TEACHERS WHO MET HMH FIDELITY CRITERIA

Survey question	Met criteria	Did not meet criteria
How many minutes did your class spend doing math (total and *MIF* specifically) this week? (*n* = 39)	15 (38%)	24 (62%)
For each chapter, how often do you complete each element of the instructional pathway when prescribed by the teacher's edition? (*n* = 39)	32 (82%)	7 (18%)
For the last chapter you completed, how often did you incorporate each of the following components into your math lesson? Concrete, Pictorial, Abstract (*n* = 34)	22 (65%)	12 (35%)

Using the CPA Approach

As displayed in Table 10, 65% of teachers (*n* = 22) met the criterion for using the CPA approach. In all four grade-level trainings, at least one to three teachers either reported or demonstrated a lack of understanding for how or when to use the concrete or pictorial representations. Similarly, in observations and the focus group, between one and three teachers explained this lack of understanding by stating that they only learned math in an abstract way when they were students, so it was difficult for them (and for their older students) to alter their former abstract approach and instead use concrete and pictorial representations to solve problems.

Summary of Extent of Program Implementation and Implementation Fidelity

The criterion that proved most difficult for teachers to meet was using *MIF* at least 80% of their math instruction time. This deficit in fidelity may be related to teachers' reported feelings of being pressured to supplement *MIF* with other math programs. The second most difficult criterion to pass was employing the CPA approach. Overall, implementation was less than ideal according to HMH's criteria, and this limitation in fidelity of implementation seems to be related to conflicts between *MIF* and district pacing, as well as between the CPA approach and teachers' former approaches to teaching and learning math in a more abstract, algorithmic manner.

Comparison of Classroom Practices between *MIF* and Control Groups

We asked both control and *MIF* teachers about instruction and planning practices to determine whether expected changes were taking place in *MIF* classes. Here we describe the survey data comparing these practices in *MIF* and control classrooms. Specifically, we compare *MIF* and control teachers' use of manipulatives and transition materials, as well as time spent planning math instruction, the types of discussions teachers had with other math teachers, and whether teachers were teaching Nevada math standards at the designated time.

In each of Surveys 2 through 9 we posed questions regarding the amount of time spent using math manipulatives (defined as any concrete materials such as base ten blocks and place value chips) during specified weeks. As shown in Table 11, on average, *MIF* teachers reported spending 83 minutes per week using manipulatives, compared to 73 minutes among control teachers. With a *p* value of .86, we have no confidence in there being a real difference between conditions.

TABLE 11. WEEKLY MINUTES OF MANIPULATIVE USE

	Mean weekly minutes
Control (*n* = 35)	73
MIF (*n* = 39)	83
p value	.86

We also posed repeated questions regarding the number of minutes teachers spent teaching with transition materials (transition was defined as using materials from prior grade levels). On average, *MIF* teachers reported using transition materials for 53 minutes during specified weeks, as compared to 33 minutes among control teachers, as can be seen in Table 12. A test of a difference in minutes of transition materials use yielded a *p* value of .18, giving us limited confidence that the result we observe reflects a real difference beyond chance.

TABLE 12. WEEKLY MINUTES OF TRANSITION MATERIALS USE

	Mean weekly minutes
Control (*n* =35)	33
MIF (*n* = 39)	53
p value	.18

TABLE 13. WEEKLY HOURS PLANNING MATH INSTRUCTION

	Median weekly hours
Control (*n* = 35)	2.9
MIF (*n* = 39)	3.0
p value	.71

In Survey 3, 6, and 9 we asked teachers to report the number of hours they spent during a specified week planning for math instruction. The median number of hours *MIF* teachers reported planning math instruction was three hours per week, as compared to 2.9 hours among control teachers, as can be seen in Table 10. With a *p* value of .71, we have no confidence that the result reflects a real difference beyond chance.

In Surveys 3, 5, and 7, we asked *MIF* and control teachers whether they had certain types of discussions regarding *MIF*/math instruction during a specified week. There was no difference between *MIF* and control teachers in their discussions regarding: reviewing student work (p = .68), discussing problematic lessons (p = .42), discussing what helps students learn best (p = .69), sharing successful strategies for lesson implementation (p = .67), and discussing resources for *MIF*/math instruction (p = .69). We have limited confidence in there being a difference between conditions in discussions labeled as ‘other’ (p = .18).

We asked teachers in Survey 7 whether they were teaching all of the Nevada state math standards, and whether they were teaching standards at the time designated by their district pacing calendar. As can be seen in Table 14, 79% of *MIF* teachers (n = 30) reported teaching all of the Nevada state math standards, as compared to 77% of control teachers (n = 27). The high p value gives us no confidence that this reflects a real difference beyond chance.

In each of the training observations and the focus group, *MIF* teachers reported that they felt the need to cover all Nevada state standards to prepare students for the CRT assessment. However, only 29% of *MIF* teachers (n = 11) reported teaching state math standards at the designated time, as compared to 80% of control teachers (n = 28). With p < .01, we have a high level of confidence that this result reflects a real difference. Additionally, in the principal survey, three out of nine, or 33% of the principals responding to the survey, indicated that their *MIF* teachers were not teaching the Nevada state math standards at the time designated by their district pacing calendar, as compared to eight out of eight principals (100%) reporting that their control teachers were teaching the state math standards at the time designated by the district pacing calendar.

TABLE 14. TEACHING NEVADA STATE MATH STANDARDS AT THE DESIGNATED TIME

	Teaching all standards	Teaching standards at designated time
Control	27 (77%)	28 (80%)
MIF	30 (79%)	11 (29%)
p value	.88	< .01

Note. One *MIF* teacher’s response is missing from these calculations.

Summary of Classroom Practices of *MIF* and Control Groups

Though HMH representatives expected there to be certain changes in classroom practices when teaching with *MIF*, and though teachers reported in observations and the focus group that they changed their teaching as a result of *MIF*, the only significant differences that we found were as follows.

- *MIF* teachers spent significantly more time teaching with transition materials.
- More control teachers taught the Nevada math state standards at the designated time. Results from the principal survey support the idea that some *MIF* teachers may not

have necessarily felt pressured to cover those standards at the time suggested by the district’s pacing calendar, because they were perhaps teaching the standards in line with when *MIF* was suggesting those standards be taught.

We did not find differences between the conditions in terms of the amount of time students spent using manipulatives, the amount of time teachers spent planning math instruction for their classes, or in types of teacher discussions surrounding math instruction. We also did not find any difference between the conditions in terms of the amount of teachers who taught all of the Nevada state math standards.

In the focus group, the majority of *MIF* teachers reported that, with *MIF,* they spent more time using manipulatives and transition materials, preparing math lessons, and collaborating with other teachers than they did with their previous math programs. Survey results support the reported changes in use of transition materials, but not the other reported changes. In classroom observations of control teachers, we noticed that some math programs and strategies that were being used by control teachers had similar components as those found in *MIF,* such as manipulative use and questioning strategies. It is perhaps due to this that there are no significant differences between *MIF* and control teachers in some aspects of classroom practices.

Teacher Satisfaction with *MIF*

In this section, we first present survey data that compares teachers’ initial and final impressions of *MIF,* as well as teachers’ and principals’ initial and final views on how well their personal beliefs align to the *MIF* approach that was presented in *MIF* trainings. Finally, we compare *MIF* and control teachers’ responses on whether they would choose to teach with *MIF*/their current math program, and we report on how many *MIF* principals said they would recommend *MIF* to other principals and to their teachers.

On Survey 1, after the introductory training, we asked *MIF* teachers about their initial impressions of *MIF*.[15] The respondents indicated mixed expectations for the program, but the largest proportion (49%, *n* = 18) reported that their initial impression was, “Good. It has possibilities,” as depicted in Figure 11. When asked about their final impression of *MIF* in the final survey, deployed in May 2012, the largest proportion of respondents again selected “Good,” however, more teachers responded with “Great! I’m glad I taught with *Math in Focus!*” (43%, *n* = 15), and fewer responded with, “OK. It’s just another math program,” (6%, *n* = 2) or “Skeptical. Don’t know if this is what my students needed,” (3%, *n* = 1), as compared to their initial impressions.

[15] Full response options were as follows: Great! Couldn’t wait to get started!/I’m glad I taught with *Math in Focus!*; Good. It has possibilities.; Ok. It’s just another math program.; Skeptical. Don’t know if this is what my students need/needed.; and Doubtful. This isn’t going to work/This didn’t work.

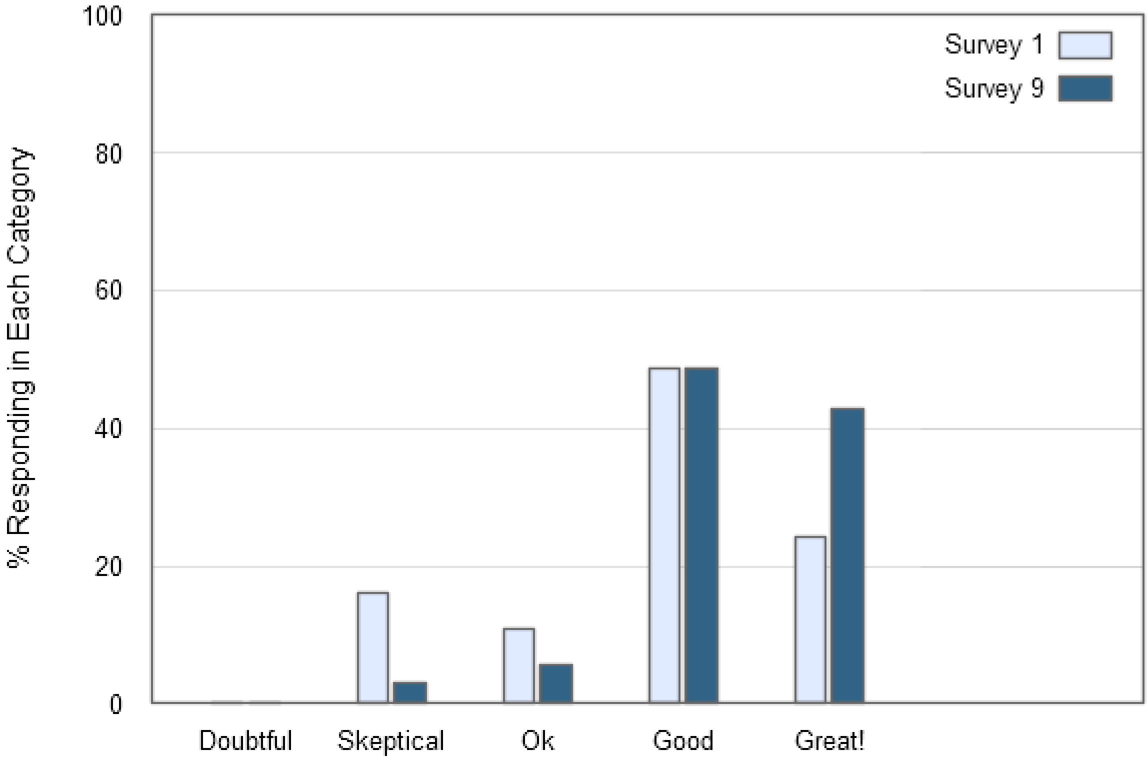

FIGURE 11. TEACHERS’ IMPRESSIONS OF *MIF*

Toward the beginning and end of the school year, in Surveys 2 and 8, we asked *MIF* teachers how well their personal beliefs regarding teaching and learning math aligned with the *MIF* approach presented in the trainings. In both surveys, the majority of teachers (53% and 69% in Surveys 2 and 8, respectively) responded that they had high levels of alignment to the *MIF* approach, as depicted in Figure 12. However, while no teachers selected “Low” at the beginning of the year, 3% selected “Low” on Survey 8, and the percent of teachers selecting “Completely” decreased by one percentage point at the end of the year. In contrast, toward the end of the school year, 16% (*n* = 7) more teachers selected, ”High” than at the beginning of the year.

Principals were asked the same question regarding alignment to *MIF* in surveys deployed in December of 2011 and May of 2012. Of the 11 principals responding to the December survey, five principals responded with “I don’t know,” two responded with “Medium,” and four responded with “High” levels of alignment. Of the ten principals responding to the May survey, two principals responded with “I don’t know,” two responded with “Medium,” five responded with “High,” and one principal responded with “Completely.”

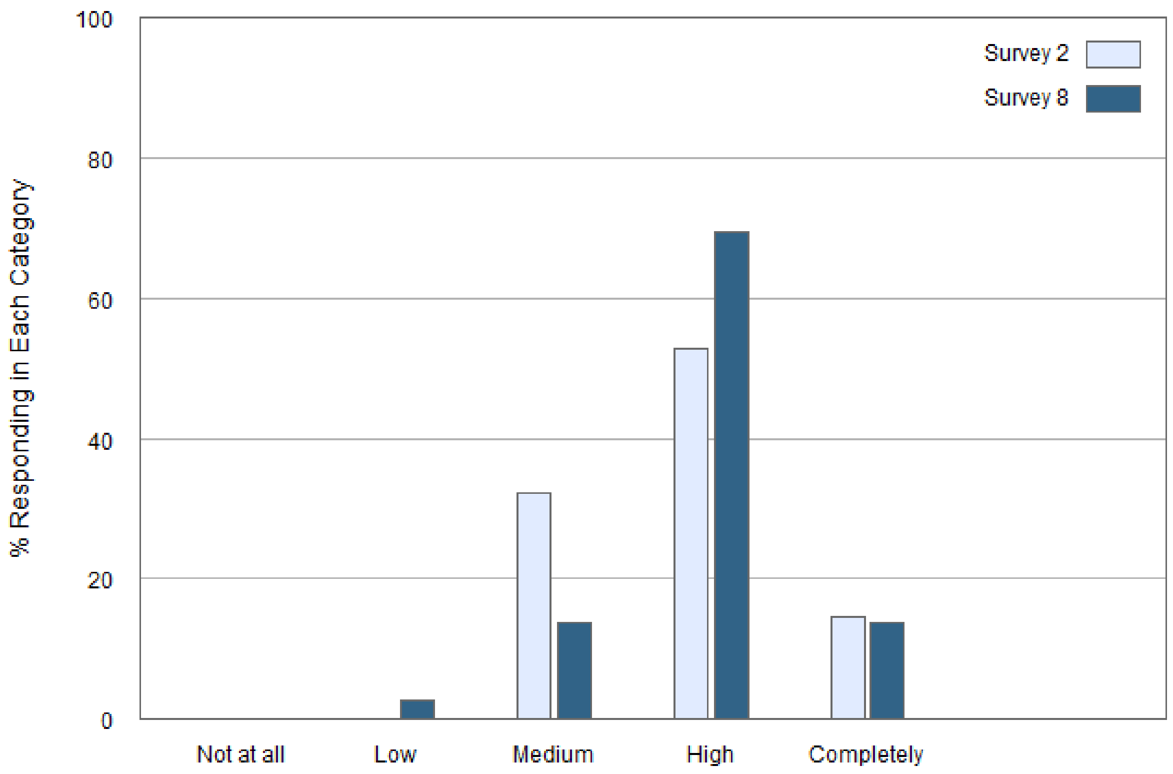

FIGURE 12. TEACHERS' ALIGNMENT TO THE *MIF* APPROACH

Surveys deployed at the end of the study asked teachers whether, given the option, they would choose to use *MIF*/their current math program. As displayed in Table 15, 76% ($n = 28$) of *MIF* teachers who responded said they would, 14% ($n = 5$) said they would not, and the remaining 11% ($n = 4$) of teachers said they did not know.[16] By comparison, only 43% ($n = 15$) of control teachers responded saying they would choose to teach with their current math program, while 31% ($n = 11$) of control teachers said they would not, and the remaining 26% ($n = 9$) of teachers responded with "I don't know." With $p < .05$, we have a high level of confidence that this result is not due to chance.

[16] Due to the rounding of decimals not all, percentages add up to 100%.

TABLE 15. IF YOU HAD THE OPTION, WOULD YOU CHOOSE TO TEACH MATH USING *MIF* / YOUR CURRENT MATH PROGRAM?

	Yes	No	I don't know
Control	15 (43%)	11 (31%)	9 (26%)
MIF	28 (76%)	5 (14%)	4 (11%)
p value			.02

Note. Due to the rounding of decimals, not all percentages add up to 100%. Two *MIF* teachers' responses are missing from this analysis.

The estimation algorithm that accounts for clustering of students in teams did not converge, therefore we present a result without adjustment for clustering.

Of the nine principals asked in May 2012, if they would recommend *MIF* to other principals and to their teachers, eight principals responded to each question saying they would, and one principal did not respond to the questions.

Summary of Teacher Satisfaction with the *MIF* Program

The majority of the teachers expressed optimism about *MIF* and indicated that their beliefs aligned with the *MIF* approach to teaching and learning math. On average, teachers' impressions of *MIF* became more favorable after they spent more time teaching with the program. Additionally, there was a general trend of increased personal alignment to the *MIF* approach as teachers spent more time teaching with the program (and as principals became more exposed to the program). Finally, there were significantly more *MIF* teachers who reported that they would teach with *MIF* than there were control teachers who reported they would teach with their math program, and all but one of the nine *MIF* principals reported that they would recommend *MIF* to other principals and to their teachers.

The above conclusions are further supported by our informal observation data. We noticed from our training observations that teachers' attitudes toward the *MIF* curriculum and research study shifted slightly from the September to the March trainings. In September, many teachers stated that they were overwhelmed because they didn't feel adequately trained before starting the curriculum, their materials were late, and the curriculum was a difficult one that necessitated more planning and thinking on the teacher's part than other common math curricula. However, in the November and subsequent trainings, most teachers reported that, though the curriculum was difficult and there was a large gap to fill for student learning, they better understood what was being asked of them and the rationale behind the teaching and learning of the program. We observed in the later training observations that many teachers acted enthusiastic about what the *MIF* curriculum was doing for them as teachers, and for their students' conceptual understanding of math. In the March training, teachers reported things such as students' improved conceptual understanding of probability, increased vocabulary when having math discussions, and higher engagement and confidence levels when using bar models to solve math word problems. Over time, teachers' views toward *MIF* became increasingly positive.

Student-level Impact Results

Overview

The primary goal of our experiment was to understand the impact of *MIF* on student math 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. We address the impact on the SAT 10 Problem Solving subscale, the primary outcome measure, and achievement on the SAT 10 Procedures subscale and the CRT state test.

Moderation of the impact: For the three outcome scales; SAT 10 Problem Solving, SAT 10 Procedures, and CRT, we examine whether the impact of *MIF* 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 varies depending on the students' pretest scores—do pretest scores moderate the impact? The other moderator we examine is minority status.

Mediation of the impact: We examine whether *MIF* has an impact on classroom practices, and if it does, whether this effect mediates subsequent impact on achievement. If there is no impact on the mediating variables we still examine if there is an association between the intermediate variable and the outcome. The potential mediator we will examine is the average percentage of Nevada state math standards covered throughout the school year, a teacher-level variable.

Program Impact on Students

SAT 10 Problem Solving

In this section we address the impacts of *MIF* on the SAT 10 Problem Solving scale.[17] Table 16 provides a summary of the samples used in the analysis and the results for the comparison of the scale scores for students in *MIF* and control groups.[18] The ‘Unadjusted’ row includes the raw means and standard deviations, as well as counts for students, teams, 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 *MIF* 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.[19]

[17] We include a series of covariates to improve precision (thus, an ANCOVA analysis) and model random team effects to reflect the cluster randomized design. The covariates included the pretest. For each grade the pretest is the year-end score from the previous grade modeled at the individual level. A CRT pretest was not available for 3rd grade (i.e., scores from the end of 2nd grade), therefore we used 2nd grade SAT 10 Problem Solving as the pretest for third grade. To accommodate this, we performed within-grade z-transformations of the pretest scores. That is, separately at each grade, we subtracted from each score the grade-specific average score for the control group and divided the result by the control standard deviation of the pretest. Thus, at each grade level the pretest is expressed as a difference from the control mean for the grade in standard deviation units for the controls. This allows us to accommodate the two pretest types (SAT 10 pretest for 3rd grade, CRT pretest for grades 4 and 5) using a common metric. We rescale pretests in this way for the remainder of the analyses in this report as well. Because the CRT is not vertically scaled, we z-transformed the posttest scores separately by grade. That is, separately at each grade, we subtracted from each score the grade-specific average posttest score for the control group and divided the result by the control standard deviation of the posttest. Thus, at each grade level the posttest is expressed as a difference from the control mean in standard deviation units for the control for that grade level. We then analyzed together the results from all three grade levels.

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

[19] SAT 10 outcomes are vertically scaled, therefore we express the raw means and regression-adjusted results in scale score units of the SAT 10.

TABLE 16. EFFECT SIZES FOR THE SAT 10 PROBLEM SOLVING OUTCOME

	Condition	Means[c]	Standard deviations	No. of students	No. of teams	No. of schools	Effect size	*p* value	Percentile standing
Unadjusted effect size[a]	Control	639.96	43.05	784	9	9	- 0.14	.55	- 5%
	MIF	634.23	40.62	857	9	9			
Adjusted effect size[b]	Control	639.96	As above				0.12	.05	5%
	MIF	644.47							

[a] The unadjusted effect size is Hedges’ g adjusted for clustering of students in teams (Hedges, 2007).

[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 *MIF* in the regression model. The program mean was obtained by adding the regression-adjusted estimate of the average one-year effect of *MIF* to the unadjusted control mean.

[c] Modeling separate school effects leads to estimates of control-group performance which are specific to schools. 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 *MIF* effect, which is constrained to be constant for each grade block (i.e., it is modeled as fixed), is added to this estimate to show the relative advantage or disadvantage to being in the *MIF* group.

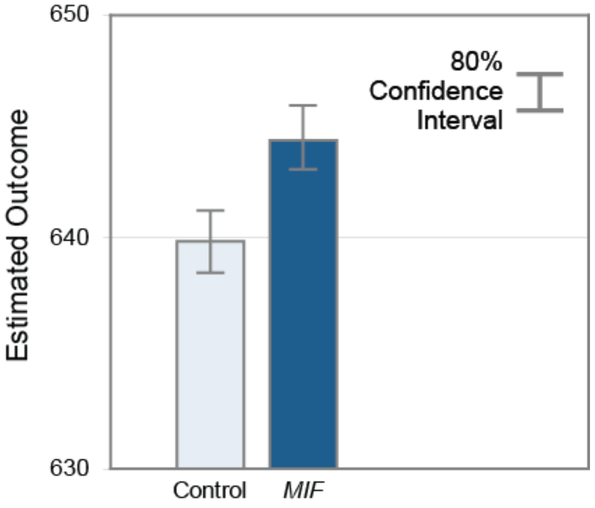

FIGURE 13. EFFECT OF *MIF* ON SAT 10 PROBLEM SOLVING OUTCOME

The adjusted analysis shows a positive impact of *MIF* on student achievement on the SAT 10 Problem Solving scale. The overall effect size (in standard deviation units) is 0.12. The low *p* value for the effect (.05) gives a high level of confidence that the result reflects a real difference beyond chance.[20]

Figure 13 is an alternative representation of the results from the benchmark analysis. The lack of overlap in the 80% confidence intervals that are added to the tops of the bars reflects the result from the benchmark analysis that we should have confidence that the result reflects a real difference attributable to the program and not just chance.

[20] We conducted a series of sensitivity analyses (see Appendix C) to determine the robustness of our result from the benchmark model to small changes in specification of the analytic model or sample. The *p* values for the impact ranged between .05 and .26. We conclude that the result from the benchmark model is not robust to all alternative specifications of the analytic model.

SAT 10 Procedures

In this section we address the impacts of *MIF* on the SAT 10 Procedures scale. Table 17 provides a summary of the samples used in the analysis and the results for the comparison of the scale scores for students in *MIF* 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 Problem Solving outcome.[21]

TABLE 17. EFFECT SIZES FOR THE SAT 10 PROCEDURES OUTCOME

	Condition	Means[c]	Standard deviations	No. of students	No. of teams	No. of schools	Effect size	*p* value	Percentile standing
Unadjusted effect size[a]	Control	634.28	47.33	792	9	9	- 0.08	.69	- 3%
	MIF	630.59	45.05	835	9	9			
Adjusted effect size[b]	Control	634.28	As above				0.14	.10	6%
	MIF	640.38							

[a] The unadjusted effect size is Hedges' g adjusted for clustering of students in teams (Hedges, 2007).

[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 *MIF* in the regression model. The program mean was obtained by adding the regression-adjusted estimate of the average one-year effect of *MIF* to the unadjusted control mean.

[c] Modeling separate school effects leads to estimates of control-group performance which are specific to schools. 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 *MIF* effect, which is constrained to be constant for each grade block (i.e., it is modeled as fixed), is added to this estimate to show the relative advantage or disadvantage to being in the *MIF* group.

The adjusted analysis shows a positive impact of *MIF* on student achievement on the SAT 10 Procedures scale. The overall effect size (in standard deviation units) is 0.14. The low *p* value for the effect (.10) gives some confidence that the result reflects a real difference attributable to the program and not just chance.[22]

[21] SAT 10 outcomes are vertically scaled, therefore we express the raw means and regression-adjusted results in scale score units of the SAT 10.

[22] We conducted a series of sensitivity analyses (see Appendix C) 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 .09 and .68. We conclude that the result from the benchmark model is not robust to all alternative specifications of the analytic model.

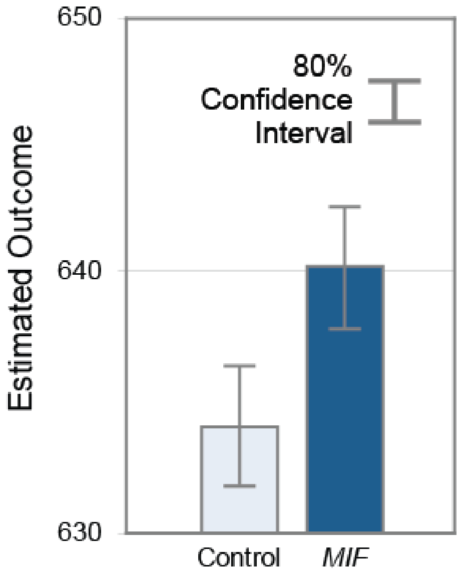

FIGURE 14. EFFECT OF *MIF* ON SAT 10 PROCEDURES OUTCOME

Figure 14 is an alternative representation of the results from the benchmark analysis. The lack of overlap in the 80% confidence intervals that are added to the tops of the bars, reflects the result from the benchmark analysis that we should have some confidence that the result reflects a real difference attributable to the program and not just chance.

CRT

In this section we address the impact of *MIF* on performance on the math CRT.

Table 18 provides a summary of the samples used in the analysis and the results for the comparison of the scale scores for students in *MIF* and control groups. The ‘Unadjusted’ row and the ‘Adjusted’ row exhibit the same kind of information as was described in the previous sections addressing the SAT 10 outcomes.[23]

[23] Because the CRT is not vertically scaled, we z-transformed the posttest scores separately by grade. That is, separately at each grade, we subtracted from each score the grade-specific average posttest score for the control group and divided the result by the control standard deviation of the posttest. Thus, at each grade level the posttest is expressed as a difference from the control mean in standard deviation units for the control for that grade level. We then analyzed together the results from all three grade levels.

TABLE 18. EFFECT SIZES FOR THE CRT OUTCOME

	Condition	Means[c]	Standard deviations	No. of students	No. of teams	No. of schools	Effect size	*p* value	Percentile standing
Unadjusted effect size[a]	Control	0.00	1.00	941	9	9	- 0.02	.88	- 1%
	MIF	- 0.02	1.07	1126	11	11			
Adjusted effect size[b]	Control	0.00	As above				0.05	.54	2%
	MIF	0.05							

[a] The unadjusted effect size is Hedges’ g adjusted for clustering of students in teams (Hedges, 2007).

[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 *MIF* in the regression model. The program mean was obtained by adding the regression-adjusted estimate of the average one-year effect of *MIF* to the unadjusted control mean.

[c] Modeling separate school effects leads to estimates of control-group performance that are specific to schools. 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 *MIF* effect, which is constrained to be constant for each grade block (i.e., it is modeled as fixed), is added to this estimate to show the relative advantage or disadvantage to being in the *MIF* group.

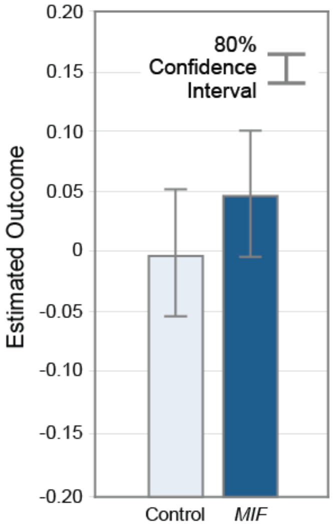

FIGURE 15. EFFECT OF *MIF* ON CRT OUTCOME

The adjusted analysis shows a small positive difference in CRT outcomes favoring *MIF*. The overall effect size is 0.05 standard deviation units. The *p* value for the effect of .54 gives no confidence that the observed difference reflects a real difference beyond chance.[24]

Figure 15 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 overlap in these intervals further indicates that we can have no confidence that the observed result reflects a real difference beyond chance.

Moderation of the Impact

Next, we report the results of our analysis of the moderating effects of pretest performance and minority status. That is, we consider whether the impact of *MIF* varies depending on a student’s pretest score or his or her

[24] We conducted a series of sensitivity analyses (see Appendix C) to determine the robustness of our result from the benchmark model to small changes in specification of the analytic model or sample. The *p* values for the impact ranged between .28 and .91. We conclude that the specifications all are consistent with the benchmark finding of no impact.

minority status.

Including Pretests as a Moderator

SAT 10 Problem Solving Pretest and SAT 10 Problem Solving Achievement

Here we assess whether the impact of *MIF* varies for students at different levels of prior achievement on SAT 10 Problem Solving. The 'Fixed Effects' in Table 19 provide the estimates of primary interest, including an estimate of the change in the impact of *MIF* for each within-grade, one-standard-deviation-unit increase on the SAT 10 Problem Solving pretest.

The moderating effect of the pretest score on the impact of *MIF*, 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 .97 indicates that we can have no confidence that the true differential impact is different from zero. The impact of *MIF* does not vary depending on the student's pretest score.

TABLE 19. MODERATING EFFECT OF SAT 10 PROBLEM SOLVING PRETEST ON THE IMPACT OF *MIF* ON SAT 10 PROBLEM SOLVING 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 school and grade with zero values for the covariates.	717.01	6.93	6	103.49	< .01
Change in outcome for the control student for a one standard deviation unit increase on the pretest	4.19	2.26	6	1.86	.11
Effect of *MIF* for a student with an average pretest	29.31	0.96	1494	30.43	< .01
Change in the effect of *MIF* for a one standard deviation unit increase on the pretest	0.04	1.29	1494	0.03	.97
Random effects	**Estimate**	**Standard error**		***z* value**	***p* value**
Team mean achievement	9.23	10.59		0.87	.19
Within-team variation	570.99	20.88		27.35	< .01

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

Note. The pretest score was z-transformed within grade by subtracting the control mean from each score and dividing this difference by the standard deviation for the controls.

SAT 10 Procedures Pretest and SAT 10 Procedures Achievement

Here we assess whether the impact of *MIF* varies for students at different levels of prior achievement on SAT 10 Procedures. The ‘Fixed Effects’ in Table 20 provide the estimates of primary interest, including an estimate of the change in the impact of *MIF* for each within-grade, one-standard-deviation-unit increase on the SAT 10 Procedures pretest.

The moderating effect of the pretest score on the impact of *MIF*, 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, 1.52, is a very small difference in the impact associated with each one-unit increase on the pretest. The *p* value of .35 indicates that we can have no confidence that the true differential impact is different from zero. The impact of *MIF* does not vary depending on the student’s pretest score.

TABLE 20. MODERATING EFFECT OF SAT 10 PROCEDURES PRETEST ON THE IMPACT OF *MIF* ON SAT 10 PROCEDURES 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 school and grade with zero values for the covariates.	659.42	10.23	6	64.47	< .01
Change in outcome for the control student for a one standard deviation unit increase on the pretest	27.83	1.23	1488	22.71	< .01
Effect of *MIF* for a student with an average pretest	6.42	3.65	6	1.76	.13
Change in the effect of *MIF* for a one standard deviation unit increase on the pretest	1.52	1.61	1488	0.95	.35
Random effects	**Estimate**	**Standard error**		***z* value**	***p* value**
Team mean achievement	31.55	28.83		1.09	.14
Within-team variation	976.08	35.77		27.29	< .01

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

Note. The pretest score was z-transformed within grade by subtracting the control mean from each score and dividing this difference by the standard deviation for the controls.

Pretest and CRT Achievement

Here we analyze whether the impact of *MIF* varies for students at different levels of prior achievement on CRT relative to the average performance for their grade level, as measured by the students’ CRT scores from the prior school year. The ‘Fixed Effects’ in Table 21 provide the estimates of primary interest, including an estimate of the change in the impact of *MIF* for each within-grade, one-standard-deviation-unit increase on the pretest.

The moderating effect of the pretest score on the impact of *MIF*, 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, 0.01, is a very small difference in the impact associated with each one-standard-deviation-unit increase on the pretest. The *p* value of .82 indicates that we can have no confidence that the true differential impact is different from zero. The impact of *MIF* does not vary depending on the student’s pretest performance relative to the mean for the grade level.

TABLE 21. MODERATING EFFECT OF PRETEST ON THE IMPACT OF *MIF* ON CRT 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 school and grade with zero values for the covariates.	0.39	0.20	7	1.97	.09
Change in outcome for the control student for a one standard deviation unit increase on the pretest	0.70	0.02	1859	30.93	< .01
Effect of *MIF* for a student with an average pretest	0.01	0.07	7	0.18	.87
Change in the effect of *MIF* for a one standard deviation unit increase on the pretest	0.01	0.03	1859	0.22	.82
Random effects	**Estimate**	**Standard error**		***z* value**	***p* value**
Team mean achievement	0.02	0.01		1.27	.10
Within-team variation	0.41	0.01		30.50	< .01

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

Note. The pretest score was z-transformed within grade by subtracting the control mean from each score and dividing this difference by the standard deviation for the controls.

Including Minority Status as a Moderator

Minority Status and SAT 10 Problem Solving

Here we assess whether the impact of *MIF* on the SAT 10 Problem Solving scale varies for students depending on minority status. The ‘Fixed Effects’ in Table 22 provide the estimates of primary interest, including an estimate of the difference between non-minority and minority students in the impact of *MIF*.

TABLE 22. THE MODERATING EFFECT OF MINORITY STATUS ON THE IMPACT OF *MIF* ON SAT 10 PROBLEM SOLVING ACHIEVEMENT

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Intercept: Outcome for minority control students with an average pretest in the reference school and grade with zero values for the covariates.	709.30	6.54	6	108.48	< .01
Change in outcome for a one standard deviation unit increase on the pretest	29.09	0.71	1611	41.02	< .01
Control group difference (non-minority minus minority) in the outcome	2.07	2.15	1611	0.96	.34
Effect of *MIF* for minority student	3.94	2.18	6	1.81	.12
Average difference (non-minority minus minority) in the effect of *MIF*	1.57	2.96	1611	0.53	.60
Random effects	**Estimate**	**Standard error**		***z* value**	***p* value**
Team mean achievement	5.03	8.07		0.62	.27
Within-team variation	625.15	22.02		28.40	< .01

[a] We do not display the fixed effect estimates for schools or covariates used in the analytic model to improve precision. The intercept value refers to a specific school (arbitrarily chosen by the estimation routine), the reference grade, 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 pretest score was z-transformed within grade by subtracting the control mean from each score and dividing this difference by the standard deviation for the controls.

The estimate of the moderating effect of minority status on the impact of *MIF*, that is, whether the intervention was differentially effective for non-minority and minority students is shown in the fifth row. The coefficient is 1.57. The *p* value of .60 indicates that we can have no confidence that the true differential impact is different from zero.

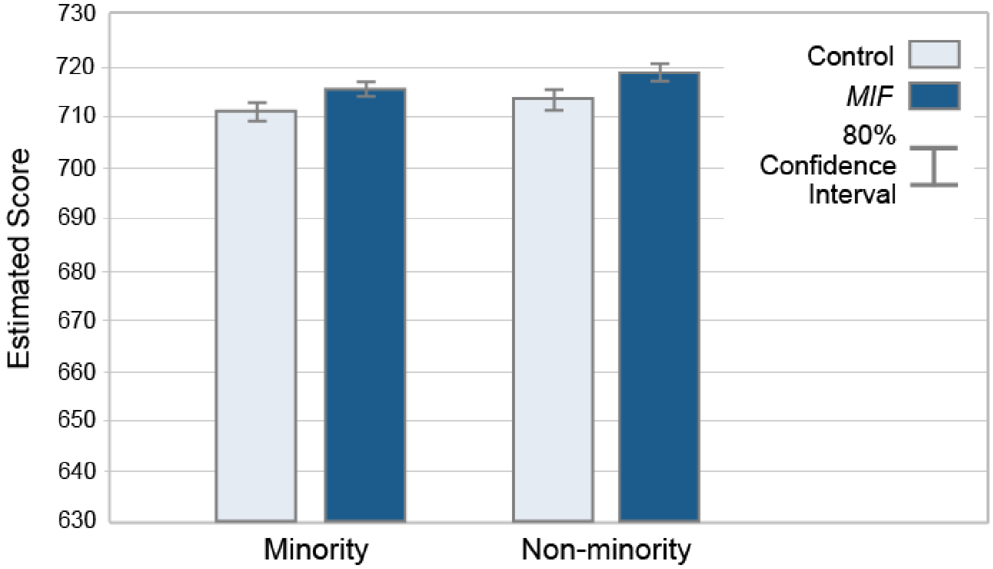

FIGURE 16. THE MODERATING EFFECT OF MINORITY STATUS ON THE IMPACT OF *MIF* ON SAT 10 PROBLEM SOLVING ACHIEVEMENT

Minority Status and SAT 10 Procedures

Here we assess whether the impact of *MIF* on SAT 10 Procedures scale varies for students depending on minority status. The 'Fixed Effects' in Table 23 provide the estimates of primary interest, including an estimate of the difference between non-minority and minority students in the impact of *MIF*.

TABLE 23. THE MODERATING EFFECT OF MINORITY STATUS ON THE IMPACT OF *MIF* ON SAT 10 PROCEDURES ACHIEVEMENT

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Intercept: Outcome for minority control students with an average pretest in the reference school and grade with zero values for the covariates.	658.39	9.79	6	67.27	< .01
Change in outcome for a one standard deviation unit increase on the pretest	28.47	0.84	1597	34.03	< .01
Control group difference (non-minority minus minority) in the outcome	- 0.47	2.70	1597	- 0.17	.86
Effect of *MIF* for minority student	4.66	3.56	6	1.31	.24
Average difference (non-minority minus minority) in the effect of *MIF*	4.79	3.76	1597	1.27	.20
Random effects	**Estimate**	**Standard error**		***z* value**	***p* value**
Team mean achievement	25.37	23.93		1.06	.15
Within-team variation	992.13	35.09		28.27	< .01

[a] We do not display the fixed effect estimates for schools or covariates used in the analytic model to improve precision. The intercept value refers to a specific school (arbitrarily chosen by the estimation routine), the reference grade 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 pretest score was z-transformed within grade by subtracting the control mean from each score and dividing this difference by the standard deviation for the controls.

The estimate of the moderating effect of minority status on the impact of *MIF*, that is, whether the intervention was differentially effective for non-minority and minority students is shown in the fifth row. The coefficient is 4.79. The *p* value of .20 indicates that we should have limited confidence that the true differential impact is different from zero.

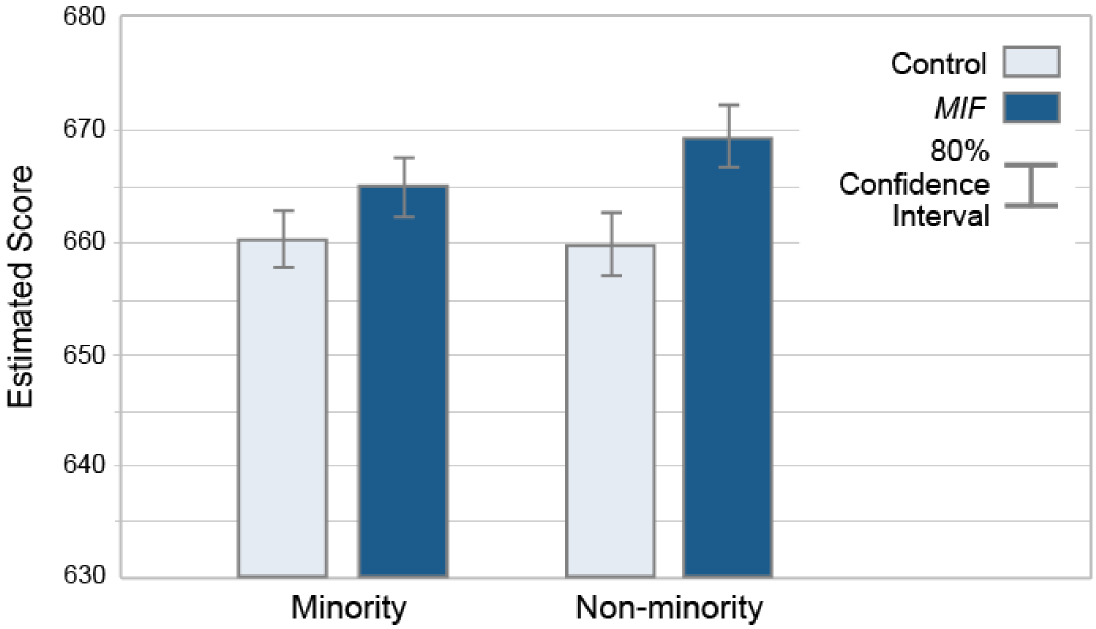

FIGURE 17. THE MODERATING EFFECT OF MINORITY STATUS ON THE IMPACT OF *MIF* ON SAT 10 PROCEDURES ACHIEVEMENT

Minority Status and CRT Achievement

Here we assess whether the impact of *MIF* on CRT achievement varies for students depending on minority status. The ‘Fixed Effects’ in Table 24 provide the estimates of primary interest, including an estimate of the difference between non-minority and minority students in the impact of *MIF*.

TABLE 24. THE MODERATING EFFECT OF MINORITY STATUS ON THE IMPACT OF *MIF* ON CRT ACHIEVEMENT

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Intercept: Outcome for minority control students with an average pretest in the reference school and grade with zero values for the covariates.	0.31	0.21	8	1.48	.18
Change in outcome for a one standard deviation unit increase on the pretest	0.70	0.02	2035	42.82	< .01
Control group difference (non-minority minus minority) in the outcome	0.01	0.05	2035	0.09	.93
Effect of *MIF* for minority student	0.01	0.08	8	0.14	.89
Average difference (non-minority minus minority) in the effect of *MIF*	0.14	0.07	2035	1.91	.06
Random effects	**Estimate**	**Standard error**		***z* value**	***p* value**
Team mean achievement	0.02	0.01		1.36	.09
Within-team variation	0.45	0.01		31.90	< .01

[a] We do not display the fixed effect estimates for schools or covariates used in the analytic model to improve precision. The intercept value refers to a specific school (arbitrarily chosen by the estimation routine), the reference grade, 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 pretest score was z-transformed within grade by subtracting the control mean from each score and dividing this difference by the standard deviation for the controls.

The estimate of the moderating effect of minority status on the impact of *MIF*, that is, whether the intervention was differentially effective for non-minority and minority students is shown in the fifth row. The difference in impact is 0.14 standard deviation units. The *p* value of .06 indicates that we can have some confidence that the true differential impact is different from zero, favoring non-minority students.

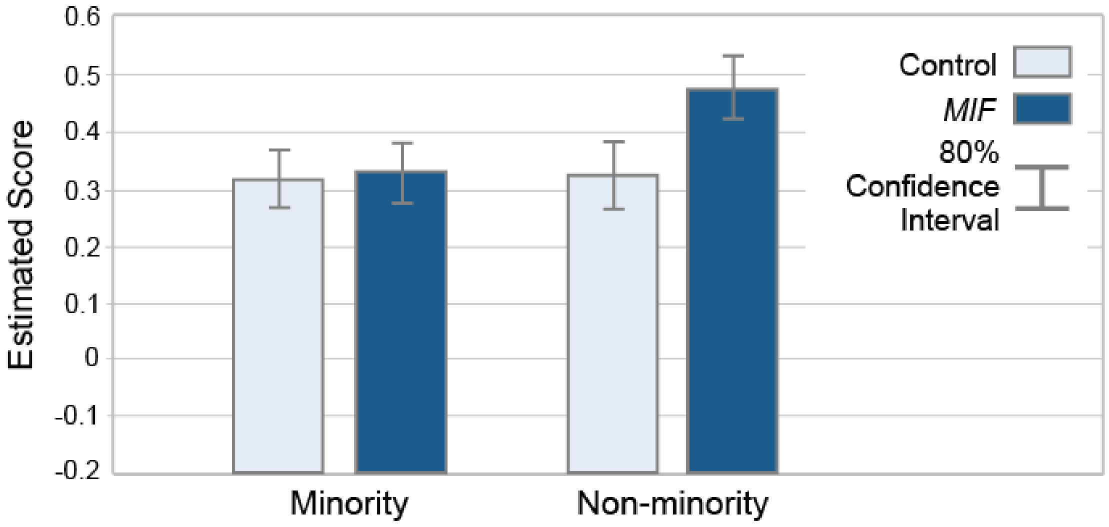

FIGURE 18. THE MODERATING EFFECT OF MINORITY STATUS ON THE IMPACT OF *MIF* ON CRT ACHIEVEMENT

Mediation of the Impact of *MIF*

Mediation occurs when an impact of the program on student achievement is accounted for in whole or in part 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.

We examined whether an impact of *MIF* on the percentage of Nevada state math standards covered mediates impact on SAT 10 Problem Solving, SAT 10 Procedures, and CRT achievement. We proposed the research hypothesis that *MIF* would support a ‘depth-over-breadth’ approach to math instruction, leading to deeper coverage of fewer standards and that this would mediate a positive impact of *MIF* on problem solving achievement and a negative impact of the program on procedures achievement. This hypothesis was not borne out, as we did not find an impact of *MIF* on the percent of Nevada standards covered. The mean percentages of standards covered were 74.58 (sd = 18.05) and 70.65 (sd = 15.31) in the control and *MIF* groups, respectively. Based on this result we have no confidence that the difference in standards covered reflects an effect of the program beyond chance (p = .33).

We also examined the association between the percentage of Nevada state math standards covered and math achievement. 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. We applied an HL model parallel to the impact model used to assess the effect of *MIF* on achievement, but where we replaced the variable indicating *MIF* status with a measure of the percentage of standards covered. For SAT 10 Problem Solving, a one percentage point increase in standards covered was associated with a .06 scale score increase ($p = .21$). For SAT 10 Procedures, a one percentage point increase in standards covered was associated with a .23 scale score increase ($p < .01$). For the CRT, a one percentage point increase in standards covered was associated with a .003 score increase ($p = .02$) (where the posttest was z-transformed using the control mean and standard deviation within each grade, as described previously). Thus, an increase in the percentage of Nevada state math standards covered is associated with increased scores on the two assessments that test for breadth over depth (SAT 10 Procedures [$p < .01$] and CRT [$p = .02$]), and it is not associated with increased scores on the assessment that tests for depth over breadth (SAT 10 Problem Solving [$p = .21$]).

Discussion

Overview

This report presents the findings of a one-year randomized control trial investigating the effectiveness of *MIF,* a math curriculum based on the pedagogical approach used in Singapore. This approach is typified by a carefully sequenced and paced instructional style that focuses on fewer topics in greater depth at each grade level to ensure mastery and support conceptual understanding. The study took place in third, fourth, and fifth grade math classrooms during the 2011 - 2012 school year in Clark County School District, Las Vegas, Nevada. We randomly assigned either fourth and fifth grades or third grade in participating schools to the program condition, in which they used *MIF*. The remaining grade(s) formed the control group assigned to use their current math program. The study investigates whether *MIF* is effective at increasing math achievement and whether impact varies for students depending on their characteristics.

Our primary outcome measure for math achievement is the SAT 10 Problem Solving assessment, which was considered to be most sensitive to the kind of instruction fostered by *MIF*. Additionally, we use the state of Nevada's Criterion Referenced Test (CRT) and the SAT 10 Procedures assessment as outcome measures. Since *MIF* emphasizes depth over breadth in the content covered, we were interested in the potential for negative effects on the CRT, which covers the full content of Nevada's math standards. To examine this issue more specifically, we also explore whether impact on achievement is associated with the percentage of Nevada state math standards covered over the course of the school year. SAT 10 Procedures is of interest because of a concern that programs such as *MIF* may put less emphasis on computation procedures than the regular math program in the control classrooms. Finally, we gathered implementation data via teacher and principal surveys, classroom and training observations, and a teacher focus group to inform outcome results.

STUDENT IMPACT RESULTS

We found a positive impact of *MIF* on math achievement. Taking into consideration both the benchmark and sensitivity analyses, we can have some confidence in a positive impact of *MIF* on problem solving skills but more limited confidence in a positive impact on procedural skills, where we found some inconsistent results when testing alternative statistical models. We did not find an impact of *MIF* on math achievement as measured by the CRT state test. Thus on the primary measure associated with *MIF* there is evidence of a positive impact. On the additional outcomes, we can say there is no evidence that *MIF* was detrimental.

Additional exploratory analyses provide descriptions of associations between the percentage of Nevada state math standards covered and math achievement. Though there was not a difference between groups in percentage of standards covered, we did find that increases in percentage of standards covered were associated with increases in CRT and SAT 10 Procedures achievement. Increases in percentage of standards covered were not associated with increases in SAT 10 Problem Solving achievement.

The impact of *MIF* was not different depending on the student's pretest scores (i.e., as deviations from the grade-level means of the pretest) on the SAT 10 Problem Solving assessment, the SAT 10 Procedures assessment, or the CRT. There was also no moderating effect of minority status on the SAT 10 Problem Solving assessment. However, in the case of CRT and SAT 10 Procedures assessments, we found a negative differential effect of the program favoring non-minority students (based on the p values we should have some confidence in there being a differential effect for CRT, and limited confidence in a difference of impact for SAT 10 Procedures.)

IMPLEMENTATION RESULTS

After initial issues with *MIF* trainings and materials were resolved in the first two months of the school year, conditions for math instruction were good across both groups. All *MIF* teachers received the necessary materials by October 2011, and what they considered to be adequate training by the end of September 2011. The delay in sufficient materials and adequate training may have resulted in less than one full school year's exposure to the *MIF* curriculum. However, *MIF* teachers did receive significantly more professional development than control teachers, and throughout the remainder of the school year, the majority of both control and *MIF* teachers reported feeling moderately or more prepared by their math professional development/*MIF* trainings to implement their math program in their classrooms.

The majority of *MIF* teachers did not meet HMH's criteria for implementing *MIF* with fidelity. The majority of teachers supplemented *MIF* with other math programs more than HMH intended, and many teachers did not use the CPA approach in the way that HMH intended.

MIF teacher practices also did not change their practice as much as HMH expected them to. We detected no differences between *MIF* and control teachers’ use of manipulatives, time spent planning math instruction, discussions with other math teachers, and coverage of all Nevada state math standards. However, *MIF* teachers did use more transition materials, and more control teachers reported teaching Nevada state math standards at the designated time.

On average, teachers’ impressions of and alignment to *MIF* became more favorable as the school year progressed. There was a significant difference between *MIF* and control teachers in that more *MIF* teachers reported they would teach with *MIF* than control teachers said they would teach with their current math program, if given the option.

Conclusions

After a one-year pilot implementation with *MIF* we have evidence of a positive effect of the program on math problem solving but less confidence in an effect on math procedures achievement. We saw no difference between the groups on student achievement as measured by the state CRT assessment. These results largely correspond with the expectations we have had from the beginning of the study.

We found a correlation between increased percentage of math standards covered and increased SAT 10 Procedures and CRT achievement, though unexpectedly there was no difference between *MIF* and control groups in the percentage reported by the teachers of standards covered.

We found that the benefits of *MIF* on CRT and SAT 10 Procedures achievement appeared stronger for non-minority students. This finding warrants further exploration since it suggests that programs such as *MIF* may have differential value for basic procedural skills.

Teachers did not receive sufficient materials or training at the beginning of the school year, and many reported on the pressures they faced while implementing with *MIF,* caused by a disconnect between *MIF* pacing and the district pacing that was suggested for preparing students for the CRT. These reported pressures help to explain the few differences found in classroom practices between *MIF* and control teachers, as well as the large proportion of teachers who did not implement *MIF* with fidelity, as prescribed by HMH. At the end of the study, though, the majority of teachers and principals reported satisfaction with *MIF* and a desire to have more time to implement with the program. With increased training on and time with the program, teachers began to better understand what *MIF* required of them. Though many did not meet HMH expectations for fidelity, teachers did come close to meeting some fidelity cutoffs, such as using *MIF* as their core curriculum. And though the majority of *MIF* teachers did report teaching all Nevada math standards, which were often at odds with the *MIF* approach, many of these teachers taught those standards at a time more aligned to the *MIF* pacing. In this way, teachers became more capable at adapting *MIF* to their current school context in order to navigate the conflicting demands of the curriculum and various district and state testing pressures.

In later training observations and the end-of-the-year focus group, *MIF* teachers repeatedly reported an increase in their students’ conceptual understanding, as well as increased student confidence and engagement while explaining and solving math problems. Though these reports were anecdotal and were not captured by our teacher survey data, they do support the positive effect of the program on *MIF* students' problem solving skills. Follow-on research might explore whether implementation by teachers who have more time with the program, and who are in school contexts that are more aligned to the Common Core State Standards (and therefore to the pacing set forth by *MIF*), results in further gains in math achievement. Research of this nature will become especially important as state tests and district pacing calendars change and increasingly accommodate the Common Core State Standards in the coming years.

References

Center for Advancing Research & Communications (ARC). (n.d.). Variance Almanac of Academic Achievement. Retrieved on November 5, 2012 from http://arc.uchicago.edu/reese/variance-almanac-academic-achievement

Common Core State Standards Initiative. (n.d.). *Common Core State Standards for Mathematics*. Retrieved on November 5, 2012 from http://www.corestandards.org/assets/CCSSI_Math%20Standards.pdf

Education plan & budget fiscal year 2010-2011. (2011). *Clark County School District*. Retrieved on July 29, 2012 from http://ccsd.net/resources/budget-finance-department/pdf/publications/quick-facts/2011.pdf

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics*, *32*(4), 341-370. doi: 10.3102/1076998606298043

Nevada Department of Education. (2012). *Procedures for the Nevada Proficiency Examination Program 2012-2013.* Retrieved November 5, 2012, from http://www.doe.nv.gov/Assessment_Resources/

Pearson. (n.d.). *Assessment and Information*. Retrieved on November 5, 2012 from http://education.pearsonassessments.com/haiweb/cultures/en-us/productdetail.htm?pid=SAT10C

Puma, M. J., Olsen, R. B., Bell, S. H., & Price, C. (2009). *What to Do When Data Are Missing in Group Randomized Controlled Trials* (NCEE 2009-0049). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Raudenbush, S. W., Spybrook, J., Liu, X., & Congdon, R. (2006). *Optimal design* Krull, J. L., & MacKinnon, D. P. (2001). Multilevel modeling of individual and group level mediated effects. *Multivariate Behavioral Research, 36,* 249–277.

SAS Institute. (2006). *SAS/STAT software: Changes and enhancements.* (Through release 9.1). Cary, NC: Author.

Spybrook, J., Bloom, H., Congdon, R., Hill, C., Martinez, A., & Raudenbush, S. (2011). *Optimal design plus empirical evidence: Documentation for the "Optimal Design" software*. William T. Grant Foundation. Retrieved on November 5, 2012 from http://www.wtgrantfoundation.org/resources/consultation-service-and-optimal-design

State of Nevada Department of Education. (n.d.). *Criterion Referenced Tests (CRT)*. Retrieved on November 5, 2012 from http://www.doe.nv.gov/Assessments_CRT/

U.S. Census Bureau. (2010). Clark County QuickFacts from the U.S. Census Bureau. Retrieved on November 5, 2012 from http://quickfacts.census.gov/qfd/states/32/32003.html

What Works Clearinghouse (WWC). (2008). *What works clearinghouse procedures and standards handbook (version 2.0).* Retrieved November 5, 2012, from http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_procedures_v2_standards_handbook.pdf

Appendix A: *MIF* Teachers’ Preparation Levels by Grade

TABLE A1. THIRD GRADE *MIF* TEACHERS’ PREPARATION LEVELS FOR PLANNING MATH LESSONS

Preparation Level	September (*n* = 15)	November (*n* = 15)	January (*n* = 16)	March (*n* = 15)
Completely Prepared	1 (7%)	5 (33%)	3 (19%)	5 (33%)
More Than Moderately Prepared	7 (47%)	4 (27%)	2 (13%)	6 (40%)
Moderately Prepared	7 (47%)	5 (33%)	8 (50%)	4 (27%)
Less than Moderately Prepared	0 (0%)	1 (7%)	2 (13%)	0 (0%)
Not At All Prepared	0 (0%)	0 (0%)	1 (6%)	0 (0%)
Average Prep Level	3.6	3.9	3.3	4.1

Note. Percentage totals may not equal 100% because of rounding.
Two teachers did not respond in September, November, and March.
One teacher did not respond in January.

TABLE A2. FOURTH GRADE *MIF* TEACHERS’ PREPARATION LEVELS FOR PLANNING MATH LESSONS

Preparation Level	September (*n* = 9)	November (*n* = 9)	January (*n* = 9)	March (*n* = 10)
Completely Prepared	2 (22%)	1 (11%)	2 (22%)	3 (30%)
More Than Moderately Prepared	6 (67%)	4 (44%)	5 (56%)	6 (60%)
Moderately Prepared	1 (11%)	4 (44%)	2 (22%)	1 (10%)
Less than Moderately Prepared	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Not At All Prepared	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Average Prep Level	4.1	3.7	4.0	4.2

Note. Percentage totals may not equal 100% because of rounding.
Four teachers did not respond in September, November, and January.
Three teachers did not respond in March.

TABLE A3. FIFTH GRADE *MIF* TEACHERS’ PREPARATION LEVELS FOR PLANNING MATH LESSONS

Preparation Level	September (*n* = 9)	November (*n* = 10)	January (*n* = 10)	March (*n* = 8)
Completely Prepared	1 (11%)	2 (20%)	2 (20%)	1 (13%)
More Than Moderately Prepared	8 (89%)	5 (50%)	2 (20%)	3 (38%)
Moderately Prepared	0 (0%)	3 (30%)	2 (20%)	2 (25%)
Less than Moderately Prepared	0 (0%)	0 (0%)	1 (10%)	2 (25%)
Not At All Prepared	0 (0%)	0 (0%)	3 (30%)	0 (0%)
Average Prep Level	4.1	3.9	2.9	3.4

Note. Percentage totals may not equal 100% because of rounding.
Three teachers did not respond in September. Two teachers did not respond in November and January. Four teachers did not respond in March.

TABLE A4. THIRD GRADE *MIF* TEACHERS’ PREPARATION LEVELS FOR USING THE CPA APPROACH

Preparation Level	September (*n* = 16)	November (*n* = 16)	January (*n* = 17)	March (*n* = 15)
Completely Prepared	1 (6%)	5 (31%)	2 (12%)	4 (27%)
More Than Moderately Prepared	11 (69%)	4 (25%)	8 (47%)	6 (40%)
Moderately Prepared	4 (25%)	7 (44%)	5 (29%)	5 (33%)
Less than Moderately Prepared	0 (0%)	0 (0%)	1 (6%)	0 (0%)
Not At All Prepared	0 (0%)	0 (0%)	1 (6%)	0 (0%)
Average Prep Level	3.8	3.9	3.5	3.9

Note. Percentage totals may not equal 100% because of rounding.
One teacher did not respond in September and November. Two teachers did not respond in March.

TABLE A5. FOURTH GRADE *MIF* TEACHERS’ PREPARATION LEVELS FOR USING THE CPA APPROACH

Preparation Level	September (*n* = 8)	November (*n* = 9)	January (*n* = 9)	March (*n* = 10)
Completely Prepared	1 (13%)	0 (0%)	2 (22%)	3 (30%)
More Than Moderately Prepared	5 (63%)	8 (89%)	5 (56%)	4 (40%)
Moderately Prepared	2 (25%)	1 (11%)	2 (22%)	3 (30%)
Less than Moderately Prepared	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Not At All Prepared	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Average Prep Level	3.9	3.9	4.0	4.0

Note. Percentage totals may not equal 100% because of rounding.
Five teachers did not respond in September. Four teachers did not respond in November and January. Three teachers did not respond in March.

TABLE A6. FIFTH GRADE *MIF* TEACHERS’ PREPARATION LEVELS FOR USING THE CPA APPROACH

Preparation Level	September (*n* = 9)	November (*n* = 10)	January (*n* = 10)	March (*n* = 8)
Completely Prepared	0 (0%)	2 (20%)	3 (30%)	1 (13%)
More Than Moderately Prepared	6 (67%)	2 (20%)	2 (20%)	2 (25%)
Moderately Prepared	3 (33%)	6 (60%)	1 (10%)	3 (38%)
Less than Moderately Prepared	0 (0%)	0 (0%)	2 (20%)	2 (25%)
Not At All Prepared	0 (0%)	0 (0%)	2 (20%)	0 (0%)
Average Prep Level	3.7	3.6	3.2	3.3

Note. Percentage totals may not equal 100% because of rounding.
Three teachers did not respond in September. Two teachers did not respond in and November and January. Four teachers did not respond in March.

TABLE A7. THIRD GRADE *MIF* TEACHERS' PREPARATION LEVELS FOR USING MANIPULATIVES DURING CLASS

Preparation Level	September (*n* = 16)	November (*n* = 16)	January (*n* = 17)	March (*n* = 15)
Completely Prepared	3 (19%)	2 (13%)	1 (6%)	4 (27%)
More Than Moderately Prepared	8 (50%)	8 (50%)	9 (53%)	6 (40%)
Moderately Prepared	3 (19%)	6 (38%)	4 (24%)	5 (33%)
Less than Moderately Prepared	2 (13%)	0 (0%)	2 (12%)	0 (0%)
Not At All Prepared	0 (0%)	0 (0%)	1 (6%)	0 (0%)
Average Prep Level	3.8	3.8	3.4	3.9

Note. Percentage totals may not equal 100% because of rounding.
One teacher did not respond in September and November. Two teachers did not respond in March.

TABLE A8. FOURTH GRADE *MIF* TEACHERS' PREPARATION LEVELS FOR USING MANIPULATIVES DURING CLASS

Preparation Level	September (*n* = 9)	November (*n* = 9)	January (*n* = 9)	March (*n* = 10)
Completely Prepared	2 (22%)	0 (0%)	2 (22%)	2 (20%)
More Than Moderately Prepared	3 (33%)	6 (67%)	4 (44%)	5 (50%)
Moderately Prepared	4 (44%)	2 (22%)	2 (22%)	3 (30%)
Less than Moderately Prepared	0 (0%)	0 (0%)	1 (11%)	0 (0%)
Not At All Prepared	0 (0%)	1 (11%)	0 (0%)	0 (0%)
Average Prep Level	3.8	3.4	3.8	3.9

Note. Percentage totals may not equal 100% because of rounding.
Four teachers did not respond in September, November, and January. Three teachers did not respond in March.

TABLE A9. FIFTH GRADE *MIF* TEACHERS' PREPARATION LEVELS FOR USING MANIPULATIVES DURING CLASS

Preparation Level	September (*n* = 9)	November (*n* = 10)	January (*n* = 10)	March (*n* = 8)
Completely Prepared	0 (0%)	1 (10%)	2 (20%)	1 (13%)
More Than Moderately Prepared	2 (22%)	2 (20%)	2 (20%)	1 (13%)
Moderately Prepared	5 (56%)	4 (40%)	2 (20%)	3 (38%)
Less than Moderately Prepared	1 (11%)	3 (30%)	0 (0%)	3 (38%)
Not At All Prepared	1 (11%)	0 (0%)	4 (40%)	0 (0%)
Average Prep Level	2.9	3.1	2.8	3.0

Note. Percentage totals may not equal 100% because of rounding.
Three teachers did not respond in September. Two teachers did not respond in November and January. Four teachers did not respond in March.

TABLE A10. THIRD GRADE *MIF* TEACHERS' PREPARATION LEVELS FOR FACILITATING METACOGNITIVE REASONING

Preparation Level	September (*n* = 16)	November (*n* = 16)	January (*n* = 17)	March (*n* = 16)
Completely Prepared	2 (13%)	2 (13%)	2 (12%)	4 (25%)
More Than Moderately Prepared	4 (25%)	9 (56%)	6 (35%)	8 (50%)
Moderately Prepared	8 (50%)	4 (25%)	7 (41%)	4 (25%)
Less than Moderately Prepared	2 (13%)	1 (6%)	1 (6%)	0 (0%)
Not At All Prepared	0 (0%)	0 (0%)	1 (6%)	0 (0%)
Average Prep Level	3.4	3.8	3.4	4.0

Note. Percentage totals may not equal 100% because of rounding.
One teacher did not respond in September, November and March.

TABLE A11. FOURTH GRADE *MIF* TEACHERS’ PREPARATION LEVELS FOR FACILITATING METACOGNITIVE REASONING

Preparation Level	September (*n* = 9)	November (*n* = 9)	January (*n* = 9)	March (*n* = 10)
Completely Prepared	1 (11%)	0 (0%)	4 (44%)	2 (20%)
More Than Moderately Prepared	5 (56%)	6 (67%)	4 (44%)	6 (60%)
Moderately Prepared	3 (33%)	3 (33%)	1 (11%)	2 (20%)
Less than Moderately Prepared	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Not At All Prepared	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Average Prep Level	3.8	3.7	4.3	4.0

Note. Percentage totals may not equal 100% because of rounding.
Four teachers did not respond in September, November and January. Three teachers did not respond in March.

TABLE A12. FIFTH GRADE *MIF* TEACHERS’ PREPARATION LEVELS FOR FACILITATING METACOGNITIVE REASONING

Preparation Level	September (*n* = 9)	November (*n* = 10)	January (*n* = 10)	March (*n* = 8)
Completely Prepared	1 (11%)	2 (20%)	2 (20%)	1 (13%)
More Than Moderately Prepared	4 (44%)	2 (20%)	2 (20%)	2 (25%)
Moderately Prepared	4 (44%)	3 (30%)	5 (50%)	5 (63%)
Less than Moderately Prepared	0 (0%)	3 (30%)	0 (0%)	0 (0%)
Not At All Prepared	0 (0%)	0 (0%)	1 (10%)	0 (0%)
Average Prep Level	3.7	3.3	3.4	3.5

Note. Percentage totals may not equal 100% because of rounding.
Three teachers did not respond in September. Two teachers did not respond in November and January. Four teachers did not respond in March.

TABLE A13. THIRD GRADE *MIF* TEACHERS' PREPARATION LEVELS FOR FACILITATING STUDENT COLLABORATION

Preparation Level	September (*n* = 15)	November (*n* = 16)	January (*n* = 17)	March (*n* = 16)
Completely Prepared	3 (20%)	3 (19%)	4 (24%)	4 (25%)
More Than Moderately Prepared	2 (13%)	6 (38%)	3 (18%)	7 (44%)
Moderately Prepared	8 (53%)	6 (38%)	6 (35%)	4 (25%)
Less than Moderately Prepared	2 (13%)	1 (6%)	3 (18%)	1 (6%)
Not At All Prepared	0 (0%)	0 (0%)	1 (6%)	0 (0%)
Average Prep Level	3.4	3.7	3.4	3.9

Note. Percentage totals may not equal 100% because of rounding.
Two teachers did not respond in September. One teacher did not respond in November and March.

TABLE A14. FOURTH GRADE *MIF* TEACHERS' PREPARATION LEVELS FOR FACILITATING STUDENT COLLABORATION

Preparation Level	September (*n* = 9)	November (*n* = 9)	January (*n* = 9)	March (*n* = 10)
Completely Prepared	1 (11%)	0 (0%)	3 (33%)	2 (20%)
More Than Moderately Prepared	4 (44%)	7 (78%)	5 (56%)	6 (60%)
Moderately Prepared	3 (33%)	2 (22%)	1 (11%)	2 (20%)
Less than Moderately Prepared	1 (11%)	0 (0%)	0 (0%)	0 (0%)
Not At All Prepared	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Average Prep Level	3.6	3.8	4.2	4.0

Note. Percentage totals may not equal 100% because of rounding.
Four teachers did not respond in September, November and January. Three teachers did not respond in March.

TABLE A15. FIFTH GRADE *MIF* TEACHERS’ PREPARATION LEVELS FOR FACILITATING STUDENT COLLABORATION

Preparation Level	September (*n* = 9)	November (*n* = 10)	January (*n* = 10)	March (*n* = 8)
Completely Prepared	2 (22%)	2 (20%)	2 (20%)	1 (13%)
More Than Moderately Prepared	4 (44%)	3 (30%)	2 (20%)	2 (25%)
Moderately Prepared	2 (22%)	4 (40%)	4 (40%)	5 (63%)
Less than Moderately Prepared	1 (11%)	1 (10%)	1 (10%)	0 (0%)
Not At All Prepared	0 (0%)	0 (0%)	1 (10%)	0 (0%)
Average Prep Level	3.8	3.6	3.3	3.5

Note. Percentage totals may not equal 100% because of rounding.
Three teachers did not respond in September. Two teachers did not respond in November and January. Four teachers did not respond in March.

TABLE A16. THIRD GRADE *MIF* TEACHERS’ PREPARATION LEVELS FOR DIFFERENTIATING INSTRUCTION

Preparation Level	September (*n* = 16)	November (*n* = 15)	January (*n* = 16)	March (*n* = 16)
Completely Prepared	2 (13%)	2 (13%)	2 (13%)	4 (25%)
More Than Moderately Prepared	6 (38%)	5 (33%)	2 (13%)	5 (31%)
Moderately Prepared	3 (19%)	5 (33%)	5 (31%)	4 (25%)
Less than Moderately Prepared	5 (31%)	3 (20%)	5 (31%)	3 (19%)
Not At All Prepared	0 (0%)	0 (0%)	2 (13%)	0 (0%)
Average Prep Level	3.3	3.4	2.8	3.6

Note. Percentage totals may not equal 100% because of rounding.
One teacher did not respond in September, January, and March. Two teachers did not respond in November.

TABLE A17. FOURTH GRADE *MIF* TEACHERS’ PREPARATION LEVELS FOR DIFFERENTIATING INSTRUCTION

Preparation Level	September (*n* = 8)	November (*n* = 9)	January (*n* = 9)	March (*n* = 10)
Completely Prepared	1 (13%)	1 (11%)	1 (11%)	2 (20%)
More Than Moderately Prepared	4 (50%)	4 (44%)	6 (67%)	3 (30%)
Moderately Prepared	2 (25%)	4 (44%)	1 (11%)	4 (40%)
Less than Moderately Prepared	1 (13%)	0 (0%)	1 (11%)	1 (10%)
Not At All Prepared	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Average Prep Level	3.6	3.7	3.8	3.6

Note. Percentage totals may not equal 100% because of rounding.
Five teachers did not respond in September. Four teachers did not respond in November and January. Three teachers did not respond in March.

TABLE A18. FIFTH GRADE *MIF* TEACHERS’ PREPARATION LEVELS FOR DIFFERENTIATING INSTRUCTION

Preparation Level	September (*n* = 9)	November (*n* = 10)	January (*n* = 10)	March (*n* = 8)
Completely Prepared	0 (0%)	1 (10%)	1 (10%)	1 (13%)
More Than Moderately Prepared	1 (11%)	2 (20%)	2 (20%)	3 (38%)
Moderately Prepared	6 (67%)	3 (30%)	2 (20%)	4 (50%)
Less than Moderately Prepared	2 (22%)	3 (30%)	3 (30%)	0 (0%)
Not At All Prepared	0 (0%)	1 (10%)	2 (20%)	0 (0%)
Average Prep Level	2.9	2.9	2.7	3.6

Note. Percentage totals may not equal 100% because of rounding.
Thee teachers did not respond in September. Two teachers did not respond in November and January. Four teachers did not respond in March.

TABLE A19. THIRD GRADE *MIF* TEACHERS’ PREPARATION LEVELS FOR GETTING STUDENTS CAUGHT UP TO GRADE-LEVEL MATERIAL

Preparation Level	September (*n* = 16)	November (*n* = 16)	January (*n* = 17)	March (*n* = 16)
Completely Prepared	1 (6%)	3 (19%)	2 (12%)	6 (38%)
More Than Moderately Prepared	5 (31%)	4 (25%)	5 (29%)	2 (13%)
Moderately Prepared	6 (38%)	6 (38%)	4 (24%)	5 (31%)
Less than Moderately Prepared	4 (25%)	3 (19%)	5 (29%)	3 (19%)
Not At All Prepared	0 (0%)	0 (0%)	1 (6%)	0 (0%)
Average Prep Level	3.2	3.4	3.1	3.7

Note. Percentage totals may not equal 100% because of rounding.
One teacher did not respond in September, November, and March.

TABLE A20. FOURTH GRADE *MIF* TEACHERS’ PREPARATION LEVELS FOR GETTING STUDENTS CAUGHT UP TO GRADE-LEVEL MATERIAL

Preparation Level	September (*n* = 9)	November (*n* = 9)	January (*n* = 9)	March (*n* = 10)
Completely Prepared	0 (0%)	1 (11%)	2 (22%)	3 (30%)
More Than Moderately Prepared	3 (33%)	3 (33%)	5 (56%)	3 (30%)
Moderately Prepared	6 (67%)	4 (44%)	1 (11%)	4 (40%)
Less than Moderately Prepared	0 (0%)	1 (11%)	1 (11%)	0 (0%)
Not At All Prepared	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Average Prep Level	3.3	3.4	3.9	3.9

Note. Percentage totals may not equal 100% because of rounding.
Four teachers did not respond in September, November, and January. Three teachers did not respond in March.

TABLE A21. FIFTH GRADE *MIF* TEACHERS' PREPARATION LEVELS FOR GETTING STUDENTS CAUGHT UP TO GRADE-LEVEL MATERIAL

Preparation Level	September (*n* = 9)	November (*n* = 10)	January (*n* = 10)	March (*n* = 8)
Completely Prepared	0 (0%)	1 (10%)	1 (10%)	2 (25%)
More Than Moderately Prepared	1 (11%)	2 (20%)	1 (10%)	3 (38%)
Moderately Prepared	5 (56%)	4 (40%)	5 (50%)	3 (38%)
Less than Moderately Prepared	3 (33%)	1 (10%)	2 (20%)	0 (0%)
Not At All Prepared	0 (0%)	2 (20%)	1 (10%)	0 (0%)
Average Prep Level	2.8	2.9	2.9	3.9

Note. Percentage totals may not equal 100% because of rounding.
Three teachers did not respond in September. Two teachers did not respond in November and January. Four teachers did not respond in March.

TABLE A22. THIRD GRADE *MIF* TEACHERS' PREPARATION LEVELS FOR FACILITATING TEACHER COLLABORATION

Preparation Level	September (*n* = 16)	November (*n* = 16)	January (*n* = 17)	March (*n* = 16)
Completely Prepared	2 (13%)	3 (19%)	1 (6%)	7 (44%)
More Than Moderately Prepared	7 (44%)	8 (50%)	10 (59%)	9 (56%)
Moderately Prepared	5 (31%)	5 (31%)	5 (29%)	0 (0%)
Less than Moderately Prepared	2 (13%)	0 (0%)	1 (6%)	0 (0%)
Not At All Prepared	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Average Prep Level	3.6	3.9	3.6	4.4

Note. Percentage totals may not equal 100% because of rounding.
One teacher did not respond in September, November, and March.

TABLE A23. FOURTH GRADE *MIF* TEACHERS' PREPARATION LEVELS FOR FACILITATING TEACHER COLLABORATION

Preparation Level	September (*n* = 9)	November (*n* = 9)	January (*n* = 9)	March (*n* = 10)
Completely Prepared	1 (11%)	2 (22%)	5 (56%)	4 (40%)
More Than Moderately Prepared	6 (67%)	5 (56%)	4 (44%)	5 (50%)
Moderately Prepared	2 (22%)	2 (22%)	0 (0%)	1 (10%)
Less than Moderately Prepared	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Not At All Prepared	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Average Prep Level	3.9	4.0	4.6	4.3

Note. Percentage totals may not equal 100% because of rounding.
Four teachers did not respond in September, November, and January. Three teachers did not respond in March.

TABLE A24. FIFTH GRADE *MIF* TEACHERS' PREPARATION LEVELS FOR FACILITATING TEACHER COLLABORATION

Preparation Level	September (*n* = 9)	November (*n* = 10)	January (*n* = 10)	March (*n* = 8)
Completely Prepared	0 (0%)	0 (0%)	2 (20%)	2 (25%)
More Than Moderately Prepared	5 (56%)	3 (30%)	1 (10%)	2 (25%)
Moderately Prepared	4 (44%)	5 (50%)	5 (50%)	4 (50%)
Less than Moderately Prepared	0 (0%)	2 (20%)	0 (0%)	0 (0%)
Not At All Prepared	0 (0%)	0 (0%)	2 (20%)	0 (0%)
Average Prep Level	3.6	3.1	3.1	3.8

Note. Percentage totals may not equal 100% because of rounding.
Three teachers did not respond in September. Two teachers did not respond in November and January. Four teachers did not respond in March.

Appendix B: Effect Estimates

TABLE B1. ESTIMATES OF FIXED EFFECTS FROM THE MULTILEVEL ANALYSIS OF THE IMPACT OF *MIF* ON THE CRT

Fixed effects model	Estimate	Standard error	Degrees of freedom	*t* value	*p* value
Adjusted grand mean outcome on CRT for control students	0.35	0.21	8	1.68	.13
Effect of Math in Focus intervention on performance on CRT	0.05	0.07	8	0.64	.54
Effect associated with being a student of school #1 (relative to school #11)	-0.24	0.23	8	-1.03	.34
Effect associated with being a student of school #2 (relative to school #11)	-0.35	0.23	8	-1.52	.17
Effect associated with being a student of school #3 (relative to school #11)	-0.16	0.23	8	-0.72	.49
Effect associated with being a student of school #4 (relative to school #11)	-0.07	0.23	8	-0.31	.77
Effect associated with being a student of school #5 (relative to school #11)	-0.20	0.23	8	-0.85	.42
Effect associated with being a student of school #6 (relative to school #11)	-0.14	0.23	8	-0.62	.55
Effect associated with being a student of school #7 (relative to school #11)	-0.20	0.23	8	-0.85	.42
Effect associated with being a student of school #8 (relative to school #11)	-0.44	0.23	8	-1.96	.09
Effect associated with being a student of school #9 (relative to school #11)	-0.47	0.22	8	-2.09	.07
Effect associated with being a student of school #10 (relative to school #11)	-0.54	0.25	8	-2.20	.06
Effect associated with being in 3th grade (relative to 5th grade)	0.11	0.08	5	1.34	.24
Effect associated with being in 4th grade (relative to 5th grade)	0.04	0.04	5	0.94	.39
Effect associated with each unit increase on the pretest	0.69	0.02	2032	41.91	< .01
Effect associated with dummy variable indicating missing value for the pretest (relative to non-missing)	-0.28	0.07	2032	-3.86	< .01

TABLE B1. ESTIMATES OF FIXED EFFECTS FROM THE MULTILEVEL ANALYSIS OF THE IMPACT OF *MIF* ON THE CRT

Fixed effects model	Estimate	Standard error	Degrees of freedom	*t* value	*p* value
Effect associated with being male (relative to female)	-0.00	0.03	2032	-0.01	0.99
Effect associated with dummy variable indicating missing value for gender (relative to non-missing)	-0.45	0.68	2032	-0.66	.51
Effect associated with being a disabled student (relative to a nondisabled student)	-0.27	0.05	2032	-5.55	< .01
Effect associated with being eligible for Free or Reduced Price Lunch (relative to not being eligible)	-0.10	0.03	2032	-2.90	< .01
Effect associated with dummy variable indicating missing value for Free or Reduced Price Lunch eligibility (relative to non-missing)	0.10	0.10	2032	0.92	.36
Effect associated with being Asian (relative to non-minority)	0.04	0.07	81	0.63	.53
Effect associated with being Hispanic (relative to non-minority)	-0.08	0.04	81	-2.00	.05
Effect associated with being Native Indian (relative to non-minority)	0.07	0.20	81	0.33	.74
Effect associated with being designated of Mixed Ethnicity (relative to non-minority)	-0.10	0.07	81	-1.42	.16
Effect associated with being Black (relative to non-minority)	-0.17	0.06	81	-2.91	< .01
Effect associated with having a teacher with less than 4 years experience (relative to having a teacher with 4 or more years' experience)	0.04	0.06	2032	0.69	.49

[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 B2. ESTIMATES OF RANDOM EFFECTS FROM THE BENCHMARK MULTILEVEL ANALYSIS OF THE IMPACT OF *MIF* ON CRT ACHIEVEMENT

Random effects model	Estimate	Standard error	*z* value	*p* value
Variance component for sections	0.02	0.01	1.36	0.09
Variance component for students within sections	0.45	0.01	31.88	< .01

TABLE B3. ESTIMATES OF FIXED EFFECTS FROM THE MULTILEVEL ANALYSIS OF THE IMPACT OF *MIF* ON SAT 10 PROBLEM SOLVING

Fixed effects model	Estimate	Standard error	Degrees of freedom	*t* value	*p* value
Adjusted grand mean outcome on SAT 10 Problem Solving for controls	711.96	6.46	6	110.29	< .01
Effect of Math in Focus intervention on performance on SAT 10 Problem Solving	4.51	1.86	6	2.43	.05
Effect associated with being a student of school #1 (relative to school #11)	-49.69	6.65	6	-7.47	< .01
Effect associated with being a student of school #2 (relative to school #11)	-58.42	6.64	6	-8.80	< .01
Effect associated with being a student of school #3 (relative to school #11)	-45.49	6.49	6	-7.01	< .01
Effect associated with being a student of school #4 (relative to school #11)	-45.00	6.82	6	-6.60	< .01
Effect associated with being a student of school #5 (relative to school #11)	-51.50	6.94	6	-7.42	< .01
Effect associated with being a student of school #6 (relative to school #11)	-50.17	6.57	6	-7.63	< .01
Effect associated with being a student of school #7 (relative to school #11)	-49.54	6.67	6	-7.43	< .01
Effect associated with being a student of school #8 (relative to school #11)	-52.61	6.57	6	-8.01	< .01
Effect associated with being a student of school #9 (relative to school #11)	-48.79	7.38	6	-6.61	< .01
Effect associated with being a student of school #10 (relative to school #11)	-53.90	6.91	6	-7.80	< .01
Effect associated with being in 3th grade (relative to 5th grade)	-47.69	2.01	5	-23.77	< .01

TABLE B3. ESTIMATES OF FIXED EFFECTS FROM THE MULTILEVEL ANALYSIS OF THE IMPACT OF *MIF* ON SAT 10 PROBLEM SOLVING

Fixed effects model	Estimate	Standard error	Degrees of freedom	*t* value	*p* value
Effect associated with being in 4th grade (relative to 5th grade)	-22.62	1.64	5	-13.82	< .01
Effect associated with each unit increase on the pretest	28.77	0.72	1608	40.15	< .01
Effect associated with dummy variable indicating missing value for the pretest (relative to non-missing)	-8.38	2.51	1608	-3.34	< .01
Effect associated with being male (relative to female)	-1.07	1.25	1608	-0.85	.39
Effect associated with dummy variable indicating missing value for the gender (relative to non-missing)	16.68	25.43	1608	0.66	.51
Effect associated with being a disabled student (relative to a nondisabled student)	-6.66	2.04	1608	-3.26	< .01
Effect associated with being eligible for Free or Reduced Price Lunch (relative to not being eligible)	-5.56	1.42	1608	-3.92	< .01
Effect associated with dummy variable indicating missing value for Free or Reduced Price Lunch eligibility (relative to non-missing)	-0.51	4.05	1608	-0.13	.90
Effect associated with being Asian (relative to non-minority)	3.57	2.64	71	1.35	.18
Effect associated with being Hispanic (relative to non-minority)	-4.57	1.71	71	-2.67	< .01
Effect associated with being Native Indian (relative to non-minority)	-4.57	8.04	71	-0.57	.57
Effect associated with being designated of Mixed Ethnicity (relative to non-minority)	-1.52	2.75	71	-0.55	.58
Effect associated with being Black (relative to non-minority)	-4.92	2.45	71	-2.01	.05
Effect associated with having a teacher with less than 4 years experience (relative to having a teacher with no less than 4 years experience)	0.05	2.47	1608	0.02	.99

[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 B4. ESTIMATES OF RANDOM EFFECTS FROM THE BENCHMARK MULTILEVEL ANALYSIS OF THE IMPACT OF *MIF ON* SAT 10 PROBLEM SOLVING

Random effects model	Estimate	Standard error	*z* value	*p* value
Variance component for sections	3.75	7.04	0.53	.30
Variance component for students within sections	622.50	21.94	28.37	< .01

TABLE B5. ESTIMATES OF FIXED EFFECTS FROM THE MULTILEVEL ANALYSIS OF THE IMPACT OF *MIF* ON SAT 10 PROCEDURES

Fixed effects model	Estimate	Standard error	Degrees of freedom	*t* value	*p* value
Adjusted grand mean outcome on SAT 10 Procedures for control students	658.87	9.57	6	68.86	< .01
Effect of *Math in Focus* intervention on SAT 10 Procedures performance	6.10	3.16	6	1.93	.10
Effect associated with being a student of school #1 (relative to school #11)	-3.95	10.07	6	-0.39	.71
Effect associated with being a student of school #2 (relative to school #11)	-5.83	10.05	6	-0.58	.58
Effect associated with being a student of school #3 (relative to school #11)	4.11	9.91	6	0.41	.69
Effect associated with being a student of school #4 (relative to school #11)	0.44	10.27	6	0.04	.97
Effect associated with being a student of school #5 (relative to school #11)	-1.42	10.82	6	-0.13	.90
Effect associated with being a student of school #6 (relative to school #11)	-8.89	9.97	6	-0.89	.41
Effect associated with being a student of school #7 (relative to school #11)	-1.59	10.08	6	-0.16	.88
Effect associated with being a student of school #8 (relative to school #11)	-9.62	9.98	6	-0.96	.37
Effect associated with being a student of school #9 (relative to school #11)	-10.68	11.26	6	-0.95	.38
Effect associated with being a student of school #10 (relative to school #11)	-6.97	10.73	6	-0.65	.54
Effect associated with being in 3th grade (relative to 5th grade)	-46.95	3.30	5	-14.23	< .01

TABLE B5. ESTIMATES OF FIXED EFFECTS FROM THE MULTILEVEL ANALYSIS OF THE IMPACT OF *MIF* ON SAT 10 PROCEDURES

Fixed effects model	Estimate	Standard error	Degrees of freedom	*t* value	*p* value
Effect associated with being in 4th grade (relative to 5th grade)	-8.53	2.08	5	-4.11	< .01
Effect associated with each unit increase on the recalibrated pretest[a]	28.36	0.85	1594	33.28	< .01
Effect associated with dummy variable indicating missing value for the pretest (relative to non-missing)	-16.89	3.38	1594	-5.00	< .01
Effect associated with being male (relative to female)	-2.66	1.59	1594	-1.68	.09
Effect associated with dummy variable indicating missing value for the gender (relative to non-missing)	-17.20	32.13	1594	-0.54	.59
Effect associated with being a disabled student (relative to a nondisabled student)	-6.42	2.57	1594	-2.49	.01
Effect associated with being eligible for Free or Reduced Price Lunch (relative to not being eligible)	-8.42	1.79	1594	-4.69	< .01
Effect associated with dummy variable indicating missing value for Free or Reduced Price Lunch eligibility (relative to non-missing)	-9.50	5.18	1594	-1.83	.07
Effect associated with being Asian (relative to non-minority)	-0.29	3.37	71	-0.09	.93
Effect associated with being Hispanic (relative to non-minority)	-2.18	2.15	71	-1.01	.31
Effect associated with being Native Indian (relative to non-minority)	-8.90	10.17	71	-0.87	.39
Effect associated with being designated of Mixed Ethnicity (relative to non-minority)	0.94	3.44	71	0.27	.79
Effect associated with being Black (relative to non-minority)	-5.35	3.07	71	-1.74	.09
Effect associated with having a teacher with less than 4 years experience (relative to having a teacher with no less than 4 years experience)	-1.55	3.22	1594	-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.

TABLE B6. ESTIMATES OF RANDOM EFFECTS FROM THE BENCHMARK MULTILEVEL ANALYSIS OF THE IMPACT OF *MIF ON* SAT 10 PROCEDURES

Random effects model	Estimate	Standard error	*z* value	*p* value
Variance component for sections	20.96	20.88	1.00	.16
Variance component for students within sections	993.30	35.17	28.24	< .01

Appendix C: Sensitivity Analyses

TABLE C1. SENSITIVITY ANALYSES FOR IMPACT OF *MIF* ON CRT

Model	Effect estimate	*p* value
Benchmark Model	0.05	.54
Listwise deletion of cases with missing values for any covariates	0.01	.91
Analysis limited to intact blocks	0.05	.53
Analysis using team-average outcomes and covariates	0.08	.28
Run benchmark model on grades 4 and 5 only	-0.08	.49
Benchmark Model with separate random effects per cluster within each grade 4 and 5 unit	0.06	.56

TABLE C2. SENSITIVITY ANALYSES FOR IMPACT OF *MIF* ON SAT 10 PROBLEM SOLVING

Model	Effect estimate	*p* value
Benchmark Model	4.51	.05
Listwise deletion of cases with missing values for any covariates	4.38	.10
Benchmark with intact blocks only	4.78	.05
Analysis using team-average outcomes and covariates	-6.06	.26
Benchmark Model with separate random effects per cluster within each grade 4 and 5 unit	4.45	.06

TABLE C3. SENSITIVITY ANALYSES FOR IMPACT OF *MIF* ON SAT 10 PROCEDURES

Model	Effect estimate	*p* value
Benchmark Model	6.10	.10
Listwise deletion of cases with missing values for any covariate	7.12	.11
Benchmark with intact blocks only	6.50	.09
Analysis using team-average outcomes and covariates	2.40	.68
Benchmark Model with separate random effects per cluster within each grade 4 and 5 unit	6.02	.25