Replicating the CGI Experiment in Diverse Environments

Effects on Grade 1 and 2 Student Mathematics Achievement in the First Program Year

Robert C. Schoen
Mark LaVenia
Amanda M. Tazaz
Kristy Farina
Juli K. Dixon
Walter G. Secada

2020

Research Report No. 2020-02

The research and development reported here were supported by the Institute of Education Sciences, U.S. Department of Education, through Award No. R305A120781 to Florida State University. The opinions expressed are those of the authors and do not represent views of the institute or the U.S. Department of Education.

This work was reviewed and overseen by the Florida State University Institutional Review Board (FWA No. IRB00000446) as HSC number 2012.8326.

Suggested citation: Schoen, R. C., LaVenia, M., Tazaz, A., Farina, K., Dixon, J. K., & Secada, W. G. (2020). *Replicating the CGI Experiment in Diverse Environments: Effects on Grade 1 and 2 Student Mathematics Achievement in the First Program Year* (Research Report No. 2020–02). Florida State University. https://doi.org/10.33009/fsu.1601237075

Replicating the CGI Experiment in Diverse Environments

Effects on Grade 1 and 2 Student Mathematics Achievement in the First Program Year

Research Report No. 2020-02

Robert C. Schoen

Mark LaVenia

Amanda M. Tazaz

Kristy Farina

Juli K. Dixon

Walter G. Secada

Florida Center for Research in Science, Technology, Engineering, and Mathematics (FCR-STEM)
Learning Systems Institute
Florida State University
Tallahassee, FL 32306
(850) 644-2570

Acknowledgements

This study was made possible by the collaboration and hard work of many individuals beside the study authors. We are grateful to the leaders in two school districts, principals, teachers, and thousands of first- and second-grade children and their parents or guardians who agreed to participate in the study and responded to the many waves of data requests that form the basis of this report. Many unnamed school secretaries, registrars, district-level data managers, and institutional review boards at the two school districts played important roles in providing access to critical data. We are particularly appreciative of the willingness of Linda Levi and the other members of Teachers Development Group to take part in this evaluation.

We were fortunate to receive thoughtful feedback and advice from members of our expert advisory board, including Thomas Carpenter (University of Wisconsin–Madison), Victoria Jacobs (University of North Carolina at Greensboro), Susan Empson (University of Missouri), Hank Kepner (University of Wisconsin–Milwaukee), and David Purpura (Purdue University).

Kristopher Childs was integrally involved in weekly management-team meetings throughout the startup and implementation and data collection period. Lisa Brooks provided assistance in the recruitment and enrollment phase.

Many graduate research assistants and research faculty across the three university partners (Florida State University, University of Miami, University of Central Florida) provided critical support in gathering, entering, and verifying accuracy of PD-implementation data and student tests: Alain Benochoa, Wendy Bray, Zachary Champagne, Kristopher Childs, Heidi Eisenreich, Kristy Farina, Uma Gadge, Rebecca Gault, Vernita Glenn-White, Katie Harshman, Naomi Iuhasz-Velez, Karon Kerr, Edward Knote, Shelby McCrackin, Erika Moore, Magnolia Placido, Nesrin Sahin, Melissa Soto, Laura Tapp, Harlan Thrailkill, Gillian Trombley, Pooja Vaswani, Maureen Warner, and Ian Whitacre.

Christopher Rhoads and Anthony Gambino provided consultation on statistical modeling decisions and reproduction of results of the confirmatory study, sensitivity analysis, and subgroup analyses. In addition to that role, they reviewed the drafts of the report and provided important feedback on the scientific merit. Eva Yujia Li also reviewed the full draft manuscript and provided valuable feedback. Anne Thistle provided valuable assistance with copy editing.

Our cognizant program officer at IES, Wai-Ying Chow, deserves recognition for serving her role in providing expert guidance and support throughout the process and for continually challenging us to strive for ever-higher levels of quality and advancement of the education sciences.

The remaining errors or omissions are entirely the responsibility of the authors.

Table of Contents

List of Appendices

List of Tables

List of Figures

Executive Summary

This report presents interim results from the *Replicating the CGI Experiment in Diverse Environments* study. Sponsored by the Institute of Education Sciences (IES), the study involves a third-party evaluation of a highly regarded professional development (PD) program in mathematics called Cognitively Guided Instruction (CGI). This report presents results from the first year of program implementation. The focus of this report is on the impact of the CGI PD program on student achievement in mathematics. Future reports will present findings on the impacts on schools, teachers, and students after the first and second years of the program.

Background and Motivation for the Study

School districts and other educational agencies spend billions of dollars each year on teacher PD (Fermanich, 2002; Odden et al., 2002; TNTP, 2015; U.S. Department of Education, 2014; Wei et al., 2010). Although the number is growing, relatively few teacher PD programs in mathematics have undergone rigorous evaluations of efficacy (Garet et al., 2016; Gersten et al., 2014). For those that have, very small or null effects on student achievement are typical (Garet et al., 2011, 2016; Jacob et al., 2017, Kennedy, 2016a).

One of the few PD programs in mathematics that has been the subject of a randomized controlled trial and resulted in positive effects on student learning outcomes is Cognitively Guided Instruction (CGI; Carpenter et al., 1989). On the basis of a large body of theory and empirical research, the theory of change for the CGI PD program hypothesizes that involvement in the program affects teachers' mathematical knowledge for teaching, beliefs about teaching and learning, and knowledge of their individual students. These changes occur over an extended period through participation in the PD workshops and interaction with their students. They occur through an interactive and iterative process and can result in changes in teachers' approaches to mathematics instruction and, ultimately, increases in student learning in mathematics.

CGI PD programs have been implemented with tens of thousands of teachers over more than 30 years. Many models of CGI-related PD are in use. We conducted a third-party evaluation of the first two years of a three-year model designed and implemented by Teachers Development Group (TDG). At the time of the study, TDG was one of largest providers of CGI PD in the world.

The present study focuses on the effects of the CGI PD program on student achievement at the end of the first year of the program. In addition to examining whether the program increased students' performance on tests of their mathematics ability, we explored whether the program had a differential effect on various subgroups of the student population.

Study Participants and Setting

The participants in the present study included grade-1 and grade-2 teachers and their students in 22 schools in two public school districts in Florida during the summer 2013 and the 2013–14 school year. Schools, teachers, and students participated in the study voluntarily. The two school districts used the same textbook series (Dixon et al., 2013) for mathematics. Florida lawmakers had recently adopted the Common Core Standards for Mathematics (NGA & CCSSO, 2010).

Study Design

The present study used a multisite cluster-randomized controlled trial evaluation design. It involved randomization of schools to an intervention or comparison condition. Under certain conditions (e.g., sufficiently low differential attrition), this type of design can support causal inference. That is, positive or negative outcomes can be attributed to the intervention—the intervention can be said to have caused the differences.

Teachers in the 11 participating schools that were randomly assigned to the intervention condition were invited to participate in the CGI program. Teachers in the 11 comparison-group schools were invited to participate in a different professional development program, chosen by the school district, which did not focus on teaching students in the domain of number, operations, or algebraic thinking.

Research Questions

The present report addresses two central research questions:

> RQ1. What is the effect of the CGI teacher professional-development program on grade-1 and grade-2 student achievement as measured by the *Mathematics Performance and Cognition* test and the *Iowa Test of Basic Skills Math Problems* and *Math Computation* tests at the end of the first year of implementation of the program?
>
> RQ2. To what extent does the effect of the CGI program at the end of the first year of the program vary by baseline student characteristics?

Outcome Measures

Focused on the domain of number, operations, and equality at the early elementary level, a one-on-one mathematics interview called the Mathematics Performance and Cognition (MPAC; Schoen, LaVenia, Champagne, Farina, & Tazaz, 2016) test was administered to students by trained members of the research team in spring 2014. Two standardized, group-administered, paper-pencil, selected-response tests—the Iowa Test of Basic Skills Math Problems and Math Computation (ITBS–MP, ITBS–MC; Dunbar et al., 2008)—were administered to students by trained members of the research team in spring 2014. A cross-grade, vertically scaled score was used in the analysis for each outcome.

Summary of Key Findings

The present study produced the following main results:

- The CGI PD program was implemented as intended, and participants perceived the program to be of high quality.
- Overall, effect sizes for student achievement in the first year of implementation were positive for the tests that focused on problem solving, applications of mathematics, and algebraic thinking, and they were negative for the computation-focused tests.
- The intervention had a large, positive effect on grade-1 students' mathematics abilities as measured by the MPAC interview and the ITBS–MP at the end of the first year of implementation.
- The intervention had a large, negative effect on grade-2 students' mathematics abilities as measured by the ITBS–MC at the end of the first year of implementation.

The focus of the content in the first year of the program is the most likely explanation for the positive effect on grade-1 students and the differential effects by grade level. The content in the first year of the three-year CGI program focused on grade-1 mathematics and frameworks for student thinking that (as defined by the state curriculum standards for grade 1). The negative effect on grade-2 students' computational abilities might also be explained by the content focus, because grade-level expectations for grade-2 students in the area of whole-number computation are substantially higher than for grade-1 students.

The resulting credibility intervals for the main-effect of the CGI treatment in the subgroup analyses all included zero, as did the moderation analyses interaction parameter estimates. The subgroup analyses grouped students by characteristics that predated the randomization to treatment conditions and were conducted on the early-and-late-joiners sample. Notwithstanding the absence of statistical significance, several effect-size estimates may be considered substantively important.

Limitations and Next Steps

On the basis of these findings, we recommend that the program developers take swift action to adjust the content and delivery of the first year of the program to address important concerns about the potential negative effect on second-grade students' computational abilities.

The findings reported in the present report do not meet the standard cutoff for statistical significance (e.g., 95% confidence). Although the effect-size estimates represent the true effect, the lack of statistical significance may be the result a study design that was slightly underpowered for the magnitude of the observed treatment effects.

A study with sufficient statistical power for moderation analyses should be conducted to permit exploration of the potential differential effects on subgroups of the population, especially for those students who are identified as having a disability, students with different baseline achievement levels, and students from different levels of socioeconomic backgrounds. Examination of moderation and subgroup analyses also suggest that complex effects are occurring for students with limited English proficiency, and the mechanisms that may be influencing those effects should be explored further.

1. Overview of the Study

1.1. Background

Although school districts and other educational agencies spend billions of dollars each year on teacher professional development (Fermanich, 2002; Odden et al., 2002; TNTP, 2015; U.S. Department of Education, 2014; Wei et al., 2010), few teacher professional-development programs in mathematics have been the subject of rigorously designed evaluations of their effect on student learning (Garet, et al., 2016; Gersten et al., 2014; Kennedy, 2016a). Cognitively Guided Instruction (CGI; Carpenter, et al., 1999) professional-development programs are among the few mathematics-teacher professional-development programs that have been the subject of randomized controlled trials designed to evaluate the impact of the program on student learning outcomes.

At least one small experimental study of the first version of the CGI PD program found potentially positive effects on mathematics achievement of grade-1 students of CGI-trained teachers as compared with other students (Carpenter et al., 1989). A subsequent quasi-experimental test was conducted in the early 1990s in an urban setting (Villaseñor & Kepner, 1993), and the results suggested that the CGI program had a positive effect on student learning. More recently, Jacobs et al. (2007) found that a CGI professional-development program had a positive effect on student achievement in algebraic and relational thinking. Dozens of other qualitative and correlational studies published in both peer-reviewed and other sources consistently indicate promise of a positive effect on various outcome measures, including teacher knowledge of mathematics content, student thinking, and student problem-solving abilities (e.g., Carpenter et al., 1988, 1996, 1998, 1999, 2003; Franke et al., 1998; Knapp & Peterson, 1995; L. Levi, personal communication, August 31, 2011; Peterson et al., 1989; Secada & Brendefur, 2000; Turner & Celedón-Pattichis, 2011).

The corpus of literature based around CGI has had a major influence on mathematics education research and policy. For example, the Common Core State Standards for Mathematics (CCSS-M; NGA & CCSSO, 2010) reference similar taxonomies for word problem types involving addition, subtraction, multiplication, and division and reference them in all elementary grade levels. Research published by the CGI program developers related to student understanding of the meaning of the equals sign (e.g., Falkner et al., 1999) has also influenced the content of the CCSS-M.

1.2. Cognitively Guided Instruction (CGI)

CGI professional-development programs are intended to incorporate scientific knowledge of how children think about mathematics into instructional practice by focusing teachers' attention on their students' thinking processes and by providing them with principled frameworks, or taxonomies, for mathematics problems and students' strategies for solving those problems. Teachers in the CGI program learn these taxonomies and practice using them to assess their students' understanding and to inform their mathematics instruction (Carpenter et al., 1988, 1989, 1996; Carpenter & Franke, 2004; Franke et al., 2001).

The first CGI teacher professional-development program was implemented with grade-1 teachers in the summer of 1986. Its purpose was to provide an opportunity for teachers to learn about an emerging (at the time) taxonomy for classification of word problems and a related taxonomy for identifying and describing the developmental trajectory of students' strategies for solving these problems. The taxonomies for problem types and strategies were based on decades of research on how young children

learn to perform operations on whole numbers (Carpenter, 1985; Carpenter et al., 1999; Fuson, 1992; Verschaffel et al., 2007).

The developers of CGI posit that primary-grades mathematics teachers have important knowledge of students' mathematical thinking, but this knowledge is typically isn't organized in a manner that allows it to play a central role in shaping teachers' instructional decisions (Carpenter et al., 1988). CGI supports teachers' efforts to increase student learning by offering conceptual models for mathematics content and student thinking that can provide a framework for teachers to use to engage in practical inquiry in their classrooms. The long-term goal of CGI is to help teachers pay close attention to their own students' thinking in ways that support generative learning (Carpenter & Franke, 2004).

CGI is guided by the following principles (Carpenter et al., 1989; Carpenter & Franke, 2004).

1. Instruction should develop understanding by stressing relationships between skills and problem solving, and problem-solving should serve as the organizing focus of instruction.
2. Instruction should be organized to facilitate students' active construction of their own knowledge *with understanding* (Hiebert & Carpenter, 1992), and each student should be able to relate problems, concepts, or skills being learned to the knowledge that he or she already possesses.
3. Teachers should continually assess their students' thinking processes and use the information gathered to guide their instructional plans.
4. Teachers learn about student thinking by listening to students, struggling to understand what they hear, and linking information about their own students with research-based frameworks.
5. Fundamental changes in teacher practice can result from understanding and building upon students' mathematical thinking.

Implementation of the CGI principles in mathematics classrooms contrasts sharply with typical instruction in the U.S. Typical mathematics instruction involves teachers' showing children how they should solve problems, focusing on whether answers are correct, and following an externally prescribed, predetermined sequence of problems and topics to teach. This has been called the Conventional Direct Recitation approach (Gage, 2009). Rather than supporting the Conventional Direct Recitation approach, implementation of the CGI principles in classroom instruction involves teachers' attending to students' cognitive processes as they solve problems rather than primarily attending to whether they produced a correct answer, drawing inference about students' understanding based on the strategies they use to solve problems, and determining the next steps in the instructional plan based on what they learn about students. This process necessarily involves teachers' making instructional decisions based on their individual students' cognitive processes rather than adhering to an externally imposed, fixed sequence of problems provided in their curriculum materials.

1.3. CGI Professional Development for Teachers

The CGI *Guide for Workshop Leaders* (Fennema et al., 1999) states that teachers typically take between 40 and 50 workshop hours to develop an initial understanding of the CGI framework for mathematics content and student thinking. The authors assert that ongoing support should be dispersed over a long period to allow teachers to integrate their new knowledge into their instructional practice. Over time, teacher participation shifts toward more detailed discussions of student strategies and mathematical thinking. Discussion of student thinking with colleagues becomes a greater part of the teachers' professional lives, and teacher's perceptions of themselves as engaged in inquiry about student thinking becomes part of their professional identities (Franke et al., 2001).

A longitudinal study of teachers who participated in the CGI program in the 1980s suggests that many teachers need multiple years of CGI PD support and practice before the effects on students are realized (Fennema et al., 1996). On the basis of the longitudinal study and other experiences, the CGI program developers have claimed that teachers' involvement in implementation of CGI PD results in self-sustaining changes in their knowledge and practice (Franke et al., 1998, 2001).

Over three subsequent decades since the first CGI PD program was implemented, CGI professional-development programs have taken many different forms, but "in all cases, [CGI] involves the focused and informed study of the development of students' mathematical thinking in specific content domains, and it is grounded in teachers' practice" (Carpenter & Franke, 2004, p. 51). The original CGI program focused on addition and subtraction of whole numbers and involved only grade-1 teachers and students (Carpenter et al., 1989). Subsequently, the content of the program has expanded to address other central topics in elementary-school mathematics such as multiplication, division, place value, and algebraic thinking (Carpenter et al., 1999; Jacobs et al., 2007).

In CGI, teaching is conceptualized as a problem-solving activity(Carpenter, 1989). In this conceptualization of teaching, teachers continually engage in a cycle involving defining a problem related to mathematics instruction (e.g., increasing their students' understanding of place value), gathering information relevant to the problem, making a plan, carrying out the plan, and reflecting on the results with respect to the original problem. Kennedy (2016a) asserts that the CGI model primarily focuses on facilitating enactment through increasing teachers' insight into student thinking. Although that is clearly a primary component of the program, other key elements of enactment involve increasing teacher knowledge of mathematics (including mathematics content and related conventions of mathematical notation), research-based taxonomies for types of word problems, and research-based frameworks for identifying student strategies for solving problems. The CGI PD program implemented in the present study integrates several topics in current research on mathematics teacher effectiveness, including a focus on content and pedagogical content knowledge specific to the work of teaching at the early elementary level, teacher collaboration, and ongoing formative assessment. The program is described in more detail in section 3 of the present report.

1.4. Purpose of the Overall Study

The purpose of the overall study is to evaluate the implementation and impact of a CGI teacher professional-development program on teacher knowledge and beliefs, classroom instruction, and student achievement in mathematics. In addition, the overall study examines whether subgroups of students and teachers respond to the intervention differently and seeks to identify the conditions under which the program may be most effective.

The intervention program serving as the focus of the current evaluation study is a version of the CGI professional development model designed and facilitated by Teachers Development Group (TDG) under the direction of Linda Levi. The TDG program for CGI is a three-year series of professional-development workshops for grade K–3 mathematics teachers. The focus of the mathematics content is on whole number (including place-value concepts), operations on whole numbers, and algebraic thinking at the early elementary level. The design and implementation of the TDG program for CGI and its theory of change is described in more detail in section 3 of this report.

1.5. Purpose of The Present Report

The present report focuses specifically on the effects of the CGI program on grade-1 and -2 student achievement in mathematics after the first year of the three-year professional-development program.

The present report does not investigate the effect of the program on teacher outcomes such as knowledge, beliefs, or instructional practice. Those factors will be explored in subsequent publications. We elected to share these results in the form of a report, because the format allows a full reporting on all of the specified data-analysis models and their results. The main body of the report contains a summary of those results. The results of specific models are provided in the appendixes.

As the primary, confirmatory question, we examine the main effect of the program on school mathematics achievement after the first year of implementation as measured by a vertically scaled score on the 2014 Mathematics Performance and Cognition (MPAC) interview and the vertically scaled standard scores on the Iowa Test of Basic Skills (ITBS) Math Problems and Math Computation (form C, levels 7 and 8). In pursuit of answers to exploratory questions, we examine the effect of the first year of the program on various subgroups of the student population and explore potential interactions between treatment condition and student characteristics. We examine the effect of the program on mathematics achievement at the school level, because the school was the level at which random assignment to treatment condition occurred.

1.6. Context of the Study

The *Replicating the CGI Experiment in Diverse Environments* study examined the direct impact of the intervention on grade-1 and -2 teachers and the indirect impact on their students in two public school districts in the state of Florida. Recruitment of schools and teachers occurred in spring 2013. The intervention period spanned two academic years, starting in summer 2013 and ending in spring 2015.

At the beginning of the study, the state of Florida had recently adopted the Common Core State Standards (NGA & CCSSO, 2010) for their mathematics curriculum standards. During the first year of the study, the state of Florida adopted the Mathematics Florida Standards, which are similar, but not identical, to the Common Core State Standards. One noteworthy difference between the two was the addition of content standards directly related to student understanding of the equals sign in grades 2 and 4, expanding the explicit reference to student understanding of the meaning of the equals sign beyond only grade 1, where it is found in the Common Core State Standards for Mathematics. During the 2013–14 and 2014–15 school years, no state-required mathematics assessment was in place for grade-1 or -2 students in Florida, but many districts (including the two participating districts) selected or created their own assessments for these students. The two school districts used the same mathematics textbook series for these grade levels (Dixon et al., 2013).

The purposes of the overall study were (a) to estimate the impact of the CGI program on teachers' mathematical knowledge for teaching, beliefs about mathematics teaching and learning, and instructional practice' (b) to estimate the impact of the CGI program on student learning and performance in mathematics; (c) to determine whether subgroups of students and teachers responded differently; and (d) to identify the conditions under which the program might be most effective. Stated simply, we intended to determine whether, for whom, and under what conditions the CGI program had an effect on student learning.

1.7. Research Questions

The following research questions guided the evaluation of the CGI program in the overall study.

1. On average, what is the effect of the CGI program on teacher and student outcomes?
2. For whom (teachers and students) and under what conditions does the CGI program work?
3. How does the size of the effect of the CGI program on teacher outcomes vary over time?
4. What are the causal mechanisms relating treatment and student outcomes? In other words, are the teacher and student factors interrelated according to the theory of change?
    a. Do teacher pedagogical content knowledge, knowledge of student thinking, and teacher collaboration have an effect on student knowledge and beliefs?
    b. Does instructional practice mediate any of the effects detected between teacher and student attributes?

The present report focuses on the student component of the confirmatory research question (i.e., effects on student achievement outcomes) and begins to explore the second research question through investigation of interactions between treatment condition and student characteristics. These analyses addressed two areas of investigation: (a) the effects of the CGI program on various subgroups of the student population and (b) the moderation of the effects of the CGI program by student characteristic. Thus, the research questions guiding the present report are as follows.

> **RQ1.** What is the effect of the CGI teacher professional development program on grade-1 and grade-2 student achievement as measured by the MPAC interview, ITBS Math Problems, and ITBS Math Computation tests at the end of the first year of implementation of the program?
>
> **RQ2.** To what extent does the effect of the CGI program at the end of the first year of the program vary by baseline student characteristics?

In addition to these two research questions, we performed sensitivity analyses (SA) to look for potential differences in outcomes with respect to (a) how we define the analytic sample and (b) our choice of estimator in modeling the data. These analyses were driven by the following two questions.

> **SA1.** Are the estimated effects of the CGI program after the first year of the program sensitive to whether the sample includes all students measured at follow-up or is constrained to those students who attended the respective school in which they were measured at follow-up for the entire school year?
>
> **SA2.** Are the estimated effects for the CGI program after the first year of implementation sensitive to whether a Bayesian or likelihood-based method of estimation is used for the analyses?

2. Study Design and Its Realization

2.1. Study Design

The design for the present study of the impact of the CGI program was a multisite cluster-randomized trial that was blocked on district and stratified by the percentage of students in the school who were eligible for the federal free/reduced-price lunch (FRL) program. Random assignment occurred at the school level. Based on *a priori* power estimates, the target number of schools was 22. The study took place in two adjacent school districts in Florida.

Participating schools were assigned to one of two treatment conditions. Half of the schools were assigned to participate in the CGI program. The other half were assigned to participate in a district initiative that was not directly related to number, operations, place value, or algebraic thinking in mathematics. District A selected a program they called *Bridge to STEM*. The Bridge to STEM program was based on a National Science Foundation (NSF) supported program called *Ramps and Pathways* (Zan & Escalada, 2011). The district program provided two days of teacher workshops as well as related lesson plans and materials necessary to implement the lesson plans. District B selected a program they called *Data-driven Science Instruction: Analyzing Students' Misconceptions in Science*.

2.2. Recruitment Process

2.2.1. School Recruitment

During the school-recruitment phase starting in January of 2013, the research project personnel worked with the original district partner (District A) to obtain a list of elementary schools deemed eligible by the district leaders to participate in the research study. At the request of one of the regional superintendents in District A, several elementary schools were removed from the list of eligible schools because of other obligations with the district. The resulting list of eligible schools comprised 90% of the total elementary schools in District A. The project principal investigator contacted each of the eligible school principals in District A with information about the study. The e-mail requested any interested principals to identify teachers within their schools who might be interested in participating. Principals who agreed to allow the study to occur in their schools provided the research project personnel with a list of teachers in grades 1 and 2 who taught elementary mathematics.

The initial response from principals and teachers in District A was lower than anticipated. To ensure that at least 22 schools were recruited as per the *a priori* power analysis, the research project personnel contacted the leaders in a neighboring school district (District B) to ask whether they would be interested in participating in the research study. After approval by the superintendent of District B, the principal investigator provided recruitment information to all elementary principals in District B that was similar to that provided to District A principals, and the same process was followed; principals sent contact information for interested teachers at their schools.

2.2.2. Teacher Recruitment

All the teachers identified by their principals in District A and District B through the process described above were contacted through e-mail. The message contained information about the research study and a link to an online questionnaire asking teachers for their consent to participate and some background information about them. Although randomization ultimately occurred at the school level, teachers voluntarily consented to participate in the research study in accordance with the process

approved by the FSU Institutional Review Board. Principals were not allowed to register teachers directly.

After the recruitment window closed, all teachers who consented to participate in the research study were sorted on the basis of school name. Because the developer of this particular CGI program strongly recommended that at least three teachers per school participate, the minimum participation rate for each school and grade level was set at three teachers. The result was a list of 22 eligible schools. Teachers who voluntarily consented to participate in the study but were not in schools where this minimum criterion for eligibility was achieved were excluded from the randomization sample and notified of their ineligible status. After randomization occurred, but before any participants were informed of their randomly assigned treatment condition, the research project personnel continued to recruit as many of the known, remaining, and eligible grade-1 and grade-2 teachers from the 22 randomized schools.

In addition to grade-1 and grade-2 classroom teachers, teachers serving in a support capacity, such as math coaches, curriculum resource teachers, or intervention teachers, were also enrolled as study participants. They were included, because the school was the unit of randomization, and they were part of the community in the school contributing to student learning in mathematics in grades 1 and 2. Because the current report concerns impact of the CGI program on student outcomes, discussion of the sample will be constrained primarily to that of classroom teachers and their students.

2.2.3. Student Recruitment

Before the beginning of the academic school year, all participating classroom teachers in the 22 participating schools were provided with parental consent forms to distribute to incoming students in their classrooms. The teachers distributed the consent forms to parents and collected them, then relayed the returned consent forms to project personnel.

2.2.4. Participation Incentives

Schools were reimbursed for the cost of substitute teachers on the days the teachers participated in workshops occurring on school days. Schools were paid $1,000 per year for their participation in all aspects of data collection (e.g., consent forms, student testing, video recording of classrooms, delivery of class rosters).

Teachers were remunerated for participation in professional-development workshops occurring outside of their contracted hours with the school districts and completion of web-based questionnaires on their own time. Teachers were paid $125 per day of workshops they attended in the summer or on Saturdays. Teachers were not paid an additional amount for baseline data collection, which was considered part of the registration process. Treatment-condition teachers were paid $50 to complete the web-based questionnaires at the end of the first year of the study. Comparison-condition teachers were paid $125 to complete those same questionnaires. In all, each participating teacher in the treatment condition received up to $800, and each participating teacher in the comparison condition received up to $375 for participation in the first year of the study. Treatment-condition teachers received two CGI books (Carpenter et al., 1999, 2003). Comparison-condition teachers in District A received lesson plans and a class set of blocks, ramps, and marbles to implement *Bridge to STEM* activities in their classrooms. Teachers in both districts received credit for the hours they participated in professional-development workshops, which could be applied toward the renewal of their teaching credentials.

To support the likelihood of a high rate of return of parental consent forms, students were offered a book they were invited to select from a list of age-appropriate books approved by the schools. Regardless of whether their parent or guardian agreed to their participation in the study, students received the book for returning the completed consent form.

2.3. Randomization Procedures

The 22 schools that met the eligibility requirements were each assigned at random to be in the CGI (i.e., treatment) condition or the comparison condition. All schools were drawn from one of two adjacent school districts. Fifteen schools were located in a large, urban school district (District A). Seven schools were located in an adjacent, suburban school district (District B).

2.3.1. Random Assignment of Schools to Treatment Condition

Through a stratified block-randomized design (Raudenbush, Martinez, & Spybrook, 2007), schools were ranked on percentage FRL, and within-district matched pairs were formed. With equal probability within each matched-pair randomization block, one school was randomly assigned to the CGI condition, the other to the comparison condition. In order to provide schools timely notification of their assigned condition, random assignment was conducted on a rolling basis. Table 2.1 presents the procedure taken for conducting the stratified block random assignment. All schools in the sample had a .50 probability of assignment to treatment condition. Within each district, an odd number of eligible schools was recruited, so one school within each district was not part of a matched-pair randomization block. Each of those schools was instead assigned to condition with a .50 probability through a coin-toss simulation.

Table 2.1 Timeline for Procedure of Rolling Random Assignment

Date	District block	FRL stratification procedure	Resulting assignment	
			Treatment	Comparison
April 10, 2013	District A	Sorted 12 schools by SP-FRL, paired schools by rank, and randomly assigned one school from each pair to treatment and the other to comparison.	6	6
May 15, 2013	District B	Sorted 6 schools by SP-FRL, paired schools by rank, and randomly assigned one school from each pair to treatment and the other to comparison.	3	3
May 17, 2013	District B	Used a coin-toss simulation to assign the single school randomly to a condition.	1	0
May 24, 2013	District A	Sorted 3 schools by SP-FRL, paired the two schools most similar in SP-FRL, and randomly assigned one to treatment and the other to comparison. Used a coin-toss simulation for the third school to determine condition.	1	2

Note. SP-FRL = School percentage of students eligible for free/reduced-price lunch for school year 2012–13.

The school percentage FRL for the 22 randomly assigned schools ranged from 11 to 100 percent, with a sample mean of 65.7 and standard deviation of 26.9. Table 2.2 presents the school percentage FRL for the sample schools in school year 2013–14, disaggregated by treatment condition and district. Across the two districts, the randomized sample had a sample mean school percent FRL of 62.3 and standard deviation of 30.4 for the Treatment group and a sample mean of 69.0 and standard deviation of 23.9 for the comparison group.

Table 2.2 School Percentage FRL School Year 2013–14, Disaggregated by Treatment Condition and District

	Treatment				Comparison			
	M	*SD*	Min	Max	*M*	*SD*	Min	Max
District A	73.52	27.94	38.80	100.00	74.50	23.91	44.12	100.00
District B	42.70	26.54	11.20	69.28	54.49	20.28	33.06	73.38
Total	62.31	30.36	11.20	100.00	69.04	23.87	33.06	100.00

Note. District A Treatment *n* = 7 and Comparison *n* = 8. District B Treatment *n* = 4 and Comparison *n* = 3.

2.3.2. Defining the Intent-to-Treat Sample

Intent-to-treat analysis involves analyzing participants as if they received the treatment to which they were assigned, regardless of amount of treatment actually received. With a school-level unit of assignment, the intent-to-treat sample for the study was all grade 1 and grade 2 teachers and students in the 22 participating schools during the 2013–14 school year. Recruitment of students was conducted with assistance from participating teachers. Participating teachers distributed a letter from the principal investigator to parents and guardians of students in their classes during the first two weeks of the school year. The parents or guardians were asked to sign the form and return it to their children's teacher if they consented to their children's participating in the study. We attempted to measure student mathematics achievement for all students with consent to participate in the study. Sample attrition is discussed in section 2.4.1.

School and teacher participation in the present study was voluntary. Schools and teachers were not required to participate. The intervention, or treatment, in the present study is therefore conceptualized as the *opportunity for teachers of grades 1 or 2 mathematics to participate in the CGI PD program*, and the opportunity was offered at the school level. As described in section 3, most, but not all, of the relevant teachers in the treatment-condition schools took part in the opportunity.

2.3.3. Random Selection of Students for Interview-based Mathematics Assessment

In addition to administering whole-class measures of student achievement to all participating students, we conducted one-on-one mathematics interviews with a stratified random sample of up to four students from each participating teacher's classroom. Spring 2014 interviews were conducted with students in the sample who completed baseline tests at the beginning of the 2013–14 school year.

To maintain a balanced sample within each classroom with respect to student gender, we used gender as the first stratum. Student gender data were collected along with spring class rosters provided by participating schools. These data were later confirmed by the school districts.

The second stratum involved splitting the class by baseline test achievement level on the fall 2013 Elementary Mathematics Student Assessment (EMSA; Schoen, LaVenia, Bauduin, & Farina, 2016). Class

rosters were divided into four subcategories based on gender and median achievement level: upper boy, lower boy, upper girl, lower girl. A random number was assigned to each student, and the sample was sorted by gender, baseline-test stratum, and random number. Then, a primary and an alternate student were selected from each stratum on the basis of the random number. The highest random number designated the primary student; the second highest the alternate. Alternate students were only called upon to be interviewed in instances where the primary student was absent or did not assent to being interviewed. Although all four strata were represented in almost every class, some classes did not have an alternate (or even a primary) student for every stratum, resulting in fewer than four students interviewed from those classrooms.

2.4. Characteristics of the Sample and Setting

2.4.1. Recruited Sample and Attrition Rates of Analytic Samples

Using guidelines from the What Works Clearinghouse (WWC; U.S. Department of Education, 2013, 2016), we reference their terminology of *stayers* and *joiners* to define our analytic samples. Stayers are individuals who were in clusters at the time of randomization and were measured at follow-up. Because the earliest record we obtained for student enrollment was fall 2013, and random assignment was conducted in spring 2013, we are unable to determine which of the students in the sample are true stayers and which ones joined in the fall. Accordingly, we define all students who were enrolled in their respective school as of fall 2013 as *early joiners*. To be included in an early-joiner analytic sample for Year 1 of the study, a given student must contribute outcome data at the spring 2014 follow-up for the same school in which he or she was enrolled in fall 2013. We define *late joiners* as those students who enrolled in their respective schools after August 2013. Therefore, all students contributing outcome data in the spring 2014 follow-up comprise the early- and late-joiner analytic sample for Year 1, regardless of their August 2013 school of enrollment. Table 2.3 presents sample-size information defining the reference populations and analytic samples for the current study.

Sample attrition occurred when parents did not actively consent to their child's participation or when data for students with consent were not available (i.e., measurement attrition). Because student recruitment occurred through participating teachers, mathematics achievement data were not gathered for students in nonparticipating teachers' classrooms (in either treatment- or comparison-condition schools). This decision created the largest single source of student-level attrition. Students whose parents declined to consent to participate represent are counted in the attrition rates reported in Table 2.4.

Table 2.4 presents the rates of attrition and representativeness for the MPAC and ITBS student analytic samples (see section 2.5 for a description of the MPAC interview and ITBS). Attrition rates are calculated for the MPAC and ITBS analytic samples, but because the MPAC was only administered to early joiners, no early-and-late-joiner MPAC sample is available on which to calculate a representativeness rate. Only early joiners participated in the EMSA pretest; late joiners did not have an opportunity to participate in the pretest.

Table 2.3. Student Reference Population and Analytic Sample Sizes

	Grade 1		Grade 2		Grades pooled	
	T	C	T	C	T	C
Participating schools	11	11	11	11	11	11
Schools contributing student data	10	11	11	11	11	11
Grade 1 or 2 teachers in participating schools[a]	64	80	69	78	140	158
Teachers contributing spring 2014 student data	46	49	45	43	91	92
Student membership in participating schools[a]						
Reference population for fall 2013 early joiners	1,078	1,297	1,160	1,356	2,366	2,653
Reference population for spring 2014 follow-up	1,048	1,288	1,153	1,355	2,330	2,643
Reference subpopulation for spring 2014 MPAC	256	320	276	312	560[b]	632[b]
Participating students[c]						
Early and late joiners						
With spring 2014 ITBS	576	527	547	522	1,123	1,049
With spring 2014 ITBS and fall 2013 test	535	490	511	469	1,046	959
Early joiners only						
With fall 2013 Pretest	650	576	603	544	1,253	1,120
With spring 2014 MPAC[d]	161	175	144	142	305	317
With spring 2014 MPAC and fall 2013 test	161	175	143	141	304	316
With spring 2014 ITBS	562	513	538	507	1,100	1,020
With spring 2014 ITBS and fall 2013 test	534	489	510	469	1,044	958

Note. T = Treatment condition; C = Comparison condition. MPAC = Mathematics Performance and Cognition Interview; ITBS = Iowa Test of Basic Skills.

[a]Teacher counts and student membership reported in the one school with grade 1 measurement attrition are excluded from the grade-1 column but included in the grades-pooled column.

[b]Reference subpopulation for the spring 2014 MPAC is calculated as four multiplied by the number of grade-1 or grade-2 classroom teachers in participating schools.

[c]Participating students are defined as all those in grades 1 and 2 with parental consent to participate in the study.

[d]All students with spring 2014 MPAC data were early joiners.

Table 2.4. Student Analytic Sample Attrition and Representativeness for Analyses Pooled across Grades

	Analytic sample *N*			Reference population *N*			Attrition/ Representativeness	
Outcome	Treatment	Comparison	Total	Treatment	Comparison	Total	Overall	Differential
				Early and late joiners[a]				
ITBS	1,123	1,049	2,172	2,330	2,643	4,973	56.32%	8.51%
				Early joiners only				
MPAC	305	317	622	560	632	1,192	47.82%	4.31%
ITBS	1,100	1,020	2,120	2,366	2,653	5,019	57.76%	8.04%

Note. The MPAC interview was not administered to any students in the late-joiner sample.

[a]Students present at follow-up. See Section 2.4.1. for explanation of sample composition.

2.4.2. Characteristics of Participating Teachers

Table 2.5 presents the demographic characteristics for the 2013–2014 participating teacher sample. The sample includes a total of 236 teacher participants: 103 grade-1 teachers (52 treatment; 51 comparison), 97 grade-2 teachers (49 treatment; 48 comparison), and 36 support teachers, such as math coaches (20 treatment; 16 comparison). Gender distribution was 98% female in the treatment condition and 100% female in the comparison condition. The proportions of each race/ethnicity for teachers in the treatment and comparison conditions were 13% and 6% Black, 8% and 15% Hispanic, 4% and 4% Multiracial, and 81% and 83% White, respectively. Seventy-six percent of treatment teachers and 83% of comparison teachers had four or more years of teaching experience. For 69% of the treatment teachers and 63% of the comparison teachers, the highest degree earned was a bachelor's; the remainder had earned a master's degree or higher.

Table 2.5. Teacher Sample Demographic Characteristics

	Treatment (*n* = 121)		Comparison (*n* = 115)		Total (*n* = 236)	
	n	Proportion	*n*	Proportion	*n*	Proportion
Gender						
Male	3	0.02	0	0.00	3	0.01
Female	118	0.98	115	1.00	233	0.99
Race						
Black	16	0.13	7	0.06	23	0.10
Multiracial	5	0.04	5	0.04	10	0.04
White	98	0.81	96	0.83	194	0.82
Unknown	0	0.00	2	0.02	2	0.01
Declined to answer	2	0.02	5	0.04	7	0.03
Hispanic						
Hispanic	10	0.08	17	0.15	27	0.11
Not Hispanic	107	0.88	93	0.81	200	0.85
Declined to answer	4	0.03	5	0.04	9	0.04
Grade role						
1	52	0.43	51	0.44	103	0.44
2	49	0.40	48	0.42	97	0.41
Other Support Staff	20	0.17	16	0.14	36	0.15
Years of teaching experience						
Three or fewer	29	0.24	19	0.17	48	0.20
Four or more	92	0.76	96	0.83	188	0.80
Highest degree earned						
Bachelor's degree	84	0.69	73	0.63	157	0.67
Master's degree	35	0.29	39	0.34	74	0.31
Professional diploma	2	0.02	2	0.02	4	0.02
Professional degree	0	0.00	1	0.01	1	<0.01

Note. Hispanic = Hispanic/Latino ethnicity. Proportions may not sum to exactly 1.00 as a result of rounding errors.

2.4.3. Student Demographics

The analytic samples vary by outcome measure. Whereas the ITBS analytic sample had an upper bound of all students in participating teachers' classrooms, the MPAC analytic sample was constrained by a stratified sampling procedure that restricted the sample to a maximum of four students per participating classroom.

Table 2.6 presents the student demographics for the 2014 MPAC analytic sample. The sample includes a total of 622 student participants: 305 in the treatment condition and 317 in the comparison condition. Gender distribution was 48% male in the treatment condition and 49% male in the comparison condition. The proportions of each race/ethnicity for students in the treatment and comparison conditions were 7% and 4% Asian, 20% and 20% Black, 29% and 48% Hispanic, 4% and 2% Multiracial, and 41% and 27% White, respectively. The prevalence of economic disadvantage in the student sample was 48% FRL in treatment and 74% FRL in comparison. English language learners comprised 16% of the sample in the treatment condition and 29% of the sample in the comparison condition. The proportions of student exceptionality in the conditions were 6% and 7% students with disabilities and 7% and 3% gifted for treatment and comparison, respectively. Student demographics were unknown for approximately 1% of the sample. See Table A.1 in Appendix A for student demographics for the 2014 MPAC early joiners analytic sample, disaggregated by district.

Table 2.6. Student Demographics for the 2014 MPAC Early-Joiners Analytic Sample

	Treatment (*n* = 305)		Comparison (*n* = 317)		Total (*n* = 622)	
	n	Proportion	*n*	Proportion	*n*	Proportion
Gender						
Male	147	.48	156	.49	303	.49
Female	158	.52	161	.51	319	.51
Race/Ethnicity[a]						
Asian	20	.07	13	.04	33	.05
Black	60	.20	62	.20	122	.20
Hispanic	87	.29	151	.48	238	.38
Multiracial	13	.04	6	.02	19	.03
White	123	.41	84	.27	207	.33
FRL[a]	146	.48	235	.74	381	.62
ELL[a]	49	.16	91	.29	140	.23
Exceptionality[a]						
SWD	19	.06	21	.07	40	.07
Gifted	22	.07	9	.03	31	.05

Note. Asian = Asian/Pacific Islander, non-Hispanic; Black = Black/African American, non-Hispanic; Hispanic = Hispanic/Latino ethnicity, any racial group; Multiracial = Multiracial or American Indian/Alaskan Native, non-Hispanic; White = White, non-Hispanic. FRL = Eligible for free/reduced-price lunch. ELL = English language learners. SWD = Students with disabilities. Gifted = Gifted and talented. Unknown = Missing demographic data.
[a]This information was unavailable for 3 students in the sample, 2 in the treatment condition and 1 in the comparison condition.

Table 2.7 presents the student demographics for the 2014 ITBS analytic sample. The sample consisted of 2,172 students: 1,123 in the treatment condition and 1,149 in the comparison condition. Gender distribution was 51% male in the treatment condition and 49% male in the comparison condition. The proportions of each race/ethnicity in the conditions were 6% and 4% Asian, 17% and 19% Black, 29%

and 45% Hispanic, 3% and 3% Multiracial, and 45% and 29% White, for Treatment and Comparison, respectively. The prevalence of economic disadvantage in the student sample was 48% FRL in Treatment and 73% FRL in Comparison. English language learners constituted 18% of the Treatment sample and 27% of the Comparison. The proportions of student exceptionality in the conditions were 6% and 8% students with disabilities and 5% and 3% gifted, for Treatment and Comparison, respectively. Student demographics were unknown for approximately 1% of the sample.

Table 2.7. Student Demographics for the 2014 ITBS Early- and Late-Joiners Analytic Sample

	Treatment (*N* = 1,123)		Comparison (*N* = 1,049)		Total (*N* = 2,172)	
	n	Proportion	*n*	Proportion	*n*	Proportion
Gender						
Male	574	.51	509	.49	1,083	.50
Female	547	.49	539	.51	1,086	.50
Unknown	2	<.01	1	<.01	3	<.01
Race/Ethnicity						
Asian	68	.06	40	.04	108	.05
Black	187	.17	202	.19	389	.18
Hispanic	327	.29	472	.45	799	.37
Multiracial	30	.03	30	.03	60	.03
White	503	.45	300	.29	803	.37
Unknown	8	.01	5	<.01	13	.01
FRL	541	.48	770	.73	1,311	.60
Unknown	8	.01	5	<.01	13	.01
ELL	204	.18	288	.27	492	.23
Unknown	8	.01	5	<.01	13	.01
Exceptionality						
SWD	70	.06	88	.08	158	.07
Gifted	61	.05	30	.03	91	.04
Unknown	8	.01	5	<.01	13	.01

Note. Asian = Asian/Pacific Islander, non-Hispanic; Black = Black/African American, non-Hispanic; Hispanic = Hispanic/Latino ethnicity, any racial group; Multiracial = Multiracial or American Indian/Alaskan Native, non-Hispanic; White = White, non-Hispanic. FRL = Eligible for free/reduced-price lunch. ELL = English language earners. SWD = Students with disabilities. Gifted = Gifted and talented. Unknown = Missing demographic data. Because of rounding, categories may not sum to 1.00.

See Table A.2 in Appendix A for student demographics for the 2014 ITBS early- and late-joiners analytic sample, disaggregated by district. See Table A.3 in Appendix A for student demographics for the 2014 ITBS early-joiners analytic sample. See Table A.4 in Appendix A for student demographics for the 2014 ITBS early-joiners analytic sample, disaggregated by district.

2.5. Data Sources and Data Collection Procedures

Data collection efforts for the presently described study served four main purposes: to form blocks for the purpose of randomizing schools to treatment condition, to allow examination of baseline equivalence of student mathematics achievement for the treatment and control conditions, to define subgroups for use as covariates or moderators in the statistical models, and to permit estimation of mathematics achievement for the student outcomes of interest.

2.5.1. Student Mathematics Achievement

2.5.1.1. Fall 2013 Student Baseline Tests: Grades 1 and 2.

Students with consent to participate completed a written, whole-class-administered mathematics test named the Fall 2013 Elementary Mathematics Student Assessment (EMSA; Schoen, LaVenia, Bauduin, & Farina, 2016) at the beginning of the 2013–14 school year. The purpose of the test was to permit examination of baseline equivalence of the students in treatment and comparison schools and to serve as a covariate in the statistical models estimating impact of the treatment and exploring potential moderators. The fall 2013 student tests were designed to measure student ability to answer correctly questions related to counting, solving word problems, and performing computation involving addition or subtraction. The tests were designed to be aligned with the learning expectations in the Common Core State Standards for Mathematics (NGA & CCSSO, 2010). The content and format of items and scales were reviewed by experts in mathematics and mathematics education (Schoen, LaVenia, Bauduin, & Farina, 2016).

The test materials were delivered to participating schools during the week of teacher preplanning for the school year. Teachers were asked to administer the tests within the first three weeks of the school year. Along with class rosters, tests were retrieved by research project personnel approximately 4–6 weeks after the beginning of the school year.

The full research report for the Fall 2013 EMSA provides information about test items, administration instructions, data modeling and scoring procedures, and diagnostic and supplementary analyses of scales and subscales, including ordinal forms of Revelle's beta and McDonald's omega hierarchical coefficients and IRT information-based reliability estimates (Schoen, LaVenia, Bauduin, & Farina, 2016). The student baseline test data generated by the Fall 2013 EMSA were modeled by means of a second-order factor analysis model with Counting, Word Problems, and Computation as three lower-order factors and Mathematics as the single higher-order factor. The test forms at the two grade levels were not vertically scaled. The chi-square and root mean square error of approximation (RMSEA) statistics and the comparative fit (CFI) and Tucker-Lewis (TLI) indices for the grade 1 model were $\chi^2(87) = 1159.026$, $p < .001$; RMSEA = .100, 90% CI [.095, .105]; CFI = .929; and TLI = .914. The corresponding numbers for the grade 2 model were $\chi^2(62) = 276.759$, $p < .001$; RMSEA = .055, 90% CI [.048, .062]; CFI = .962; and TLI = .952. The composite reliability estimates for the higher-order Mathematics scores for the grade 1 and grade 2 samples were .84 and .89, respectively.

2.5.1.2. Mathematics Performance and Cognition (MPAC) Interview.

The achievement score generated by the MPAC student interview was used as one of three primary outcomes of interest in the confirmatory study. Focused on the domain of number, operations, and equations, the MPAC interview was designed to be used (a) to measure student achievement in mathematics and (b) to gather information about the cognitive strategies students used to solve the mathematics problems (Schoen, LaVenia, Champagne, Farina, & Tazaz, 2016).

The MPAC interview consisted of a series of mathematics problems that the students were asked to solve in a one-on-one interview setting. The interviewer posed a fixed set of problems to the student, observed how the student solved the problems, asked the student to report the strategies he or she used, and recorded the student's responses. The MPAC interview used a semistructured format. The sequence and wording of the general instructions and the mathematics problems were designed to be presented in the same order and spoken exactly from the interviewer's script. Subsequent follow-up questions varied and depended upon the interviewer's ability to perceive and understand the student's

strategy as well as the student's ability to demonstrate or articulate how he or she arrived at the given answer. The interview lasted on average approximately 45 minutes and ranged from about 30 minutes to about 60 minutes.

The development process for this interview involved expert review that verified the alignment of the content of the interview with current research and with fundamentally important ideas in grade-1 or grade-2 mathematics that are consistent with the content of the CCSS-M (NGA & CCSSO, 2010).

Interviews were conducted by a team of research faculty with mathematics teaching experience and graduate students in mathematics education. Interviewer training occurred in several phases over a period of approximately 6 weeks. Each interview was video recorded. The video recordings of a stratified random sample of 79 interviews were also coded by an additional trained reviewer as a check for consistency among interviewers of the implementation of the protocol and coding of data. The overall rate of interrater agreement for whether students provided correct or incorrect answers on individual items was .96 (Schoen, LaVenia, Champagne, Farina, & Tazaz,, 2016).

The student test data generated by the Spring 2014 MPAC were modeled by means of a second-order factor analysis model with Number Facts, Word Problems, Operations on Both Sides of the Equals Sign, Equals Sign as a Relational Symbol, and Computation as five lower-order factors and Mathematics as the single higher-order factor. The RMSEA, CFI, and TLI goodness-of-fit statistics indicated that the models provided a close fit to the data. The grade 1 higher-order model-fit statistics were $\chi^2(204) = 281.69$, $p < .001$; RMSEA = .03, 90% CI [.02, .04]; CFI = .98; and TLI = .98. The grade 2 higher-order model fit statistics were $\chi^2(225) = 301.75$, $p < .001$; RMSEA = .04, 90% CI [.02, .04]; CFI = .98; and TLI = .98. The composite reliability estimates for the higher-order Mathematics scores for the grade 1 and grade 2 samples were each .92. The full research report for the Spring 2014 MPAC (Schoen, LaVenia, Champagne, Farina, & Tazaz, 2016) presents test items, administration instructions, model specifications, and diagnostic and supplementary analyses of scales and subscales, including ordinal forms of Revelle's beta and McDonald's omega hierarchical coefficients and IRT information-based reliability estimates.

The grade 1 and grade 2 MPAC scales comprised 22 and 23 items, respectively, among which 20 items were used at both grade level scales. The high proportion of items common to the two scales was used to scale the two forms vertically, allowing analyses that pool across grade level. We employed Bayesian measurement invariance modeling (Muthén & Asparouhov, 2013) to calculate a cross-grade vertically scaled score based on the higher-order Math factor. Within a vertical scaling context, Bayesian measurement invariance modeling involved specifying approximate invariance between grades for factor loadings (i.e., metric invariance) and item thresholds (i.e., scalar invariance). Finding of approximate metric invariance indicated that the items were related to the latent factors equivalently across grades, ensuring the same latent factors were being measured in each grade. Finding of approximate scalar invariance indicated that items had the same expected response at the same absolute level of the trait, meaning the observed differences in the proportion of responses for each grade were due to factor mean differences only. For the structural portion of the model, factor means were allowed to vary freely across grade, reflecting the expectation that grade 2 students would have higher factor means than grade 1 students.

2.5.1.3. Iowa Test of Basic Skills: Math Problems and Math Computation.

Iowa Test of Basic Skills (ITBS; Dunbar et al., 2008) is a whole-class-administered, paper-pencil, norm-referenced, vertically scaled test designed to measure skills and achievement in fundamental content areas of school curricula. Many states have used the ITBS as the primary assessment test in their school accountability systems. The ITBS therefore serve as a policy-relevant outcome that can be used to

estimate potential effect on tests that school leaders might use to compare potential effects on their own standardized tests.

Two of the mathematics tests selected from the complete battery were administered by the evaluation team in the spring of each year: ITBS Math Problems (ITBS–MP), which measures students' abilities to perform symbolic computation, and ITBS Math Computation (ITBS–MC), which measures students' abilities to solve word problems. These tests were administered for the purpose of measuring students' achievement on widely used standardized tests in mathematics. Level 7, form C, was used with grade 1 students. Level 8, form C, was used with grade 2 students.

The ITBS Standard Scores on the ITBS–MP and ITBS–MC use a Rasch-based model. Each item on the tests is scored dichotomously (correct or incorrect). Missing item data are considered incorrect for scoring purposes. The Kuder-Richardson Formula 20 was used to estimate reliability, and the internal consistency estimates for the ITBS–MP were .83 and .85 for the Level 7 and Level 8 tests, respectively. The reliability estimates for the Level 7 and Level 8 ITBS–MC were .84 and .83, respectively (Buros Institute, 2010). Reliability estimates based on the Kuder-Richardson 20 formula and our sample of 1,159 grade 1 students and 1,144 grade 2 students are provided in Table 2.8.

Table 2.8. Reliability Estimates for the ITBS–MP and ITBS–MC by Means of the Kuder-Richardson 20 Statistic with Data from the Spring 2014 Sample

Test level	Item *N*	Person *N*	ITBS–MP	ITBS–MC
7	28	1,159	.836	.858
8	30	1,144	.849	.844

Note. MP = Math Problems; MC = Math Computation; Item N = number of items; Person N = number of examinees.

The ITBS tests were administered by research faculty and graduate student members of the evaluation team in sample schools between April 23–May 23, 2014. To avoid introducing bias due to timing of tests, testing occurred on the same day or on adjacent days for schools in matched pairs created during the process of randomization of schools to treatment condition. Test booklets were mailed to the publisher for data entry and scoring. We used the vertically scaled ITBS Standard Score for each of the two ITBS tests.

2.5.2. School and Student Characteristics

For the purpose of stratifying schools by percent FRL in the random assignment procedure, school FRL data were obtained from publicly available data files housed on the Florida Department of Education's website (http://www.fldoe.org/accountability/data-sys/edu-info-accountability-services/pk-12-public-school-data-pubs-reports/index.stml).

For the purpose of determining the school enrollment size to serve as the reference population in attrition calculations for the early-joiner analytic sample, grade 1 and grade 2 enrollment data were obtained from publicly available data files housed on the same Florida Department of Education web page. The enrollment data we drew from pertained to Survey 2, which was conducted the week of October 14–18, 2013.

Participating schools provided data on the number of grade 1 and grade 2 students who were enrolled during the last two weeks of the 2013–14 school year. These enrollment data served as the reference population in attrition calculations for the early- and late-joiner analytic samples. Participating teachers

returned class rosters, which provided sufficient information about consenting students to allow us to obtain demographic data from their districts. Data requests through the districts also provided demographic information on individual students, including gender, race, ethnicity, FRL eligibility, ELL eligibility, disability status, gifted status, and date of enrollment for the 2013–14 school year.

2.6. Test of Baseline Equivalence for Student Mathematics Achievement between Treatment and Comparison Conditions

We assessed the baseline equivalence for the analytic sample of clusters by fitting the baseline test data to multilevel regression models with teacher and school random effects and a dichotomous indicator for treatment as the only independent variable. As reported in Table 2.9, results indicated a small, but nonnegligible, group difference of 0.11 standard deviations favoring the treatment group. According to WWC (U.S. Department of Education, 2013) guidelines, although not statistically significant at $p < .05$, baseline group differences of this size warrant covariate adjustment for these factors when the effect of treatment on outcome measures is estimated.

Table 2.9. Cluster-Adjusted Baseline Equivalence of Student Mathematics

	Unadjusted baseline test descriptives				Multilevel regression estimates		
	Treatment		Comparison				
Dependent variable	*N*	*SD*	*N*	*SD*	*Coeff*	*g*	*p*
Grade 1 baseline test	650	0.716	576	0.701	0.080	0.11	.497
Grade 2 baseline test	603	0.730	544	0.764	0.081	0.11	.584

Note. *g* = Hedges' *g* effect size. Multilevel regression estimates are based on models with teacher and school random effects and with treatment as the only independent variable. The baseline population for fall 2013 early joiners was used as the reference. Representativeness of the grade 1 baseline-test analytic sample indicated an overall attrition of 48.4% and differential attrition of 15.9%; for the grade 2 pretest analytic sample, overall attrition was 54.4% and differential attrition was 11.9%.

3. Description and Implementation of the First Year of CGI Intervention

3.1 Selection of the CGI Professional Development Provider for the Efficacy Study

The CGI PD program evaluated in the present efficacy study was created and taught by Teachers Development Group (TDG) under the direction of Linda Levi, the Director of CGI Initiatives for TDG and a coauthor of: three CGI books (Carpenter et al., 1999; Carpenter, Franke, & Levi, 2003; Empson & Levi, 2011), a manual for CGI workshop leaders (Fennema et al., 1999), and the 2nd edition of the primary CGI book (Carpenter et al., 2015). At the time of the study, TDG was the world's largest provider of CGI professional development for teachers. Over several years implementing CGI professional development and performing formal and informal evaluations of the effect on teachers, TDG had refined its professional-development plan to consist of three years of professional development, each year consisting of workshops during the summer and the academic year. In 2005, TDG provided professional development for five cohorts of teachers; in 2010, TDG provided CGI professional development to over 80 cohorts of teachers (L. Levi, personal communication, June 8, 2011). Each workshop cohort typically comprised approximately 25–30 teachers. To meet the demand, TDG created a network with more than 30 experienced CGI teachers and university mathematics educators, who have provided CGI professional development workshops in locations across the United States. Several U.S. states, including Iowa and Arkansas, have set goals to provide CGI professional development for all of their elementary teachers through the TDG model. Thousands of teachers in these (and other) states have participated in this program. The present study focuses on the first two years of the three-year TDG CGI PD model.

3.2. Design of the CGI Professional Development Program

3.2.1. Teacher Workshops

Eight workshop days per year created the time frame for the TDG CGI program. The PD program, as designed, included a four-day summer workshop (24 hours of activities) and two two-day follow-up sessions (another 24 hours of activities) held during a single school year. The planned amount of time for teachers to attend professional development workshops in year 1 amounted to 48 hours of face-to-face, workshop-based professional development time per teacher. Teachers spend additional time outside of the workshops completing reading activities, posing problems to students, analyzing their students' work and bringing it to the follow-up workshop, and participating in CGI team meetings with their colleagues in their schools. This same basic structure is repeated in the design of the second year of the program, but the content and substance of the PD experiences become increasingly sophisticated.

The CGI program was designed to focus teachers' attention on their students' mathematical thinking and to provide teachers with principled frameworks for understanding this thinking. Teachers learned about two complementary researched-based frameworks:

- Problem Types Frameworks, which describe how the structure of a problem influences how children think about the mathematical concepts embedded in the problem, and
- Solution Strategy Frameworks, which describe the developmental progressions of children's mathematical thinking as illustrated by their strategies for solving problems within the problem-type framework.

The frameworks addressed in the CGI program describe children's thinking about (a) addition and subtraction, (b) multiplication and division, (c) base-ten number concepts, and (d) early algebraic ideas (Carpenter et al., 1999, 2003). Created by the program developers, Table 3.1 provides an overview of the content focus for each of the eight days of workshops in the first year of the program.

Table 3.1. CGI Year 1 Agenda Overview

Day 1	Day 2	Day 3	Day 4
• Introduction to CGI • Direct modeling • CGI problem types	• Strategies for addition and subtraction problems	• Multiplication and division problem types and students' strategies • Interviewing children	• Introduction to students' strategies for multidigit addition and subtraction • Getting started with using CGI in your own classroom
Day 5 (Fall)	**Day 6 (Fall)**	**Day 7 (Spring)**	**Day 8 (Spring)**
• Classroom-embedded work—base-ten number concepts	• Problem types and students' strategies for developing understanding of the base-ten number system	• Classroom-embedded work—multidigit addition	• Students' strategies for solving addition and subtraction problems with large numbers

The four-day summer workshop provided participants with an introduction to CGI problem-types and solution-strategies frameworks. Participants first experienced these frameworks by analyzing video of students solving problems. After they demonstrated an initial understanding of the perspectives of children at various points in a developmental learning trajectory, they were asked to anticipate the strategies that students might use to solve particular problems. The four-day summer workshop also provided participants with an introduction to using information about children's thinking to guide mathematics instruction.

Although the PD did not prescribe specific teaching practices, participants were encouraged to focus on understanding problem solving as described in the standards for mathematical practice (NGA & CCSSO, 2010) and were encouraged to increase their practice and skill in using children's thinking to guide instruction. Participants in this CGI program are introduced to the *purposeful pedagogy model* (Jaslow & Evans, 2012), a tool intended to help teachers use information about students' thinking to guide

instruction, in the beginning of their second year of the program. The purposeful pedagogy model consists of the following steps to be used in planning and implementing mathematics instruction:

1. Analyze a student's current level of understanding.
2. Set a learning goal for each students based on his or her understanding and the grade level standards.
3. Design instruction to engage children in this learning goal.

Participants practiced step one of this model when they viewed video of students' solving problems. Midweek during the four-day summer workshop, teachers continued their practice in analyzing a student's current level of understanding when the workshops provided them with an opportunity to interact with grade 1 and grade 2 students through live, one-on-one mathematics interviews. Participants were given approximately one hour to interview students using a predetermined set of word problems. They were asked to refrain from teaching, as they were there only to observe how the students were solving these types of problems and to ask clarifying questions to come to understand the child's thinking. After the student interviews, participants returned to the workshop session and engaged with all three steps of the purposeful pedagogy model when they linked data from their interviews to the CGI solution strategies frameworks (Carpenter et al., 1999, 2003) and discussed how they could use what they learned in this interview to help advance students learning if this was an everyday classroom.

Grade-level teams from each school were given the task of holding a *CGI team meeting*. This meeting is a time for a team of teachers to plan a problem to pose to the students and then a place for those teachers later to discuss student work/thinking. The summer institute culminated with participants' planning for the beginning of the school year and setting goals for how and when they intended to accomplish CGI team meetings at their schools.

Each two-day follow-up session included one *classroom-embedded* workshop day (Levi, 2017; Nielsen et al., 2016). The classroom-embedded workshop day engaged participants with the purposeful pedagogy model within the context of an actual classroom. The day began with each teacher's interviewing and/or observing one (or two) students from a volunteer host classroom solving a set of facilitator-determined problems. Participants were once again reminded that they were not to teach the student how to solve the problem; rather they were there to observe how they solved the problem and to ask questions (when necessary) to gain more information on how students were thinking about the problem. After the interviews, the cohort of teacher participants analyzed the students' thinking, linked students' thinking to CGI frameworks, set learning goals for these students, and designed instruction that would engage students with the content in the learning goals. The learning goal was typically different for children at different developmental levels, but all learning goals were linked to the same mathematics concept. In the afternoon (typically after lunch), the cohort returned to the host classroom and observed as the facilitator implemented the instructional plan that was developed by the cohort. Participants were asked to observe the students they had interviewed earlier that day, but from a distance, to see whether any changes appeared in the way the students were thinking and solving the selected problem in the instructional plan. The instructional plan typically included a component in which, at the beginning of the lesson, the facilitator called on students who had been purposefully selected by the workshop participants, to explain their thinking while solving the interview problems with the class. Once all purposefully selected students shared their solutions to the interview problem, the class was presented with the newly developed problem from the instructional plan for them to solve while the participants observed. Note, the purposefully selected students are not always those who generated a correct solution to the problem. The workshop facilitator sometimes selected incorrect responses to be

presented if the solution was valid and they felt the classroom should observe the solution. At the conclusion of the classroom lesson, the participants returned to the meeting room, where they related and discussed their observations during the classroom lesson. To close these follow-up days, participants repeated portions of this protocol using student work from their own classes. They analyzed their students' written strategies, set a learning goal for their students, designed a problem to engage students with this learning goal, and planned how they would lead a whole-class discussion of student strategies for solving that problem.

3.2.2. Eligibility/Target Participants/Setting

The TDG CGI program is designed for teachers of grades K through 3 mathematics. Teachers of the various grade levels participate in the same program. Classroom teachers and other instructional support personnel are also encouraged to participate in the program. In the current study, facilitators worked with mathematics teachers, coaches, and other mathematics instructional support personnel in the participating treatment-condition schools.

On the basis of her previous experiences, the Director of CGI Initiatives at TDG (Linda Levi) expressed a preference for having at least three teachers per school participate, a number that would provide opportunities for teachers to continue discussions and collaborate in their schools on a regular basis (rather than only during the CGI workshops). As described in section 2.2.2, this minimum was incorporated into the eligibility criteria for school participation, and the minimum was met in every school participating in the study.

The TDG model set a limit of 30 participants per cohort. As a result, the year 1 program consisted of 5 cohorts of grade-1 and -2 teachers. TDG expressed a strong preference to keep each cohort of up to 30 participants intact for the full year, and the TDG workshop leader remained the same throughout the year. The design of the TDG program did not warrant that these cohorts (and matched TDG workshop leaders) stay intact across multiple years of the program. Although the participants can continue through the multiyear program in consecutive years, the program need not be completed in three consecutive years.

Ideally, schools have at least three teachers who are participating in the CGI program together, and each individual teacher participant should complete all three years of the program. Work environments that encourage teachers to make continual adjustments to the instructional plan based on their professional judgment about the instructional needs of students are more likely to encourage implementation of CGI than environments where teachers are asked or required to follow a rigid pacing guide created by an external person or committee.

3.2.3. Learning Objectives for Teachers

Participants are provided with many opportunities to learn mathematics throughout every session in the CGI PD. They spend considerable time and attention on learning how to express mathematical ideas using written notation, including formally acknowledged conventions of mathematical notation as well as mathematical notation invented by students and teachers to express mathematical ideas in the moment of problem solving. They learn the language and vocabulary for algebraic concepts that undergird elementary arithmetic—many of which are tacitly understood by teachers. For example, they learn to recognize when student strategies are based on the commutative property of addition, and they learn that this property is a fundamental law of addition (on whole numbers). The primary way participants learn mathematics in the CGI PD is through in-depth study and discussion of students' mathematical thinking.

In the first year of the three-year TDG CGI program, the learning objectives for participants focus on their learning the problem-type taxonomies for all four operations, the base-ten number system, and computation with large numbers. Participants write problems that match these frameworks and practice posing these problems to their students. A goal is that teachers, at least sometimes, will encourage students to use their own strategies for solving these problems rather than showing students strategies that they are expected to use. Participants also develop a general understanding of the different levels of solution strategies used by students in solving mathematics problems, and they develop their skills in questioning students to gather additional information about their thought processes.

The present report focuses on the first year of the program, but readers curious about how the subsequent years compare with the first year should refer to Tazaz and Schoen (2020) for further description of the first and second years of the program.

3.2.4. Qualifications of Professional Development Facilitators

To lead a TDG CGI workshop, an individual must meet a minimum set of requirements and become certified by TDG to facilitate workshops. Certification to lead workshops begins with individuals' becoming certified to teach CGI year 1. As they become more experienced and complete further training, they become eligible to teach years 2 and 3. The requirements for certification as a CGI year 1 facilitator are

- Have a strong understanding of the CGI frameworks (e.g., problem types, solution strategies, relationship between problem types and solution strategies).
- Have at least 5 years of experience with CGI in one or more of the following ways:
    - actively implementing CGI as a classroom teacher.
    - actively supporting/implementing CGI as a math coach working with expert CGI teachers.
    - actively supporting/implementing CGI as a CGI researcher working closely with expert CGI teachers.
- Have at least 3 years of experience leading CGI PD for teachers in their own communities.
- Be able to recognize the formal mathematical concepts embedded in children's intuitive strategies.
- Be able to design a problem in real time that would engage children with a particular property within a particular number domain.
- Have strong pedagogical skills when working with adult learners.

Three facilitators delivered the workshops to the five main cohorts of participants. Each of the facilitators had at least 10 years of experience either as a CGI classroom teacher, as a mathematics coach working with expert CGI teachers, or as a CGI researcher working closely with expert CGI teachers. Each facilitator had at least 5 years of experience leading CGI PD with teachers in their own communities. All three of the facilitators met the educational criteria necessary to administer TDG CGI workshops and were certified by Linda Levi, Director of CGI Initiatives for TDG.

Facilitator 1 and 2 independently led the workshops for two cohorts each; Facilitator 3 led the workshops for only one cohort. Facilitator 1 provided two additional 4-day summer make-up workshops. The participants in those make-up workshops were then added to the cohorts containing colleagues from their schools before the beginning of day 5 of the program.

The program developer (TDG) prepared an agenda and an implementation plan for each workshop day that was split into 15–60 minute segments, specifying the content to be covered, the delivery method, the materials to be used during the training, and predicted responses from participants. In the first year of the program, this plan was approximately 45 pages: 24 pages for the four days of summer workshops, 11 pages for the first two follow-up workshop days, and 10 pages for the last two follow-up workshop days.. The timing, sequence, and emphasis of each activity in the implementation plan is expected to be adapted by the workshop leader to meet the needs of the participants in the cohort if the workshop leader determines it is necessary. To ensure the full access to requested materials at every training session, TDG made a list of materials and provided the daily handouts needed by all cohorts, and the research staff printed and delivered these materials and handouts to the PD sessions.

3.3. Implementation of the CGI Professional Development Workshops

3.3.1. Workshops Offered to Teachers in the CGI Intervention Condition

To measure the degree to which the workshops were implemented as intended, trained research project personnel observed every workshop day for all five cohorts. On each of the eight days of PD, an observation protocol tailored to the daily activities was created and used to record data describing the extent to which (a) daily activities were completed, (b) modifications to the agenda were made, and (c) teachers' homework was assigned by the workshop leaders. The observations focused on eight dimensions of the PD: learning about teaching mathematics for understanding; learning the CGI framework; watching student/classroom video; time spent reflecting on material; planning for teacher/student interactions; interviewing students; and observing teaching. The observers measured the degree to which each facilitator's plan was implemented; they did not measure or judge the quality of the delivery.

Figure 3.1 summarizes the data on the average amount of planned activities completed during year 1 of the PD program. On average, across all five cohorts, the facilitators delivered approximately 83% of the intended program.

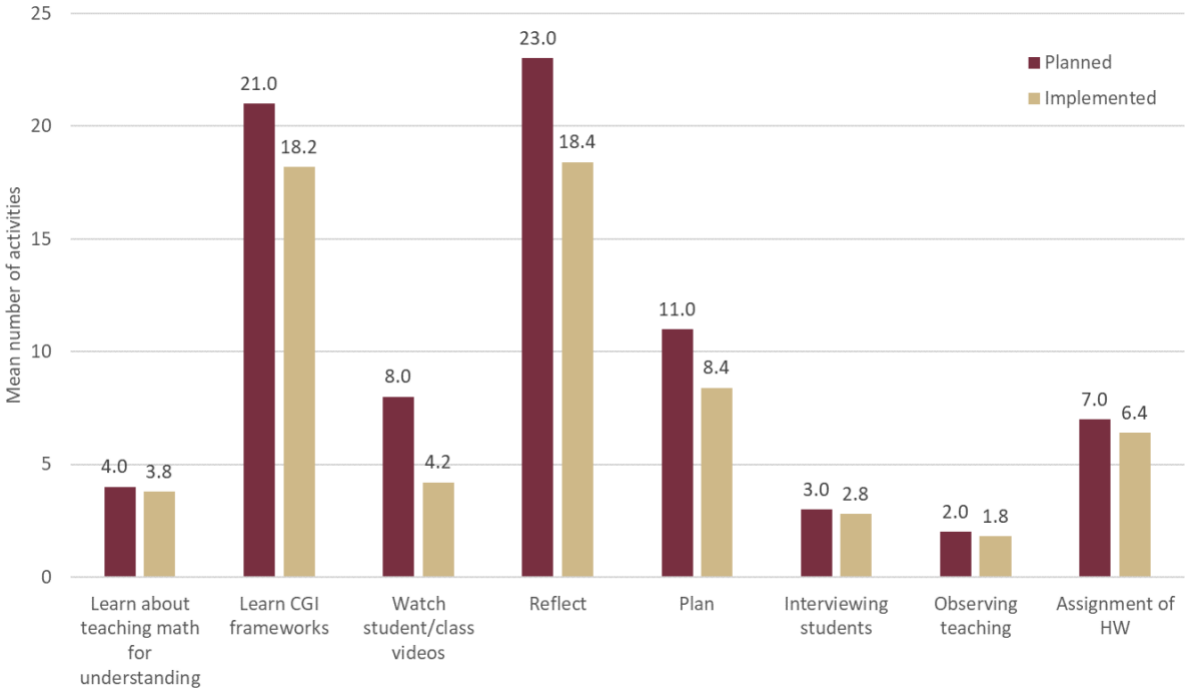


Figure 3.1. Total number of planned CGI activities and average number of implemented activities completed in year 1 of the program.

3.3.2. Teacher Attendance at the CGI Workshops

Observers determined attendance by scanning participants' nametags as they entered and left the training sessions. Paper-based sign-in sheets were also used. In instances where electronic attendance logs had missing or unusual data, physical sign-in sheets were used. The total number of hours attended by participants is reported in Table 3.2. The majority of participants in the treatment-condition schools attended more than 37 of the possible 48 hours of professional development offered between June 2013 and February 2014. Approximately 60% of the participants in the treatment-condition schools completed all 48 hours of PD offered. Six of the 12 participants who attended between 25 and 36 hours were from cohort 5 and missed the PD days when their facilitator was sick.

Five participants who attended between 13 and 24 hours of workshops comprised all the participants in one of the schools in the treatment group. These 5 participants only attended the four days of summer workshops (i.e., 24 hours of workshop time). They did not participate in any additional training during the school year but agreed to continue to participate in the study, including data collection for teachers and students at the end of the year. Fourteen of the 16 participants who attended between 0 and 12 hours did not participate in any of the four days of summer workshops as a result of personal scheduling conflicts.

Table 3.2. Cumulative Number of Hours Participants Attended CGI Workshops

	Total number of hours of CGI workshop attendance			
Condition	0–12	13–24	25–36	37–48
Intervention (N = 121)	16	10	12	83
Comparison (N = 115)	0	0	0	0

3.3.3. Reading Assignments

Part of the TDG CGI program involved asking participants to read relevant passages outside of the time they spent in workshops. The daily implementation-observation protocol included a section in which the observer documented the passages the participants in each cohort were expected to read. Such passages included chapters or sections of pages from *Children's Mathematics: Cognitively Guided Instruction* (Carpenter et al., 1999, hereafter *Children's Mathematics*) or *Thinking Mathematically: Integrating Arithmetic & Algebra in Elementary School* (Carpenter et al., 2003, hereafter *Thinking Mathematically*) and two journal articles: Behrend (2003) and Falkner et al. (1999). The readings typically followed topics introduced in the workshops. At the beginning of each PD day (days 2–8), the implementation observer asked each participant to report the extent to which he or she had completed the assigned readings. Table 3.3 summarizes the extent to which planned reading assignments were assigned by the workshop leaders to the cohorts and the proportion of all participants who reported completing the readings outside of the training sessions. In some instances data related to the completion of the assigned homework were incomplete from an entire cohort. In those cases, the data were recorded as a zero in the calculations for percentage completed, so the completion rates for those activities may be underestimated.

Table 3.3. Homework Readings Assigned and Completed by Participants in the Intervention Condition

Assigned reading	% of participants who completed readings
Children's Mathematics: Chapter 1	79.2
Children's Mathematics: Chapter 2	56.4[a]
Children's Mathematics: Chapter 3	69.3[a]
Children's Mathematics: Chapter 4	45.5
Children's Mathematics: Chapter 7	76.2
Journal articles	49.8[a]

Note. The assigned journal articles were Behrend (2003) and Falkner et al. (1999).
[a] When missing data were reported on cohort readings, data were recorded as a zero in the calculations for percentage completed, so these values may underestimate actual completion rates

Table 3.4 summarizes the percentage of participants in the cohorts who were assigned additional readings and the extent to which those participants completed the assigned readings. The additional readings were not part of the original implementation plan. These instances were recorded by the observer on the implementation observation sheet. In their reporting of homework completion, participants were also given an opportunity to report any additional readings they might have completed.

Table 3.4. Additional Homework Readings Assigned and Completed

Assigned reading	% of teachers assigned who completed additional readings
Children's Mathematics: Chapter 5	97
Children's Mathematics: Chapter 6	95
Thinking Mathematically: Chapter 1	35
Thinking Mathematically: Chapter 2	35
Thinking Mathematically: Chapter 3	35

3.3.4. CGI Team Meetings

An important component of the CGI program is the participants' opportunity to collaborate with colleagues in their schools in discussion of student thinking and associated instructional decisions. During the summer training sessions, participants were introduced to *CGI Team Meetings*. CGI Team Meetings involve the completion of the following three activities: (a) as a team, identify a mathematics problem that will give insight into student thinking with respect to learning goal(s), (b) pose the problem to their students in their individual classrooms, (c) as a team, discuss what they learned about student thinking through that problem.

As part of a task during the summer training institutes, the participants were assigned the task of holding a CGI Team Meeting cycle during the training session. Because the summer workshops occur outside of the school year, the activity was modified. Participants were asked to choose a problem based on learning goals identified by the participants. Rather than posing the problem to students and discussing the outcome, their assignment was to plan how to pose the problem and to come up with ideas on how they thought students would think about the problem. At the conclusion of this modified activity, teams were given a CGI Team Meeting Log sheet (not provided here, because it is copyrighted) and reminded to go back into their classrooms while school was in session, actually pose the problem, and meet with their teams to discuss what they observed students doing.

At two points in the year (day 5 and day 8), participants who attended the workshops were asked to report the extent to which they participated in CGI Team Meetings at their schools. During training day 5 and 8, attendees were given a CGI Team Meeting reflection sheet and asked to review their Meeting Logs and reflect on the extent to which they participated in CGI Team Meetings since their last training session. If they participated in any meetings, they were asked to reflect on and report the extent of the discussion and the teachers with whom they collaborated. These reflection sheets were then collected and analyzed by the evaluation team for the extent to which each attendee participated in a formal or informal CGI Team Meeting.

Table 3.5 summarizes the extent to which participants reported participating in CGI Team Meetings on days 5 and 8. The *formal* category indicates cases in which all three aspects of a CGI Team Meeting were reported. The *informal* category indicates cases in which at least one of the aspects of a CGI Team Meeting was reported or when the participant reported informal discussion of CGI with colleagues. *None* indicates cases in which participants reported conducting neither official CGI Team Meetings nor similar informal meetings. Participation in formal CGI Team Meetings increased between days 5 and days 8, but the total participation levels by day 8 were only approximately one-third. In some instances, participation in CGI Team Meetings was unreported. The cohort most affected was the one that was canceled because of illness; 100% was unreported on day 8.

Table 3.5. Percentage of Participants Indicating They Participated in CGI Team Meetings

Team meetings	Percentage
Day 5	
Formal	14
Informal	69
None	3
Unreported	14
Day 8	
Formal	35
Informal	31
None	0
Unreported	34

3.3.5. Teachers' Perceptions of the Quality of the Professional Development Program

The participants were asked to evaluate the quality of the professional development anonymously on a scale of one to five, one indicating "poor" and five indicating "excellent." Days 4, 6, and 8 were selected as evaluation points, because they comprised the last day of each series of workshops before the participants returned to the classroom to implement the ideas addressed in the workshops. Asking the participants to provide their evaluation of program quality of the program provided useful feedback for the individual workshop leaders and the Director of CGI Initiatives at TDG so that they could ensure participants' perspectives were considered and incorporated into the program. Table 3.6 and Figure 3.2 provide the summary of the teacher-reported professional-development quality at the three evaluation points. The data clearly indicate that the participants considered the program to be of high quality throughout the year.

Table 3.6. Participants' Evaluation of the Professional Development Program

Evaluation point	*n*	*M*	*SD*	Min	Max
Day 4 session total	111	4.5	0.6	3	5
Day 6 session total	85	4.6	0.5	3	5
Day 8 session total	57[a]	4.6	0.6	3	5

Note. Responses were registered on a scale from 1 to 5, 1 indicating "poor" and 5 indicating "excellent."
[a]PD evaluations were not reported for Cohort 5 on day 8 because of cancelation of the PD session.

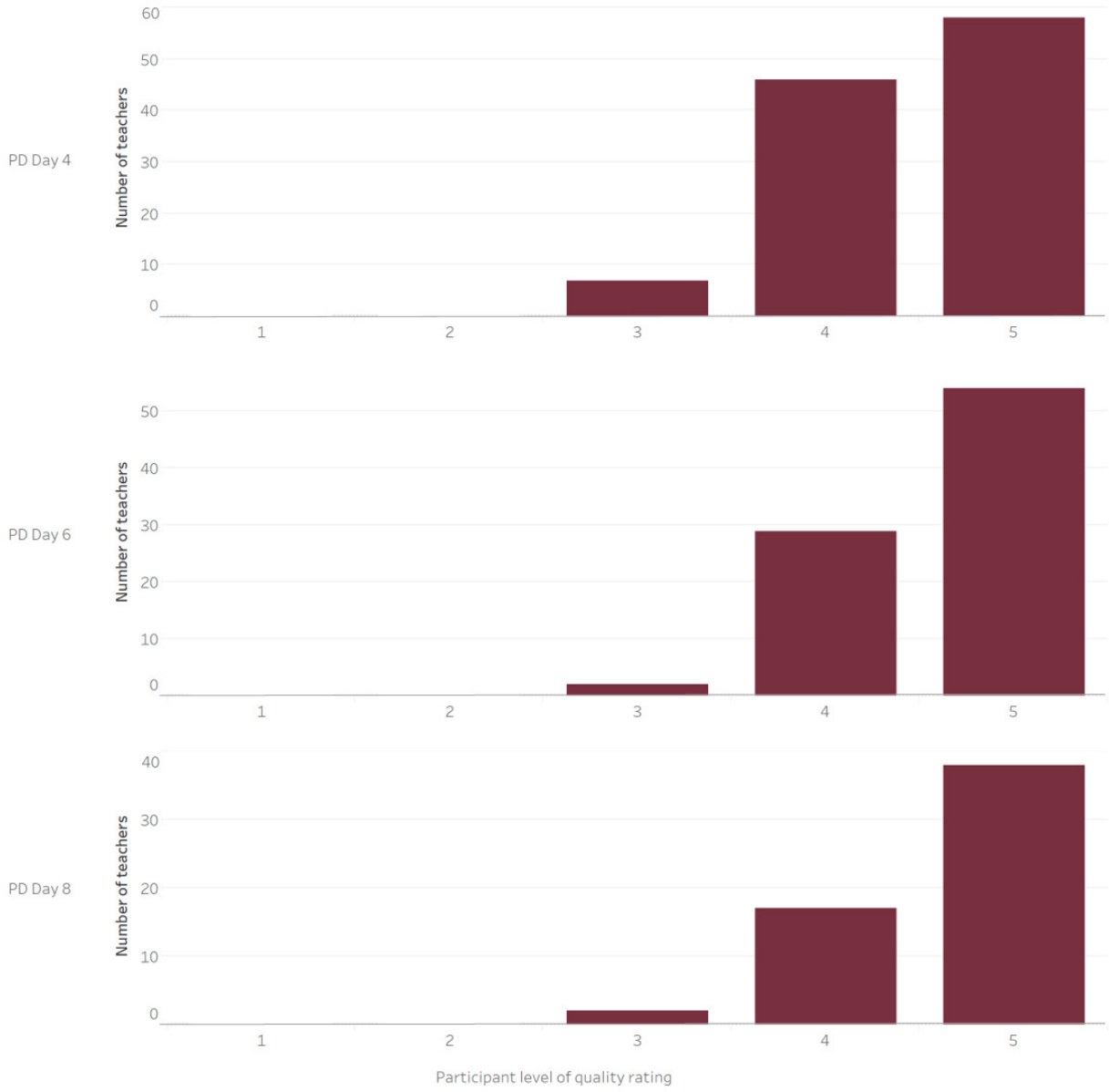


Figure 3.2. Participants evaluation of the professional development on a 5-point scale, 1 representing "poor" and 5 representing "excellent."

3.3.6. Professional Development Hours Reported by Teachers in the Treatment and Comparison Conditions During the Intervention Year

To end this section, we provide a glimpse into the differences in number of PD hours experienced between the teachers in the treatment-condition and counterfactual-condition schools during the intervention year. Teachers are expected to participate in PD programs as part of their continuing education experience. Each participant was therefore asked to respond to several online survey questions asking about their PD experiences during the 2013-14 school year. The following table draws from responses to the prompt: "*From June 2013 through May 2014, approximately how many hours of*

professional development did you participate in for each of the following subject areas?" Subject areas listed were: Math, Reading/Language Arts, Science, and Social Studies. Data reported here are drawn from responses from classroom teachers only; responses from math coaches, etc. are excluded. Of the 185 participating classroom teachers in the sample, 172 responded.

Table 3.7 presents descriptive statistics for the reported number of PD hours experienced by participants during the summer 2013 and the 2013–14 school year, disaggregated by subject area, treatment condition, and district. Values are reported in hours.

The data suggest that classroom teachers participated in more hours of PD related to Mathematics and Reading/Language Arts than in Science and Social Studies. In the comparison condition, PD in Mathematics and Reading/Language Arts appear to be nearly equal, but in the treatment condition, PD hours in Mathematics appear to be increased in proportion to the number of hours that the research study provided. The total number of PD hours in Reading/Language Arts appears to be higher in the treatment-condition schools than in the comparison-condition schools in both districts. We do not know why, but perhaps treatment-condition teachers reported more hours of Reading/Language Arts PD, because they see certain aspects of the CGI program, particularly the strong emphasis on word problems and comprehension, as supporting Reading/Language Arts.

Table 3.7. Descriptive Statistics for Reported Number of 2013–14 Professional Development Hours per Subject Area, Split by Treatment Condition and District

	Condition		
	Treatment	Comparison	Overall sample
Subject area	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)
*District A (*n *= 52 Treatment;* n *= 70 Comparison)*			
Mathematics	64.16 (68.10)	20.77 (42.01)	39.27 (58.52)
Reading/Language Arts	40.06 (66.36)	23.66 (49.64)	30.65 (57.68)
Science	5.16 (14.63)	4.59 (15.43)	4.83 (15.04)
Social Studies	0.89 (2.60)	0.70 (2.97)	0.78 (2.81)
*District B (*n *= 32 Treatment;* n *= 18 Comparison)*			
Mathematics	49.48 (31.36)	6.28 (3.27)	33.93 (32.63)
Reading/Language Arts	30.14 (40.50)	8.72 (5.13)	22.43 (33.98)
Science	3.35 (9.50)	8.11 (28.02)	5.07 (18.29)
Social Studies	0.51 (1.85)	0.50 (1.47)	0.51 (1.70)
*Districts A and B combined (*n *= 84 Treatment;* n *= 88 Comparison)*			
Mathematics	58.57 (57.17)	17.81 (37.90)	37.71 (52.29)
Reading/Language Arts	36.28 (57.81)	20.61 (44.68)	28.26 (51.96)
Science	4.47 (12.88)	5.31 (18.56)	4.90 (16.00)
Social Studies	0.75 (2.33)	0.66 (2.73)	0.70 (2.54)

Note. *n* = 172 responses out of 185 participating classroom teachers.

Further information about implementation is available in a report focused on that aspect of the intervention program (Tazaz & Schoen, 2020).

4. Analytical Approaches

The first phase of investigation in the current study was to conduct confirmatory analyses of the main effects of the CGI program on student achievement in mathematics. Next, we conducted two sensitivity analyses investigating the extent to which (a) the estimated effects of the CGI program were sensitive to whether the analytic sample was composed of all students measured at follow-up or was constrained to only students who were early joiners and (b) the estimated effects of the CGI program were sensitive to whether a Bayesian or likelihood-based method of estimation was used for the analyses. Last, we conducted exploratory analyses investigating variation in the size of the CGI program effect by student characteristic. All analyses except the last were conducted independently by external data analysts, who confirmed all results found and reported in the present document.

Separate models were fit for each student outcome measure: the MPAC interview, ITBS Math Problems (ITBS–MP) test, and ITBS Math Computation (ITBS–MC) test. By means of Mplus version 7.4 (Muthén & Muthén, 1998–2012), data were fit to multilevel models with random effects for classroom and school clusters. Independent variables consisted of indicators for the randomization blocks, student grade level, school treatment condition, student demographics, and baseline mathematics achievement.

Baseline student mathematics achievement was modeled by decomposition of the baseline test into levels of clustering to form three latent variable covariates from the observed student-level scores. As described by Asparouhov and Muthén (2007; see also Example 9.1b in Muthén & Muthén, 1998-2012) and evaluated by Lüdtke et al. (2008), this modeling technique creates a latent variable at each level of clustering with an implied group mean centering at the within level and cluster means at each between level. Muthén and colleagues note that using a latent covariate may be preferred to an observed covariate, because it avoids biased estimation associated with low reliability of observed cluster means.

The inclusion of covariates in our analyses achieved the dual purpose of (a) improving precision of the estimated treatment effect and (b) adjusting the estimates to compensate for any lack of balance between conditions at baseline. Moreover, the inclusion of covariates can increase the precision of a treatment effect by explaining extraneous variance in the outcome, reducing the standard error of the impact estimator, and correspondingly producing a lower *p*-value for the treatment coefficient (Bloom, 2005; Raudenbush, 1997). In the presence of baseline imbalance on covariates that correlate with the outcome, however, the inclusion of such covariates can increase or decrease the point estimate for the effect size. For example, the Kahan et al. (2014) review of risks and rewards of covariate adjustment in randomized trials presents an instance where covariate adjustment reduced the estimated effect by as much as 38%. Kahan and colleagues comment:

> Randomization ensures that, on average, both known and unknown covariates are well balanced between treatment conditions. However, randomization does not guarantee balance; in any individual trial, there may be large imbalances in important prognostic covariates between treatment conditions merely by chance. Any such imbalance can give an unfair advantage to one treatment condition over another if not accounted for in the analysis (p. 2).

Further, concern over balance of covariates at baseline may be of even greater concern for cluster-randomized trials than for individually randomized trials because of the varying compositional profiles of assigned units and difficulty in achieving balance on all relevant characteristics (Ivers et al., 2012).

Inclusion of covariates for student-level baseline achievement as well as classroom- and school-mean achievement at baseline has the advantage of controlling for related contextual effects, if they are present (Raudenbush & Bryk, 2002). In the present case, wherein we adjusted for baseline achievement,

contextual effects would take the form of effects on student-level achievement that resulted from the environments in lower- and higher-achieving classrooms or schools, above and beyond the effect of the respective individual's prior achievement. Castellano et al. (2014) used average socioeconomic status (SES) to illustrate how contextual effects might arise in education, noting that the process can be psychological and sociological.

> The psychological, or opportunity to learn, explanation, is that the average SES of a school may affect the style of instruction of its teachers which in turn affects individual achievement. The sociological, or normative climate, explanation is that the average SES level creates a climate that affects the individual student's motivation to learn and hence affects his or her individual achievement level. (p. 350)

Although the estimation of contextual effects was not of substantive interest for the current analyses, we attempted to improve the precision of the estimate of the effect of treatment by controlling for potential contextual or peer effects (see, e.g., Sacerdote, 2001).

In summary, we include covariates in the statistical model for the following three purposes:

1. Improve precision and reduce p-value by explaining extraneous variance in the outcome variables.
2. Compensate for lack of balance at baseline from the randomization process.
3. Control for contextual effect.

Our baseline model includes covariates for randomization blocks and grade level to account for structural aspects of the data. We used a model-building procedure comprising a sequence of nested models, with covariates added in succession. We compared across results of these models as a robustness check to understand whether the estimated effects of treatment are robust to covariate adjustment.

All analyses in the current study used Bayesian estimation, with the exceptions of the sensitivity analyses that compare results for Bayesian and likelihood-based estimation. The choice of a Bayesian approach was made after careful review of research literature on multilevel modeling when the number of clusters is relatively low (e.g., McNeish & Stapleton, 2014). An advantage of Bayesian estimation is that it does not carry the frequentist asymptotic assumptions of the number of clusters converging to infinity (Gelman, et al., 2013; Raudenbush & Bryk, 2002).

Successful model convergence for the Bayesian models was judged according to the criteria of (a) Gelman-Rubin Potential Scale Reduction (PSR) and (b) absence of evidence of discrepancy between posterior distributions in the different Markov chain Monte Carlo (MCMC) chains indicated by failure to reject the equality of posterior distributions in different chains by the Kolmogorov-Smirnov (KS) distribution test (Kaplan & Depaoli, 2012; Muthén & Muthén, 1998–2012). With all models specified with a fixed 100,000 iterations, the PSR criterion was satisfied if the value fell below and stayed below 1.05 for more than half of the iterations. Given the complexity of these models, a KS test p-value < .001 was used as the criterion for test rejection. Bayes estimated models were specified with two processors and the default two MCMC chains. All Bayesian parameter estimations used Mplus default noninformative priors (Asparouhov & Muthén, 2010a, 2010b).

Hedges' g effect sizes were used for estimation of group differences between treatment conditions. The effect-size estimates were calculated in accordance with WWC guidelines (U.S. Department of Education, 2013), where the effect size was calculated as the regression coefficient divided by the unadjusted pooled within-group standard deviation (based on the total variance summed across within- and between-cluster variance estimates), multiplied by the correction for small sample size.

Determination of substantive importance of effects follows the WWC criterion of ≥ 0.25, where the handbook stated, "Effect sizes at least this large are interpreted as a qualified positive (or negative) effect, even though they may not reach statistical significance in a given study" (p. 23).

4.1. Confirmatory Analyses

The confirmatory analyses were designed to determine the effect of the first year of the CGI teacher professional-development program on grade 1 and 2 student achievement as measured by the MPAC, ITBS Math Problems, and ITBS Math Computation after the first year of implementation of the program.

The analytic sample for the confirmatory analyses investigating the main effects of the CGI program on student mathematics included all participating students measured at follow-up (i.e., both early and late joiners), pooled across grades 1 and 2. Model building for the main-effects analyses comprised a sequence of three nested models, with covariates added in succession. Model 1 included school treatment assignment as the key predictor of interest, controlling only for student grade-level and randomization block. Model 2 included all variables from Model 1, plus covariates for student demographic characteristics of gender, free/reduced-price lunch status, English language learner status, and disability status. Model 3 included all variables from Models 1 and 2, plus covariates for student baseline mathematics performance, classroom-mean mathematics performance at baseline, and school-mean mathematics performance at baseline. Table B.1 in Appendix B provides a description of analytic models for the main effects analyses.

4.2. Sensitivity Analyses

The sensitivity analyses addressed two areas of investigation: sensitivity of the treatment effect to (a) how we define the analytic sample and (b) our choice of estimator.

4.2.1. Treatment Effect Sensitivity to Analytic Sample Definition

The first sensitivity analysis was guided by the following question. Are the estimated effects of the CGI program on student achievement outcomes after the first year of the program sensitive to whether the sample includes all students measured at follow-up or are they constrained to those students who were early joiners in the respective school in which they were measured at follow-up?

The analytic sample for the sensitivity analyses constituted participating students who contributed outcome data at spring 2014 follow-up for the same school in which they were enrolled fall 2013 (i.e., early joiners), pooled across grades 1 and 2. Model building for the sensitivity analyses was identical to that for the models used in the main effects analyses. The sensitivity analyses therefore model the main effects of the CGI program on the early-joiner sample for the purpose of inspecting the discrepancy between those estimates and the main effects when data from the early- and late-joiner samples were used.

4.2.2. Treatment Effect Sensitivity to Method of Estimation

The second sensitivity analysis was intended to determine whether the estimated effects of the CGI program after the first year of implementation were sensitive to whether a Bayesian or a likelihood-based method of estimation was used for the analyses. This sensitivity analysis was motivated by the research literature's caveats on multilevel modeling with samples comprising few cluster units and our concerns as to whether the estimates of treatment effects were stable across estimation methods.

Although the sample of 22 school clusters in the current study is on the low-end of what is recommended for multilevel analysis, Maas and Hox (2005) found the bias associated with having a small number of groups to be a problem primarily when group-level variance components are estimated. Motivated by Snijders and Bosker's (1999) assertion that multilevel modeling becomes plausible when the number of groups exceeds 10, Maas and Hox (2005) found that the estimation of unbiased regression coefficients and of their standard errors is tenable with sample sizes as small as 10 groups of five units, when the research focus is on estimating fixed effects.

For all models in these sensitivity analyses, we used the Mplus ML maximum likelihood with conventional standard errors estimator, as opposed to the default MLR maximum likelihood with robust standard errors estimator. Although the MLR estimator provides the advantages of accounting for nonnormality in outcomes and is assumed to be robust to unmodeled heterogeneity, analyses by Hox et al. (2010) of estimation methods in multilevel structural equation modeling concluded that MLR is more accurate than ML only when the number of groups is sufficiently large—suggesting, for some models, that a sample size of 50 groups would be sufficient when ML is used but that up to 200 groups would be needed for MLR to perform optimally.

4.3. Exploratory Analyses

The exploratory analyses addressed two areas of investigation: (a) the size of the CGI program effect on student subgroups and (b) the moderation of the effects of the CGI program by student characteristics. The student characteristics we explored were grade level, gender, race/ethnicity, eligibility for free or reduced-price lunch, classification as English-language learner, and classification as having a disability.

Two analytic phases were employed in the exploratory analyses. First, the sample was disaggregated by student characteristic for subgroup analyses. Second, models were fit to the aggregate sample with treatment-by-student-characteristic interactions included. The analytic sample for all exploratory analyses drew from the early- and late-joiner sample.

As recommended by Tanniou et al. (2016), we specify the purpose of these analyses to stem from a research interest in investigating the consistency of the treatment effect across subgroups if an effect was detected, and in the event of an overall nonsignificant trial, in exploring the treatment effect across different subgroups.

Although segmenting the sample into subgroups reduces the size of the analytic sample and probably reduces the statistical power to find an effect if one is present, it also increases the probability of false discovery resulting from conducting multiple comparisons with various subgroups (Brookes et al., 2004). Rather than adjusting for multiplicity, we therefore place our interest in these exploratory analyses on the magnitude of the effect sizes, rather than on the statistical significance of the estimates.

We also note that caution is warranted when inferences are drawn from the subgroup analyses, because our sampling procedure did not stratify by student characteristic. Thus, although the random assignment procedure did form matched pairs of schools based on percentage of student membership eligible for free or reduced-price lunch (FRL), chance variation in other student characteristics resulted in some imbalance. For example, one matched pair of schools were both Provision 2 schools (i.e., 100% of school membership was FRL-eligible), even though the two schools differed by nearly 20% in percentage of minority students and a more than 10% in students classified as English language learners. Further, because not every category of student characteristic was obtained in the sample for each school, segmenting the sample by subgroup caused some schools or entire randomization blocks to fall out of the analysis. For example, the sample for the subgroup analysis of students who were not FRL-eligible

was reduced to only 16 of the potential 22 schools for the MPAC analysis and only 17 for the ITBS analyses, because some schools had no students in the sample who were not FRL eligible.

Thus, based on guidelines in the research literature on interpreting subgroup analyses for randomized controlled trials (e.g., Rothwell, 2005) and the presence of chance variation in student characteristics across the sample, we investigate the heterogeneity of treatment effects by inspecting results for the subgroup analyses and moderation analyses in tandem, treating the analysis-by-subgroup as a descriptive phase and the moderation analysis as the inferential phase of the analyses.

4.3.1. Treatment Effects on Student Subgroups

The subgroup analyses were intended to determine whether the CGI program affected student mathematics achievement after the first year of the program for subgroups of the student sample as identified by grade level, gender, race or ethnicity, eligibility for free or reduced-price lunch, classification as English-language learner, or classification as having a disability.

Each analytic sample for the subgroup analyses was disaggregated by the subgroup of interest. The analytic model for the subgroup analyses used model specifications analogous to those of Model 3 for the main effects analyses—except that the covariate corresponding with the respective subgroup was excluded from the model. Also, for the grades 1 and 2 subgroup analyses, only the baseline test corresponding to the respective grade level was modeled. Summaries of results for subgroup analyses are provided in Table 5.2, and detailed results tables are provided in Appendix G.

4.3.2. Moderation of Treatment Effects by Student Characteristic

The exploratory analyses addressed the question of whether the effect of the CGI program after the first year of the program differs according to student baseline characteristics.

The analytic model for the moderation analyses used Model 3 (see section 4.1 for an explanation of Model 3) as the model for the main-effects analyses for each of the three outcome metrics. Then, according to a separate model for each student characteristic, the slope for the given outcome on the respective predictor variable was specified to vary randomly across clusters, and treatment was specified as a predictor of the school-level variation around the respective slope. Prediction of the random slope by treatment constitutes a cross-level interaction that indicates whether and by how much the effect of treatment varied across levels of the predictor. The analytic sample for most moderation analyses was the aggregate sample, the exception being the treatment-by-baseline-achievement interaction model, which was conducted on each grade level separately. Another difference for the treatment-by-baseline-achievement interaction model was that cluster-level latent means for student baseline test were not included as covariates. Summaries of results for moderation analyses are found in Table 5.3, and detailed results tables are found in Appendix H.

4.4. Treatment of Missing Data

Our modeling approach estimated the means and variances for covariates, which brought the covariates into the model where missing data were assumed missing at random. Accordingly, our analytic models used all cases with data on the dependent variable, including cases with incomplete data for covariates. As described by Muthén, Muthén, and Asparouhov (2016), bringing a covariate *x* into the model (making it endogenous) by mentioning its variance expands the model to the joint distribution of *y* and *x* instead of the usual approach of *y* conditional on *x* (assuming nothing about the *x* distribution). This approach of bringing covariates with incomplete data into the model was applied for the variables indicating student

demographics and baseline achievement. The variance for the variables indicating grade was also estimated, so the covariance between all Level-1 covariates was free to vary.

Bringing covariates into the model afforded the opportunity to treat the binary covariates as categorical, but doing so introduced convergence difficulties for some models. The resulting need to treat covariates as normal was not problematic given results from Muthén et al.'s (2016) simulation study, which found the advantage of Bayes' treating binary covariates as categorical was less pronounced when the missing values did not exceed 20%. For analyses presented in the current report, missing data on binary covariates were less than 1% for all analyses.

Although the modeling procedure we used accommodated incomplete data when the missing data were for covariates, all analyses were restricted to cases with observed outcome data. Tables C.1 and C.2 in Appendix C present an analysis of patterns in missing data for confirmatory analyses. Subgroup analyses and moderation analyses employed case deletion when data were incomplete for the variable used to define the subgroup or when data were incomplete for the moderating variable.

4.5. Interpreting Bayesian Statistics

Within a frequentist approach, probability is conceptualized through a framework of frequency of repeated events. Within a Bayesian approach, probability is conceptualized through a framework of degree of uncertainty about values. As a result, the interpretation of a frequentist 95% *confidence* interval is if the study was replicated an inifinite number of times, 95% of those replications of the study would produce a band that containined the true parameter value. In contrast, the interpretation of the corresponding Bayesian statistic, a 95% *credibility* interval, is that, given the observed data and information in the prior distributions, the empirical value has a 95% chance of falling within that band. Succinctly, for frequentists, parameters are fixed (but unknown) and data are random; for Bayesians, parameters are random and data are fixed.

Given that investigators typically wish to know not just the parameter estimates for a finite sample but rather the true value of the parameter in the population, the assumptions that undergird a frequentist approach are compatible with the work of scientists, but given how frequentists define probability, a 95% confidence interval does not answer the question, "Is there at least a 0.95 probability that the parameter value is not zero?"; the Bayesian solution answers that question. Extending this concept to the frequentist *p*-value, a parameter with an estimate that is $p < .05$ is interpreted to mean that the chance of observing that or a more-extreme value was less than 5%, under the hypothesis that the true parameter value was zero. Accordingly, frequentist methods can be said to test the probability of the data, given that the (null) hypothesis is true, whereas Bayesian methods test the probability that the hypothesis is true, given the data.

As indicated in the recent American Statistical Association statement on *p*-values (Wasserstein & Lazar, 2016), compared to frequentist methods, Bayesian methods "more directly address the size of an effect (and its associated uncertainty) or whether the hypothesis is correct" (p. 132). For example, inspection of the Bayesian posterior parameter distribution for an outcome on treatment regression parameter allows for a direct assessment of the probability of a hypothesized effect given the data. Therefore, for a positive parameter point estimate, the proportion of the distribution above zero in a Bayesian posterior parameter distribution plot indicates the probability that the true parameter value is above zero (this quantity is the complement of the Bayesian one-tailed *p*-value). Any hypothesized parameter value of interest could be used, however, where, for a positive parameter point estimate, the proportion of the distribution above that point can be interpreted to indicate the probability that the true parameter value is at least that large. Moreover, in the current era of statistics where investigators are encouraged

to move away from bright-line rules that guide yes/no decisions on whether treatment effects were observed (e.g., Matthews et al., 2017; Wasserstein & Lazar, 2016), Bayesian methods represent a suitable tool for evaluating program effects (Lecoutre et al., 2001).

Differences between frequentist and Bayesian traditions in reporting findings include the Bayesian preference for referring to credibility intervals rather than *p*-values when parameters are evaluated. In addition, coefficient estimates in frequentist likelihood-based estimation have standard errors, but Bayesian estimates have posterior standard deviations. Tables presenting results from Bayesian analyses will therefore show posterior standard deviations rather than standard errors and 95% credibility intervals rather than *p*-values. Conceptual differences notwithstanding, interpretation of Bayesian credibility intervals follows the same logic as interpretation of frequentist confidence intervals, where the interest is whether the interval includes zero.

5. Impact of the PD Program on Student Achievement After the First Year of Implementation

As described in section 4, model building for the main-effects analyses involved a sequence of three nested models with covariates added in succession. The analytic model for the subgroup analyses used model specifications analogous to the full covariate model from the main-effects analyses, except that the covariate corresponding to the respective subgroup was omitted from the model. All analyses used the vertically scaled outcome scores, including the MPAC grade-level subgroup analyses. The analytic models for the moderation analyses used the full covariate model from the main-effects analyses as the baseline model, with the inclusion of a cross-level interaction intended to reveal whether the slope for student characteristics differed systematically by condition. Satisfactory convergence was demonstrated for all Bayesian models, as indicated by the PSR value's falling below and staying below 1.05 for more than half the fixed 100,000 iterations and the failure to reject the equality of posterior distributions in the different MCMC chains at $p < .001$ for the KS test.

In the following sections, we provide summary tables for the results of the impact of treatment conducted within the confirmatory, sensitivity, and exploratory analyses. These tables relay the effect size for treatment, but other model information such as the model coefficient, posterior standard deviation, and credibility intervals for the treatment parameter, and any information on covariates, has been omitted for visual simplicity.

A full reporting of model statistics appears in Appendix D (for confirmatory analyses), Appendices E and F (for sensitivity analyses), and Appendices G and H (for exploratory analyses). The tables in Appendices D, E, and F for the confirmatory and sensitivity analyses provide estimates for the fixed effects, the variance components, r-square, and intraclass correlations. Both portions of the tables report estimates from Models 1–3, but the lower portions also report estimates from Model 0. (See the first section in section 4 for information about the model-building procedure used in these analyses.)

Readers unfamiliar with Bayesian approaches to data analysis may wish to refer back to section 4.5 for a primer and to read the orientation we provide after each table in the current section.

5.1. Confirmatory and Sensitivity Analyses

5.1.1 Summary of Results of Confirmatory Analyses

Table 5.1 summarizes the estimated effect sizes for the main effect of treatment on each of the three measured outcomes in Model 3 (full covariate model absent any interaction terms) after the first year of program implementation. For all analyses, treatment was coded as 1 and comparison as 0. These effect-size estimates are based on the aggregate sample and use the vertically scaled scores for the outcome measures. Table A.5 in Appendix A provides descriptive statistics (sample size, mean, standard deviation) for the baseline tests, MPAC, and ITBS tests.

Table 5.1. Summary of Treatment Effects across Outcomes for the Confirmatory and Sensitivity Analyses

	MPAC		ITBS–MP		ITBS–MC	
Analysis	*N*	*g*	*N*	*g*	*N*	*g*
Confirmatory analyses						
Bayesian estimation with early and late joiners	622	0.08	2,172	0.03	2,172	−0.11
Sensitivity analyses						
Bayesian estimation with early joiners	622	0.08	2,120	0.03	2,120	−0.12
Maximum likelihood estimation with early and late joiners	622	0.10	2,172	0.02	2,172	**−0.11**

Note. MPAC = Mathematics Performance and Cognition interview; ITBS–MP = Iowa Test of Basic Skills Math Problems; ITBS–MC = Iowa Test of Basic Skills Math Computation. Multilevel regression models included school treatment assignment as the key predictor of interest, controlling for student grade-level; randomization block; and student demographic characteristics of gender, free/reduced-price lunch status, English language learner status, and disability status, as well as student baseline mathematics performance, classroom-mean mathematics at baseline, and school-mean mathematics at baseline. Boldface indicates the 95% credibility/confidence interval does not include zero.

The model results summarized in Table 5.1 show that, overall, little discrepancy in the results occurs when the early- and late-joiner sample is constrained to include only early joiners or when we use a likelihood-based estimator for the analyses. The effect-size estimates were positive in value for the two outcomes aligned with solving word problems and algebraic thinking and negative for the outcome focused on computational abilities involving the addition and subtraction operations. The point estimate of the treatment effect as measured by the MPAC interview is greater in magnitude than the point estimate for the ITBS Math Problems (ITBS–MP), though both are relatively small.

An inspection of the estimated effect across Models 1–3, summarized in Table 5.2 (see table note for description of models), reveals that the unadjusted treatment effect for the word problem and algebraic-thinking-oriented measures (MPAC, ITBS–MP) had a point estimate higher than the adjusted estimate for the effect of treatment. For example, in the confirmatory analyses, the size of the treatment effect as measured by the MPAC interview in Model 1—controlling only for grade level and randomization block—was $g = 0.20$. The effect size estimate was reduced to $g = 0.14$ in Model 2 after student demographic covariates were added. It was reduced further to $g = 0.08$ in Model 3 after baseline achievement covariates were included. Demonstrating a similar pattern, the effect size estimates as measured by the ITBS–MP for the confirmatory analyses were 0.10, 0.05, and 0.03 for Models 1, 2, and 3, respectively. Conversely, the negative effect-size estimate as measured by the ITBS–MC for the confirmatory analyses increased in magnitude after adjustment for covariates, increasing from −0.07 in Model 1 to −0.11 in Model 2, and remaining at −0.11 in Model 3.

As reported in Table 5.2, Bayesian parameter estimates and credibility intervals for the confirmatory analyses of the effect of treatment in Model 3 were $\gamma = 0.08$, 95% CI [−0.25, 0.42] for the MPAC interview; $\gamma = 0.03$, 95% CI [−0.17, 0.24] for the ITBS–MP; and $\gamma = -0.11$, 95% CI [−0.34, 0.11] for the ITBS–MC. The only place where conventional statistical significance (i.e., $p < .05$) was demonstrated in analyses of main effects was for the likelihood-based sensitivity analysis with the ITBS–MC test, where

the Model 3 coefficient for treatment of $\gamma = -0.11$ demonstrated conventional statistical significance at $p = .02$. No other coefficients for treatment in the confirmatory or sensitivity analyses had 95% confidence/credibility intervals that did not include zero.

Table 5.2. Summary of Treatment Effects across Different Models for the Confirmatory Analyses

	Model 1			Model 2			Model 3		
Outcome	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI
MPAC (*N* = 622)									
Treatment effect	0.195 (0.132)	0.20	[−0.074, 0.449]	0.142 (0.129)	0.14	[−0.119, 0.394]	0.083 (0.168)	0.08	[−0.253, 0.417]
ITBS–MP (*N* = 2,172)									
Treatment effect	0.103 (0.086)	0.10	[−0.063, 0.279]	0.047 (0.087)	0.05	[−0.119, 0.225]	0.034 (0.104)	0.03	[−0.171, 0.244]
ITBS–MC (*N* = 2,172)									
Treatment effect	−0.070 (0.093)	−0.07	[−0.252, 0.121]	−0.105 (0.091)	−0.11	[−0.280, 0.080]	−0.110 (0.113)	−0.11	[−0.338, 0.114]

Note. Model 1 comprised school treatment assignment, controlling for student grade-level and randomization block. Model 2 comprised all variables from Model 1, plus student demographic characteristics of gender, free/reduced-price lunch status, English language learner status, and disability status as covariates. Model 3 comprised all variables from Models 1 and 2, plus student baseline mathematics performance, classroom-mean mathematics at baseline, and school-mean mathematics at baseline. Effects for covariates are omitted for visual simplicity.

5.1.2. Interpreting the Summary Table of Confirmatory and Sensitivity Analyses

Table 5.1 presents the summary of treatment effects across outcomes for the confirmatory and sensitivity analyses. For each of the three outcomes—MPAC interview, ITBS–MP, and ITBS–MC—the size of the respective analytic sample and calculated Hedges' *g* effect size for the main effect of treatment is reported. The top panel of the table reports results for the confirmatory analyses (Bayesian estimation with early and late joiners); the bottom panel reports results for the sensitivity analyses (Bayesian estimation with early joiners and maximum likelihood estimation with early and late joiners). Collectively, the table reports results for nine separate analytic models.

5.2. Exploratory Analyses

5.2.1 Initial Subgroup and Moderation Analyses

The subgroup analyses were based on prespecified student characteristic groupings and were conducted on the early-and-late-joiners sample. The credibility interval for the subgroup analyses main-effect treatment parameter point estimates all included zero, as did the moderation analyses interaction parameter point estimates. Notwithstanding the absence of statistical significance, several effect-size estimates can be considered substantively important.

Table 5.3 presents a summary of effect size estimates for the three outcome measures, disaggregated by subgroup. For the subgroup analyses, separate models were fit to each outcome measure for each subgroup. Table 5.4 presents a summary of effect-size estimates for the analyses of moderated-treatment effects. For the moderation analyses, separate models were fit to each outcome measure with the aggregate sample; each treatment-by-subgroup interaction was tested in a separate model.

The effects reported in Table 5.3 indicate the main effects of treatment, pertaining to the specific subgroup to which the sample is constrained. The effects reported in Table 5.4 show the conditional effect of treatment and the treatment-by-student-characteristic interaction. The conditional effect of treatment in the moderation analyses indicates the effect size of treatment for students at the zero point of the moderating variable (i.e., the reference category for a categorical moderator or the point of centering for a quantitative moderator). The effect size for the treatment-by-student-characteristic interaction indicates the difference in magnitude between the effect of treatment for students in the reference category of the moderating variable and those in the focal category or the difference in magnitude in the effect of treatment between integers along the scale of the moderator.

Table 5.3. Summary of Treatment Effects across Outcomes on Subgroups

	MPAC		ITBS–MP		ITBS–MC	
Subgroup	*N*	*g*	*N*	*g*	*N*	*g*
Grade level						
Grade 1	336	0.25	1,103	0.14	1,103	0.03
Grade 2	286	−0.01	1,069	−0.07	1,069	−0.29
Gender						
Female	319	0.11	1,086	0.07	1,086	−0.06
Male	303	0.05	1,083	0.01	1,083	−0.15
Race/ethnicity						
Nonminority	207	0.15	803	0.05	803	−0.03
Minority	412	0.07	1,356	0.03	1,356	−0.10
Free or reduced-price lunch						
Not eligible	238	0.23	848	0.02	848	−0.16
Eligible	381	−0.03	1,311	0.05	1,311	−0.08
English language learner						
Non-ELL	479	0.07	1,667	0.08	1,667	−0.10
ELL	140	0.12	492	−0.07	492	−0.00
Student with disabilities						
Non-SWD	579	0.11	2,001	0.03	2,001	−0.11
SWD	40	0.36	158	−0.15	158	−0.27

Note. MPAC = Mathematics Performance and Cognition interview; ITBS–MP = Iowa Test of Basic Skills Math Problems; ITBS–MC = Iowa Test of Basic Skills Math Computation. Multilevel regression models included school treatment assignment as the key predictor of interest, when student grade level, randomization block, student demographic characteristics of gender, free/reduced-price lunch status, English language learner status, and disability status, as well as student baseline mathematics performance, classroom-mean mathematics at baseline, and school-mean mathematics at baseline. The covariate corresponding with the respective subgroup was excluded from the model for the analysis of that particular subgroup. For the Grade 1 and 2 subgroup analyses, only the corresponding pretest was modeled.

5.2.2. Interpreting the Summary Table of Subgroup Analyses

Table 5.3 presents the summary of treatment effects across outcomes on subgroups. The sample size and Hedges' *g* effect size for the main effect of treatment is reported for each analytic sample disaggregated by subgroup as measured by the MPAC interview, ITBS–MP, and ITBS–MC. Because these models include no interaction term, the treatment coefficient is still interpreted as a main effect (constant or average effect), though pertaining to the specific subgroup to which the sample is constrained. Because the sample was successively disaggregated by grade level (grade 1 or grade 2), gender (female or male), race/ethnicity, free or reduced-price lunch eligibility, English language learner status, and student disability status, the table reports results for 36 separate analytic models.

Table 5.4. Summary of Moderated-Treatment Effects across Outcomes

	MPAC		ITBS–MP		ITBS–MC	
Moderation model	*N*	*g*	*N*	*g*	*N*	*g*
Grade level	622		2,172		2,172	
Treatment		0.23		0.11		−0.00
Treatment by Grade 2		−0.27		−0.14		−0.18
Gender	622		2,169		2,169	
Treatment		0.09		0.06		−0.09
Treatment by Male		−0.02		−0.05		−0.05
Race/ethnicity	619		2,159		2,159	
Treatment		0.06		0.03		−0.06
Treatment by Minority		0.06		0.01		−0.06
Free or reduced-price lunch	619		2,159		2,159	
Treatment		0.18		0.03		−0.13
Treatment by FRL		−0.21		−0.03		0.03
English language learner	619		2,159		2,159	
Treatment		0.10		0.08		−0.12
Treatment by ELL		−0.11		−0.15		0.09
Student with disabilities	619		2,159		2,159	
Treatment		0.09		0.04		−0.11
Treatment by SWD		−0.02		−0.03		−0.01
Grade 1 Pretest	336		1,025		1,025	
Treatment		0.20		0.15		0.05
Treatment by Baseline test		0.20		0.02		0.07
Grade 2 Baseline test	284		980		980	
Treatment		−0.01		−0.06		**−0.31**
Treatment by Baseline test		0.17		0.08		−0.03

Note. MPAC = Mathematics Performance and Cognition interview; ITBS–MP = Iowa Test of Basic Skills Math Problems; ITBS–MC = Iowa Test of Basic Skills Math Computation. Multilevel regression models included school treatment assignment as the key predictor of interest, controlling for student grade-level, randomization block, student demographic characteristics of gender, free/reduced-price lunch status, English language learner status, and disability status, as well as student baseline mathematics performance, classroom-mean mathematics at baseline, and school-mean mathematics at baseline (classroom and school baseline test means were omitted from the baseline test interaction models) . The slope for the covariate corresponding to the analytic subgroup was specified to vary randomly across clusters, and treatment was specified as a predictor of the school-level variation around the respective slope, constituting the treatment-by-student-characteristic interaction. Boldface indicates the 95% credibility interval does not include zero.

5.2.3. Interpreting the Summary Table of Moderation Analyses

Table 5.4 presents the summary of moderated-treatment effects across outcomes. For each model, the table presents the respective sample size, Hedges' *g* effect size for the conditional effect of treatment, and effect size for the treatment-by-student-characteristic interaction. Moderation models were fit for each of the six categorical student characteristics specified in the subgroup analyses, as well as the two

continuous covariates of grade-1 and grade-2 baseline test. The baseline test scores were approximately normally distributed. In all, the table reports results for 24 separate analytic models.

The effect of treatment in a moderated-treatment analysis is considered *conditional,* because the regression of the outcome on treatment is conditional on the value of the moderator (Aiken & West, 1991). The conditional effect of treatment indicates the effect size of treatment for students at the zero point of the moderating variable (i.e., the reference category for a categorical moderator or the point of centering for a quantitative moderator). The effect size for the treatment-by-student-characteristic interaction indicates the difference in magnitude between the effect of treatment for students in the reference category of the moderating variable and those in the focal category or the difference in magnitude in the effect of treatment between integers along the scale of the moderator. For example, for a categorical moderator such as grade level, a conditional effect of treatment of $g = 0.25$ and treatment-by-grade interaction of $g = -0.25$ would indicate a one-quarter of a standard deviation positive effect of treatment for students in the reference category (e.g., grade 1) and a zero effect of treatment for students in the focal category (e.g., grade 2). When a quantitative moderator such as baseline test is used, as an example, a conditional effect of treatment of $g = 0.20$ and treatment-by-baseline-test interaction of $g = 0.10$ would indicate a one-fifth of a standard deviation positive effect of treatment for students who performed approximately at the mean on the baseline test and a one-tenth of a standard deviation increase in the estimated effect of treatment for approximately each standard deviation above the mean of students who were at baseline.

5.2.4. Subgroup and Moderation Analyses by Grade level

The treatment effects by subgroup reported in Table 5.3 suggest that the positive effects in the aggregate sample as measured by the MPAC interview were driven by a positive effect on the grade 1 sample ($g = 0.25$), where a positive effect was not observed in the grade 2 sample ($g = -0.01$). The negative effects in the aggregate sample as measured by the ITBS–MC appeared to be driven by a negative effect in the grade 2 sample ($g = -0.29$), where a negative effect was not observed in the grade 1 sample ($g = 0.03$). We first evaluated the evidence of grade-level treatment effect heterogeneity by inspecting the posterior parameter distributions for the treatment parameters per grade level subgroup, then drew inferences based on parameter estimates in the treatment-by-grade moderation analyses.

Figures 5.1 and 5.2 present the Bayesian posterior parameter distributions for the outcome on treatment regression parameters for the subgroup analyses dissaggregated by grade level. Their inspection can assist in assessing the probability of an observed positive effect by the CGI program on the grade-1 students and a negative effect on the grade 2 students. Plots a, b, and c in Figure 5.1 show the grade-1 posterior parameter distributions for the outcome on treatment regression parameters for the MPAC, ITBS–MP, and ITBS–MC analyses, respectively. The treatment parameter posterior distribution for the grade 1 MPAC analysis displayed in plot a shows a disribution median of $\gamma = 0.25$ and posterior standard deviation of PSD = 0.24, which indicate that 85% of the distribution has a positive value and translates to a .85 probability that the treatment effect is a nonzero positive value. (Areas under the normal distribution were calculated with the online normal-distribution calculator found at http://onlinestatbook.com/2/calculators/normal_dist.html.) To extend this exercise to a hypothesized parameter estimate of $\gamma \geq 0.10$ (a small effect size but typical of those found in education research), it corresponds to a .73 probability that the true parameter value is 0.10 or more in magnitude for the grade-1 MPAC analysis. The treatment-parameter posterior distribution for the grade-1 ITBS–MP analysis displayed in Figure 5.1, plot b, shows an estimate of $\gamma = 0.13$, PSD = 0.17, which corresponds to a .78 probability that the treatment effect is a nonzero positive value. The .85 and .78 estimated probabilities of an above-zero parameter estimate for treatment in grade 1 as measured by the MPAC

interview and ITBS–MP , respectively, lend credibility to the inference that treatment had a positive effect on the problem-solving abilities of grade-1 students in the treatment-condition schools.

The use of Bayesian methods in the current report allows for an interpretation of parameter estimates that readers accustomed to likelihood-based inference may not be familiar with. As noted by Wasserstein and Lazar (2016), compared to frequentist methods, Bayesian methods "more directly address the size of an effect (and its associated uncertainty)" (p. 132). Inspection of the Bayesian posterior parameter distribution for an outcome on treatment regression parameter therefore allows for a direct assessment of the probability of a hypothesized effect. Accordingly, for a positive parameter point estimate, the proportion of the distribution above zero in a Bayesian posterior parameter-distribution plot indicates the probability that the true parameter value is above zero (this quantity is the complement of the Bayesian one-tailed p-value).

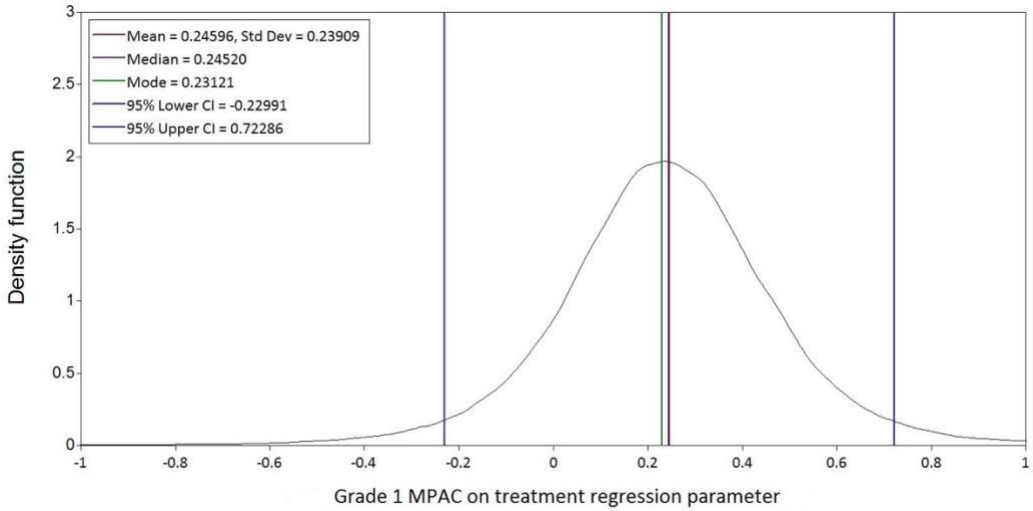
Plots a, b, and c in Figure 5.2 show the grade-2 posterior parameter distributions for the outcome on treatment regression parameters for the MPAC, ITBS–MP, and ITBS–MC analyses, respectively. The treatment parameter posterior distribution for the grade 2-ITBS–MC analysis displayed in plot c shows a disribution median of $\gamma = -0.29$ (the point estimate for the treatment effect) and posterior standard deviation of PSD = 0.15. Although the 95% credibility intervals include zero (95% CI [−0.59, 0.01]; reported in Appendix 5D), inspection of this distribution serves as a vivid example for why bright-line rules can be an impediment to sensible inference. In this case, a one-tailed evaluation of statistical significance indicates a .97 probability that the treatment effect is negative. Moreover, notwithstanding the absence of conventional statistical significance, Bayesian posterior parameter distributions for treatment in grade 2 as measured by the ITBS–MC lend credibility to the inference that treatment had a negative effect on the grade-2 sample in math computation.

Following this trail of potential treatment-effect heterogeneity leads to an inspection of the treatment-by-grade moderation analyses. Table 5.5 provides a detailed reporting of the parameters for the conditional treatment effect and treatment-by-grade moderation (additional detail of model parameters reported in Table H.1 in Appendix H).

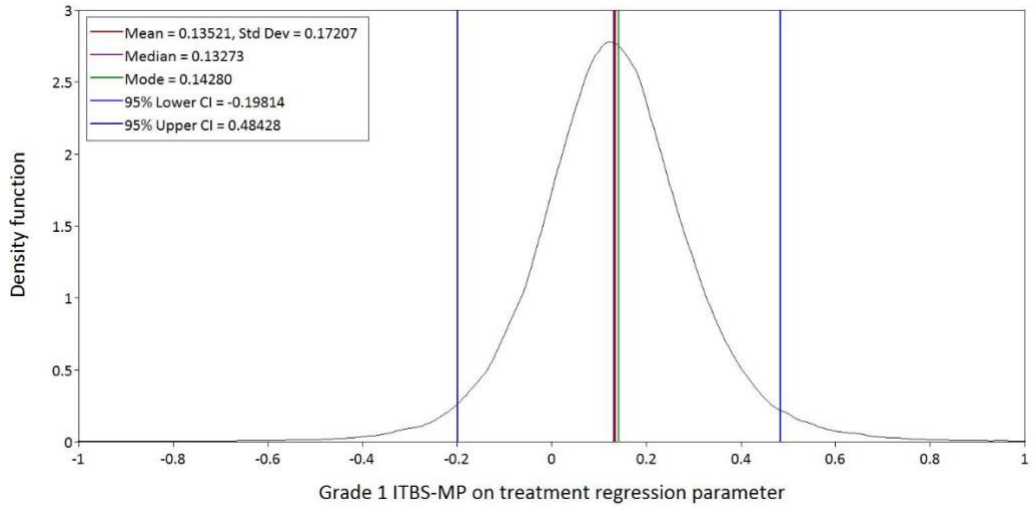
Although the estimated interaction terms were not statistically significant at the 95% credibility level for any of the models, evaluation of the parameter estimates can still yield insights into probability of program effect heterogeneity.

For the MPAC treatment-by-grade analysis, the treatment parameter of $\gamma = 0.23$, PSD = 0.19, corresponded to a .89 probability of a nonzero positive effect of treatment on the grade-1 sample, with the most probable estimate being one-fifth to one-quarter standard deviation in magnitude. The interaction parameter of $\gamma = -0.26$, PSD = 0.17, corresponded to a .94 probability of a nonzero difference in effect of treatment between grades. The magnitude and direction of the interaction term indicated an approximate one-quarter standard deviation difference in effect between grades, whereby the effect was estimated to be zero or slightly negative for grade 2 as measured by the MPAC interview.

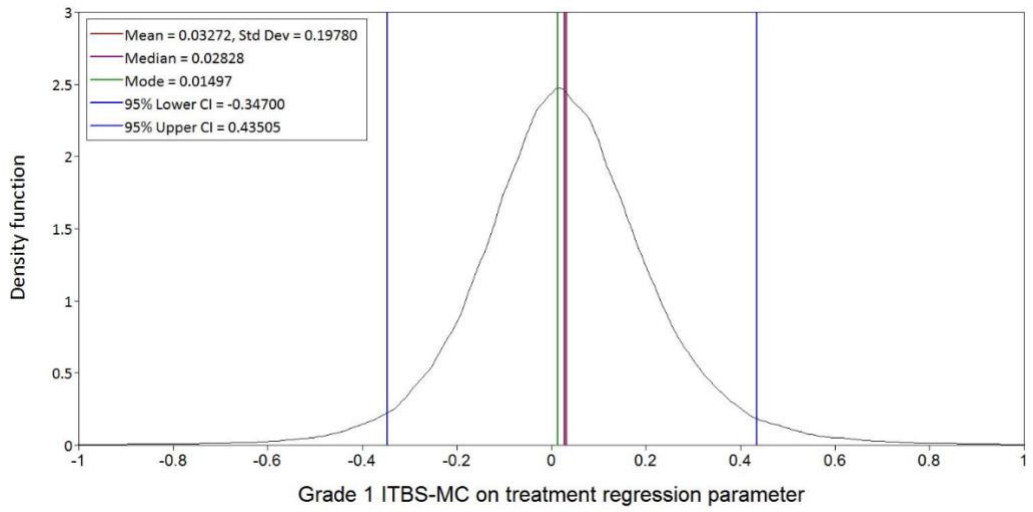
For the ITBS–MP treatment-by-grade analysis, the treatment parameter of $\gamma = 0.11$, PSD = 0.12, corresponded to a .82 probability of a nonzero positive effect of treatment on the grade 1 sample, with the most probable estimate being approximately a one-tenth standard deviation in magnitude. The interaction parameter of $\gamma = -0.14$, PSD = 0.12, corresponded to a .87 probability of a heterogeneity of effect between grades. The magnitude and direction the interaction term indicated a slightly larger than one-tenth standard deviation difference in effect between grades, whereby the effect was estimated to be zero or slightly negative for grade 2 as measured by the ITBS–MP.



(a)



(b)



(c)

Figure 5.1. Kernel density curves of the Bayesian posterior parameter distributions for the grade-1 outcome on treatment regression parameters.

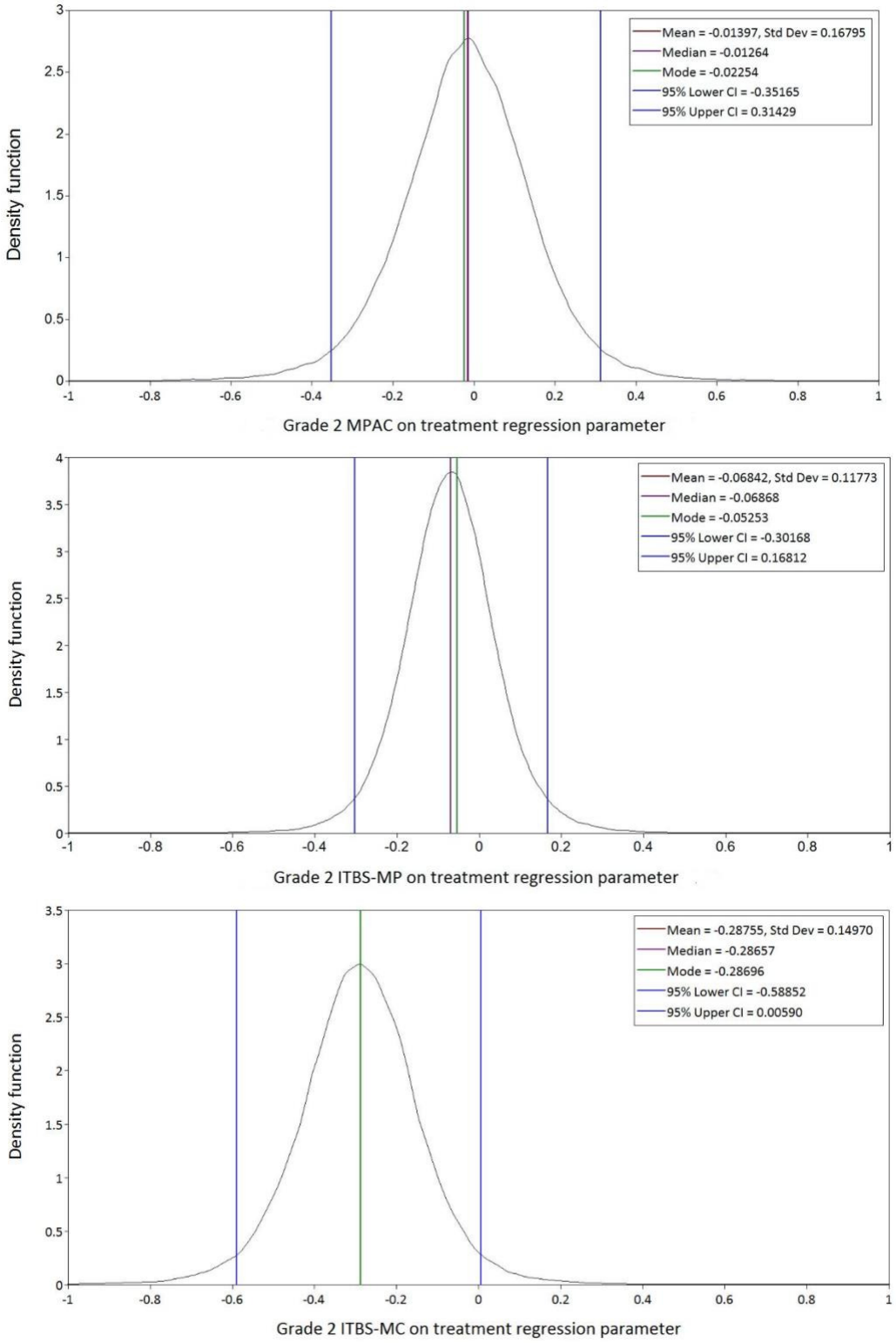


Figure 5.2. Kernel density curves of the Bayesian posterior parameter distributions for the grade-2 outcome on treatment regression parameters.

Table 5.5. Summary of Treatment-by-Grade Moderated Effects across Outcomes

	MPAC (*N* = 622)			ITBS–MP (*N* = 2,172)			ITBS–MC (*N* = 2,172)		
Parameter	Estimate (*PSD*)	*g*	95% CI	Estimate (*PSD*)	*g*	95% CI	Estimate (*PSD*)	*g*	95% CI
Treatment	0.230 (0.189)	0.23	[−0.154, 0.591]	0.111 (0.120)	0.11	[−0.119, 0.355]	−0.001 (0.118)	−0.00	[−0.232, 0.234]
Treatment by Grade 2	−0.263 (0.167)	−0.27	[−0.593, 0.065]	−0.136 (0.123)	−0.14	[−0.385, 0.104]	−0.180 (0.154)	−0.18	[−0.494, 0.118]

Note. Multilevel regression models included school treatment assignment as the key predictor of interest, controlling for student grade-level;, randomization block;, and student demographic characteristics of gender, free/reduced-price lunch status, English language learner status, and disability status, as well as student baseline mathematics performance, classroom-mean mathematics at baseline, and school-mean mathematics at baseline. The slope for grade was specified to vary randomly across clusters and treatment was specified as a predictor of the school-level variation around the grade slope, constituting the treatment-by-grade interaction.

Commensurate with results from the grade-1 and grade-2 subgroup analyses, the treatment-by-grade moderation analyses interaction estimates indicated a high probability of no effect of the program on the grade-2 sample, as measured by the MPAC interview and ITBS–MP, but the conditional effect found for the moderation analyses indicated a high to moderately high probability that the treatment effect in the grade 1 sample as measured by the MPAC interview and ITBS–MP was positive.

Likewise, the treatment-by-grade moderation analyses for the ITBS–MC revealed a conditional treatment effect parameter of γ = −0.00, PSD = 0.12, which corresponded to a .5 probability of a nonzero positive effect of treatment or probable effect of zero magnitude on the grade 1 sample. The corresponding interaction parameter of γ = −0.18, PSD = 0.15, corresponded to a .88 probability of a heterogeneity of effect between grades. The magnitude and direction of the interaction term indicated a nearly one-fifth standard deviation difference in effect between grades. With an estimated conditional effect of zero, the posterior parameter distribution for the interaction term indicated a moderately high probability that the treatment effect in the grade-2 sample as measured by the ITBS–MC was negative.

5.2.5. Subgroup and Moderation Analyses by Gender

Subgroup and moderation results for the analyses by gender were consistent in indicating a positive effect of treatment on the MPAC interview and ITBS–MP and a negative effect of treatment on the ITBS–MC. Effects appear relatively homogeneous across gender. According to the moderation analyses, results indicated estimated conditional effects of 0.09 for female students and −0.02 for male students on the MPAC interview, estimated conditional effects of 0.06 for female students and −0.05 for male students on the ITBS–MP, and estimated conditional effects of −0.09 for female students and −0.05 for male students on the ITBS–MC. The small sizes of these differences suggests any heterogeneity in the effect of treatment between female and male students is negligible—particularly given the exploratory nature of the analyses.

5.2.6. Subgroup and Moderation Analyses by Race/Ethnicity

Subgroup and moderation results for the analyses by race/ethnicity were consistent in indicating a positive effect of treatment on the MPAC interview and ITBS–MP and a negative effect of treatment on the ITBS–MC. Effects appear reasonably homogeneous with respect to minority and nonminority racial/ethnic groups. According to the moderation analyses, results indicated estimated conditional effects of 0.06 for nonminority students and 0.06 for minority students on the MPAC interview, estimated conditional effects of 0.03 for nonminority students and 0.01 for minority students on the ITBS–MP, and estimated conditional effects of −0.06 for nonminority students and −0.06 for minority students on the ITBS–MC. The small size of these differences—and lack of statistical significance—suggests any difference in the effect of treatment between nonminority and minority students is negligible.

5.2.7. Subgroup and Moderation Analyses by Economic-Disadvantage Status

Subgroup and moderation results for the analyses by economic-disadvantage status were consistent in indicating a positive effect of treatment on the MPAC interview for students not FRL-eligible, a small but negative effect of treatment on the MPAC interview for students FRL-eligible, a positive effect of treatment on the ITBS–MP for both groups, and a negative effect of treatment on the ITBS–MC for both groups. Except for the MPAC results, effects appear relatively homogeneous across the groups. According to the moderation analyses, results indicated estimated conditional effects of 0.18 for non-FRL-eligible students and −0.21 for FRL-eligible students on the MPAC interview, estimated conditional effects of 0.03 for non-FRL-eligible students and −0.03 for FRL-eligible students on the ITBS–MP; and

estimated conditional effects of −0.13 for non-FRL-eligible students and 0.03 for FRL-eligible students on the ITBS–MC. The small size of the differences for the ITBS analyses suggests any difference in the effect of treatment by economic status of students is negligible. The size of the difference for the MPAC analyses warrants attention, but we note that caution is warranted when inferences are drawn from these results, because six schools included no students in the sample who were not FRL-eligible. The imbalance in the sampling on that characteristic may therefore bias the results for this investigation.

5.2.8. Subgroup and Moderation Analyses by English-Learner Status

Subgroup and moderation results for the analyses by English learner status indicated small positive effects of treatment for non-ELL students on the MPAC Interview and ITBS–MP, small positive to small negative effects of treatment for ELL students on the MPAC Interview and ITBS–MP, and small negative effects for non-ELL and ELL students on the ITBS–MC. According to the moderation analyses, results indicated conditional effects of 0.10 for non-ELL students and −0.11 for ELL students on the MPAC interview, estimated conditional effects of 0.08 for non-ELL students and −0.15 for ELL students on the ITBS–MP; and estimated conditional effects of −0.12 for non-ELL students and 0.09 for ELL students on the ITBS–MC. The small size of the differences for the ITBS analyses suggests any difference in the effect of treatment between non-ELL and ELL students can be considered negligible.

5.2.9. Subgroup and Moderation Analyses by Disability Status

Subgroup and moderation results for the analyses by disability status demonstrated some inconsistencies. Discrepancies occurred on the MPAC Interview where a substantively important positive effect was estimated for the SWD subgroup, but only a small positive effect was estimated for SWD students with the moderation analysis. Similarly, on the ITBS–MC , a substantively important negative effect was estimated for the SWD subgroup, but only a small negative effect was estimated for SWD students with the moderation analysis. Subgroup and moderation results on the ITBS–MP were comparable.

According to the moderation analyses, results indicated conditional effects of 0.09 for non-SWD students and −0.02 for SWD students on the MPAC interview, estimated conditional effects of 0.04 for non-SWD students and −0.03 for SWD students on the ITBS–MP, and estimated conditional effects of −0.11 for non-SWD students and −0.01 for SWD students on the ITBS–MC. The small size of the differences for the ITBS analyses suggests that any heterogeneity in the effect of treatment between non-SWD and SWD students is negligible.

5.2.10. Subgroup and Moderation Analyses by Baseline Student Achievement

For all three outcomes in grade 1, the interaction term had a positive point estimate, albeit of a negligible size for the ITBS tests. Variation of the treatment effect on the grade 1 MPAC interview shown in Figure 5.3, plot a, illustrates simple slopes for treatment of approximately zero for students at the lowest end of the scale on the baseline test and nearly 0.4 for students at the highest end of the scale. Variation of the treatment effect on the grade-1 ITBS–MP shown in Figure 5.3, plot b, illustrates simple slopes for treatment of approximately 0.11 for students at the lowest end of scale on the baseline test and greater than 0.15 for students at the highest end of the scale. Variation of the treatment effect on the grade 1 ITBS–MC shown in Figure 5.3, plot c, illustrates simple slopes for treatment of approximately −0.02 for students at the lowest end of scale on the baseline test and greater than 0.12 for students at the highest end of the scale.

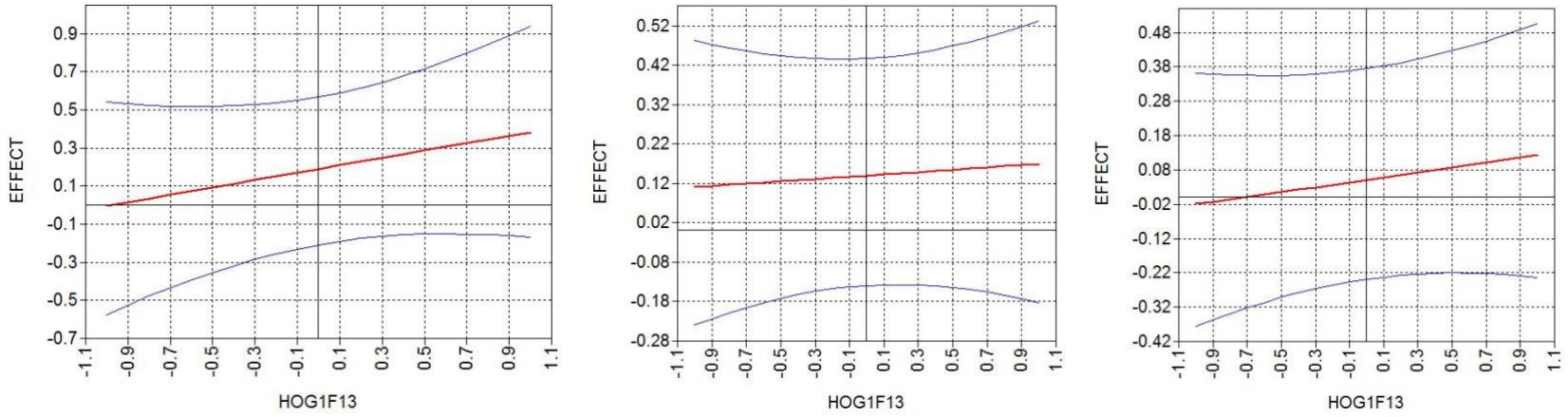


Figure 5.3. Plots illustrating variation of the size of the effect of treatment across the range of pretest scores for the grade-1 sample.

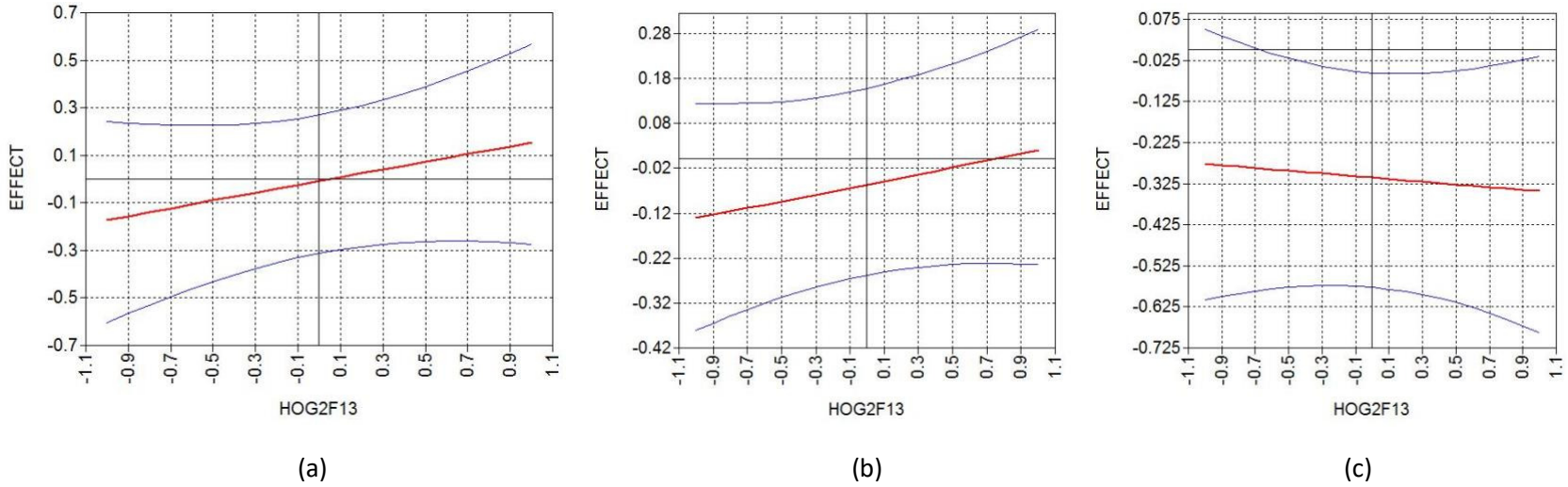


Figure 5.4. Plots illustrating variation of the size of the effect of treatment across the range of pretest scores for the grade-2 sample.

For the MPAC and ITBS–MP outcomes in grade 2, the interaction term had a positive point estimate and the interaction term for the grade-2 ITBS–MC analysis had a negative point estimate. As with grade 1, the grade-2 treatment-by-pretest interactions were of a negligible size. Variation of the treatment effect on the grade-2 MPAC interview shown in Figure 5.4, plot a, illustrates simple slopes for treatment of approximately −0.2 for students at the lowest end of the scale on the baseline test and nearly 0.2 for students at the highest end of the scale. Variation of the treatment effect on the grade-2 ITBS–MP shown in Figure 5.4, plot b, illustrates simple slopes for treatment of approximately −0.12 for students at the lowest end of the scale on the baseline test and approximately zero for students at the highest end of the scale. Variation of the treatment effect on the grade-2 ITBS–MC shown in Figure 5.4, plot c, illustrates simple slopes for treatment of approximately −0.25 for students at the lowest end of the scale on the baseline test and greater than −0.35 for students at the highest end of the scale.

The grade-2 ITBS–MC analysis results were different from those of all other moderation results, in that it was the only model that produced a statistically significant treatment effect conditional on the level of the moderator. Specifically, in the grade-2 treatment-by-pretest moderation analysis, the conditional effect for treatment indicated that treatment had a negative, statistically significant effect for students who had average achievement at baseline as measured by the baseline test. Table 5.6 provides a detailed reporting of the parameters for the conditional treatment effect and treatment-by-pretest moderation (additional detail of model parameters reported in Appendix H, Tables H.7 and H.8). Moderation analyses indicated a negative conditional effect of g = −0.31, with credibility intervals of 95% CI [−0.58, −0.06]. These results are commensurate with the grade-2 subgroup analysis, where the the main effect of treatment (across all levels of pretest) on the ITBS–MC test had an estmated coefficient of γ = −0.29, 95% CI [−0.59, 0.01].

Table 5.6. Summary of Treatment-by-Pretest Moderated Effects across Outcomes

	MPAC (Grade 1 *N* = 336; Grade 2 *N* = 284)			ITBS–MP (Grade 1 *N* = 1,025; Grade 2 *N* = 980)			ITBS–MC (Grade 1 *N* = 1,025; Grade 2 *N* = 980)		
Parameter	Estimate (*PSD*)	*g*	95% CI	Estimate (*PSD*)	*g*	95% CI	Estimate (*PSD*)	*g*	95% CI
				Grade 1					
Treatment	0.191 (0.196)	0.20	[−0.209, 0.568]	0.143 (0.147)	0.15	[−0.137, 0.443]	0.050 (0.155)	0.05	[−0.239, 0.376]
Treatment by Baseline test	0.192 (0.203)	0.20	[−0.198, 0.605]	0.022 (0.110)	0.02	[−0.202, 0.236]	0.071 (0.105)	0.07	[−0.137, 0.280]
				Grade 2					
Treatment	−0.008 (0.147)	−0.01	[−0.313, 0.271]	−0.056 (0.104)	−0.06	[−0.258, 0.157]	**−0.310 (0.132)**	**−0.31**	**[−0.577, −0.056]**
Treatment by Baseline test	0.165 (0.155)	0.17	[−0.150, 0.465]	0.076 (0.078)	0.08	[−0.077, 0.231]	−0.031 (0.106)	−0.03	[−0.251, 0.168]

Note. Two of the 622 students who participated in the MPAC interview had missing ITBS scores, so the analytic sample for these analyses was only 620. Multilevel regression models included school treatment assignment as the key predictor of interest, controlling for student grade-level; randomization block; and student demographic characteristics of gender, free/reduced-price lunch status, English language learner status, and disability status, as well as student mathematics performance at baseline. The slope for the baseline test was specified to vary randomly across clusters, and treatment was specified as a predictor of the school-level variation around the baseline-test slope, constituting the treatment-by-pretest interaction. Boldface indicates that the 95% credibility interval did not include zero.

6. Discussion

The purpose of the present study was to estimate the effect of the first year of the CGI PD intervention on first- and second-grade student achievement in mathematics and to determine whether the effects differ according to student characteristics. The CGI PD intervention evaluated in the present study was directed by one of the coauthors of the three definitive CGI books. The substance and design of the program has evolved over thirty years of large-scale implementation and research. It conforms to many of the research-based recommendations for effective teacher professional development, including content focus, extended duration, coherence, active learning, and collective participation (Desimone, 2009). Implementation of the CGI PD program was consistent with the planned program. The CGI workshop leaders were highly qualified with respect to the stringent standards specified by the program director. The overwhelming majority of teachers who participated in the program received the full dose of the planned intervention.

Using Hedges' *g* effect-size estimator with the confirmatory analyses on the combined grade-1 and grade-2 sample, we found an average effect of the program on student performance on those three outcomes to be 0.08, 0.03, and –0.11, respectively. None of the effects was statistically significant (even without any adjustment for multiple comparisons). Nevertheless, the point estimates of the main effects of the intervention can be considered the "true" effect. These effect-size estimates are in the small-medium range for causal studies of educational interventions (Kraft, 2020).

Subgroup and moderation analysis provide insight into the effect of the program on student achievement that the overall analyses do not offer. The results of both the subgroup and the moderation analyses appear to indicate that the program had a positive effect on grade-1 students' problem-solving abilities, but the effect on grade-1 students' computational abilities was approximately zero. The estimated effect on grade-2 students' abilities in problem solving was close to zero, but the point estimate of the effect on grade-2 students' computational abilities was negative. The conditional effect of the CGI program on grade-1 students' MPAC scores and ITBS–MP scores were commensurate with those of some of the stronger PD programs that have been subjected to rigorous evaluation of their effect on students (Kennedy, 2016a; 2016b; Kraft, 2020).

6.1. Exploration of Subgroup and Moderation Analyses

None of the effect-size estimates for confirmatory analyses had 95% credibility intervals that did not include zero, several noteworthy results do appear in the set of subgroup and moderation analyses. These two subgroup and moderation analyses consistently suggest potentially positive effects for grade-1 students' performance on the MPAC interview and ITBS Math Problems, whereas the ITBS Math Computation data consistently suggest potentially negative effects for grade-2 students.

The treatment effects by subgroup reported in Table 5.3 suggest that the potentially positive effects in the aggregate sample as measured by the MPAC interview appear to be driven by a positive effect on grade-1 student performance (g = 0.25), because a positive effect was not observed on grade-2 students' performance (g = −0.01). The potentially negative effect in the aggregate sample as measured by the ITBS–MC appears to be driven by a negative effect in the grade-2 sample (g = −0.29), as a negative effect was not observed in the grade-1 sample (g = 0.03). The credibility of the results from these subgroup analyses is bolstered by the result of the moderation analyses. Findings for the grade-level subgroups were replicated in the grade-level moderation analyses.

Results indicate some variation in the effect of treatment across levels of baseline achievement, where estimates were largest for students who had higher levels of achievement at baseline. Notable interaction effects were observed in grade 1 and grade 2 for the MPAC interview, where treatment effects were approximately one-fifth of one standard deviation higher for students one standard deviation above the mean at baseline than for students at the mean at baseline. This type of result, in which students with higher baseline achievement benefit more from an intervention than their lower-ability peers, is not uncommon. The result was not statistically significant, but it should prompt reflection and action by the program developers/implementers to remedy aspects of the program may be perpetuating inequity in mathematics learning opportunity for students.

Both the subgroup analysis and the moderation analysis suggest that the CGI program had a relatively large, positive effect on grade-1 students' problem-solving abilities as measured by the MPAC. This is good news for the CGI program and for researchers and practitioners alike. It provides new evidence that is largely consistent with the results reported from the first randomized trial of a CGI program implemented with grade-1 students and teachers in the mid-1980s (Carpenter et al., 1989). Because the MPAC interview measures student abilities in solving word problems and computation as well as solving problems involving algebraic thinking with respect to understanding the equals sign as a relational operator, this result is consistent with the results reported by Jacobs et al. (2007), suggesting that the CGI program has a positive impact on students' ability in the domain of algebraic thinking as well as in the domain of number and operations.

The CGI program appeared to have a larger effect on students who were not FRL-eligible than on those who were eligible. Subgroup and moderation analyses suggest a one-fourth to one-fifth standard deviation effect of treatment on non-FRL eligible students, whereas the effect on FRL-eligible students was estimated to be near zero and negative. While we think this result is worthy of note and of future study, we note that caution is warranted when inference is drawn from these results, because six schools did not have any students in the sample who were not FRL-eligible, and the resulting imbalance in the sampling on that characteristic might bias the results for this particular investigation.

The largest positive effect-size estimate occurred on the MPAC interview in the subgroup analysis for the students with disabilities. One of the largest negative effect-size estimates occurred for the same subgroup on the ITBS Math Computation test. These potential effects were not supported by the results of the treatment-by-SWD moderation analyses. Because only 40 students in the aggregate sample were identified as having disabilities, we have low confidence in the validity of the results of these particular subgroup analyses. Nonetheless, we believe this result warrants further study with a larger sample. The subsequent study should explore both the effect of the CGI program on the mathematical abilities of these students and the apparent discrepancy in these students' performance on the interview-based assessment and the group-administered, standardized test. The discrepancy between these students' performances in the interview setting and in the group-administered setting appeared greater than that for their peers.

The MPAC interview may be better suited to detecting program effects for several reasons. First, it is conducted in a semistructured, one-on-one setting, where the interviewer can observe the examinee and ask follow-up questions. This type of administration may increase reliability, and higher reliability can increase the strength of the association of factors (such as treatment condition and student outcomes) in data models. The MPAC also included items designed to measure student understanding of algebraic concepts, such as the meaning of the equals sign in mathematics. The CGI PD program focused on word problems, computation, and algebraic thinking, so the MPAC may have been better aligned with the content of the program.

The content of the MPAC was aligned with the expectations in the state curriculum standards for mathematics, so the students in the treatment and comparison conditions could reasonably be expected to have had opportunities to learn the material before the test was administered. In fact, the tests measured a broad swath of content in number, operations, and algebraic thinking, which is the mainstay of the elementary mathematics curriculum and something that all of the schools in the sample worked very hard to try to affect. In other words, the three outcome measures were not focused on a fringe topic or narrowly defined skill. This point should be considered carefully when the results the results are interpreted. It is much easier for a program to have a large effect on skills that are not emphasized in the comparison condition. As program evaluators, we took great caution to ensure that the MPAC was not overaligned with the CGI PD program. We note that the items on the MPAC were not known by the program developers or the participating teachers, and they were not used as part of the PD program.

6.2. The Importance of Content Focus in Teacher Professional Development

The positive impact on grade-1 students' achievement on the problem-solving tests might be most easily explained by an analysis of the content of the first year of the CGI PD program. The chapters from *Children's Mathematics* (Carpenter et al., 1999) that provided the focus of the workshops in the first year of the CGI PD program focused on the mathematics content and strategies that are largely consistent with the curriculum and level of understanding of grade-1 students. This result may provide further support for the well-established recommendations in favor of content focus in the design and delivery of effective professional-development programs (Desimone, 2009; Garet et al., 2001; Wilson, 2013; Yoon et al., 2007). In this case, the CGI frameworks for student thinking with respect to solving problems involving single-digit addition and subtraction provide an important focal point, because they provide teachers with a principled framework for identifying individual students' level of understanding. The in-depth focus on student thinking with respect to number, addition, subtraction, place value, and mathematical equality also served as the means to develop teachers' understanding of mathematics and mathematics-related vocabulary and notation for these topics.

Some of the chapters in *Children's Mathematics* (Carpenter et al., 1999) focus on multidigit addition and subtraction, multiplication and division, and some key ideas related to the place value in the decimal number system. The CGI frameworks for student thinking related to multidigit addition and subtraction align more with the focus of the second-grade curriculum, and they are studied more thoroughly in the second year of the CGI PD program implemented in the present study.

6.3. Alignment of Student Outcome Measures with Intervention

The three tests used to measure the effects of the program on student abilities in mathematics each served a different purpose. The two ITBS tests have been used for decades by many states and school districts to measure student achievement. They therefore provided a metric that is relevant to and understood by many school leaders and policymakers. Although the specific items and tests have surely changed over time, the ITBS tests were also used in the original randomized trial of CGI (Carpenter et al., 1989). We note that Carpenter et al. (1989) reported that the CGI program did not have a statistically significant positive effect on students' ITBS scores or their knowledge of number facts, but they did report a positive effect on students' abilities to solve nonroutine word problems.

The third test, the MPAC, used a different format and was able to focus more directly on the topics related to number, operations, and algebraic thinking that were the focus of the CGI program. The items on the MPAC test used a constructed-response format, and the items on the ITBS tests all used a selected-response format. Moreover, the content of the MPAC could be tailored to align with the content of the CGI program. To avoid overalignment, the test developers took great care not to include

any specific mathematics problems that had been part of the CGI program. To be sure the MPAC interview provided a fair comparison of the student abilities in the treatment and comparison conditions, the content of the MPAC was aligned with the content (and associated content limits) of the CCSS-M and the Mathematics Florida Standards (Florida Department of Education, 2014; NGA & CCSSO, 2010; Schoen et al., 2016). In our appraisal, all three of the tests would probably fall into Hill et al.'s (2005) *Standardized test (narrow)* category. The differences in effect-size estimates might be explained by differences in content or in the reliability of the tests, where the reliability estimate was highest for the MPAC and lowest for the ITBS–MP.

The interview setting in the MPAC provided opportunities for the assessor to observe students working on the problems and to ask follow-up questions to provide clarification about the students' responses. These features provide a different kind of information about student understanding and confer a distinct advantage of assessment in an interview setting over that of a group-administered, paper-pencil, multiple-choice test—especially for students in earlier grade levels or those who have not had extensive practice taking standardized tests. For example, when students answered "true" or "not true" to the questions about equations, the students' rationale was incorporated into the scoring. When students answered these items correctly, but they had major flaws in their reasoning, the response was scored as incorrect. For example, several items displayed statements such as 6 + 1 = 7 – 2 and asked the student to determine whether the equation was true or not true. In accordance with interview protocol, the interviewer then asked, "how did you decide that was not true?" Some students who answered "not true" would offer an explanation such as, "six plus one equals seven, but seven minus two equals five, so not true." This was scored as a correct response for correct reason. Other students who answered "not true" gave the rationale that "the minus sign cannot be on that side." These responses were scored as incorrect; the answer was correct, but the reasoning was mathematically incorrect. Although items like these can be field tested and removed from the set of items in a well-constructed test that uses a selected-response format, the interview format afforded the ability to use those types of items and use additional information about the examinee's reasoning processes in the scoring process.

6.4. Limitations

As with any individual study, the present study suffers from many limitations. Being a single, randomized controlled trial, the research design used in the present study maximizes internal validity but cannot speak to external validity. The extent to which its results can be generalized to other settings is not known.

Many of the teachers in the treatment-condition schools received the treatment, but some of them did not, and some of the participating teachers in the treatment-condition schools received less than the full year of PD. The present study therefore represents an intent-to-treat sample.

Informed by almost three decades of research on CGI, the CGI PD is designed to be a three-year program. The CGI PD program does not provide a script or curriculum for teachers to follow in the classroom, and the teachers can reasonably be expected to require some time before they can to harness the potential power of their new knowledge and beliefs about teaching and learning and use it to affect students positively. Previous studies have indicated that many teachers take multiple years to learn how to implement CGI in their classrooms. The present study focuses on impact during the first year of the intervention. Examination of the results after a second year will be important.

6.5. Future Directions

The size of the negative-effect estimate for grade-2 students' performance on the ITBS–MC is concerning. On the basis of these results, we strongly recommend modifying the program to improve its utility to grade-2 teachers and students. The program seemed to have a positive effect on grade-1 students, which is most easily explained by the focus of the first year of the PD on grade-1 material. One way to modify the program might be to lengthen the first year so that the content of the PD workshops can address multidigit computation and fluency with addition and subtraction facts.

The analytic sample in the present study only included students for whom we had follow-up test data. Future directions might involve reanalysis and the use of more sophisticated methods for handling missing data.

One goal of the larger study is to examine the theory of change for the CGI program and to revise it on the basis of empirical findings. Analyses of data on teacher knowledge, teacher beliefs, and classroom instruction will be needed to permit a more through investigation of the mechanisms at work in the theory of change. Except the results concerning the effect on teachers' mathematical knowledge for teaching, those results are not available as of this writing. Extended duration is also an important component of the CGI program. Examination of the effects of the second year of intervention on student achievement will be important.

The ideas teachers encounter in the first year of the program are complex, and the goals of the program are ambitious. We know that teachers' knowledge and beliefs were significantly affected in the first year of the program (Schoen, Kisa, & Tazaz, manuscript in preparation; Schoen, Secada, & Tazaz, 2015). Teachers may reasonably be expected to require more than a single year to learn how to use their new knowledge and perspective on mathematics teaching and learning to achieve a greater effect on their students. Many of the teachers in our sample started the program with many years of teaching experience behind them. The ability of PD to improve teaching and learning may require teachers first to learn to inhibit some habits they have developed over years of teaching. For example, a teacher whose first instinct when teaching ELL students is to teach them to identify key, individual words or phrases in a word problem may need to learn to inhibit that response in order to engage the students in mathematical problem solving that delves into the deeper meanings of the problems. Given that most textbook series continue to make use of key words, teachers must not only inhibit their own initial responses (built up over years of practice), but they must also ignore explicit cues from mathematics books if they are shifting their instructional practice to center on problem solving.

6.6. Conclusion

Overall patterns in the treatment-effect estimates reveal positive effects on the problem-solving outcomes—especially for grade-1 students—and negative effects on the computation outcome—especially for grade-2 students. The positive effects on the problem-solving and algebraic-thinking tests are encouraging. The negative effects on grade-2 students' computational ability are concerning. These results are not statistically significant, but the magnitude of the effects may be substantively important. We recommend a careful review that should result in swift and substantive adjustment to the program to address the concern about negative impact on grade-2 students' computational abilities. A follow-up study—ideally one with more statistical power—could reveal more about the generalizability of these results and examine whether the program adjustments have the desired effect.

.

References

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newberry Park, CA: Sage Publications.

Asparouhov, T., & Muthén, B. (2007). Constructing covariates in multilevel regression. *Mplus Web Notes: No. 11* (version 2). Retrieved from https://www.statmodel.com.

Asparouhov, T., & Muthén, B. (2010a). *Bayesian analysis of latent variable models using Mplus* (Technical report). Los Angeles, CA: Muthén & Muthén.

Asparouhov, T., & Muthén, B. (2010b). *Bayesian analysis using Mplus: Technical implementation* (Technical appendix). Los Angeles, CA: Muthén & Muthén.

Behrend, J. L. (2003). Learning-disabled students make sense of mathematics. *Teaching Children Mathematics, 9*(5), 269–273.

Bloom, H. S. (2005). Randomizing groups to evaluate place-based programs. In H. S. Bloom (Ed.), *Learning more from social experiments: Evolving analytic approaches* (pp. 115–172). New York: Russell Sage.

Brookes, S. T., Whitely, E., Egger, M., Smith, G. D., Mulheran, P. A., & Peters, T. J. (2004). Subgroup analyses in randomized trials: Risks of subgroup-specific analyses; power and sample size for the interaction test. *Journal of Clinical Epidemiology*, *57*, 229–236.

Buros Institute (2010). *Mental measurements yearbook and tests in print.* Lincoln, NB: University of Nebraska

Carpenter, T. P. (1985). Learning to add and subtract: An exercise in problem solving. In E. A. Silver (Ed.), *Teaching and learning mathematical problem solving: Multiple research perspectives* (pp. 17–40). Hillsdale, NJ: Erlbaum.

Carpenter, T. P. (1989). Teaching as problem solving. In R. I. Charles & E. Silver (Eds.), *Research agenda in mathematics education: The teaching and assessing of mathematical problem solving* (187–202). Reston, VA: National Council of Teachers of Mathematics; Hillside, NJ: Erlbaum.

Carpenter, T. P., Fennema, E., & Franke, M. L. (1996). Cognitively guided instruction: A knowledge base for reform in primary mathematics instruction. *The Elementary School Journal, 97*(1), 3–20.

Carpenter, T.P., Fennema, E., Franke, M. L., Levi, L., & Empson, S.B. (2015). *Children's mathematics: Cognitively guided instruction.* Portsmouth, NH: Heinemann.

Carpenter, T. P., Fennema, E., Franke, M. L., Levi, L., & Empson, S. B. (1999). *Children's mathematics: Cognitively guided instruction.* Portsmouth, NH: Heinemann.

Carpenter, T. P., Fennema, E., Peterson, P. L., & Carey, D. A. (1988). Teachers' pedagogical content knowledge of students' problem solving in elementary arithmetic. *Journal for Research in Mathematics Education, 19*(5), 385–401.

Carpenter, T. P., Fennema, E., Peterson, P. L., Chiang, C.P., & Loef, M. (1989). Using knowledge of children's mathematics thinking in classroom teaching: An experimental study. *American Educational Research Journal, 26*(4), 385–531.

Carpenter, T. P., & Franke, M. L. (2004). Cognitively guided instruction: Challenging the core of educational practice. In T. K. Glennan, S. J. Bodilly, J. R. Galegher & K. A. Kerr (Eds.), *Expanding the reach of education reforms: Perspectives from leaders in the scale-up of educational interventions* (pp. 41–80). Santa Monica, CA: RAND Corporation.

Carpenter, T. P., Franke, M. L., Jacobs, V. R., Fennema, E., & Empson, S. B. (1998). A longitudinal study of invention and understanding in children's multidigit addition and subtraction. *Journal for Research in Mathematics Education, 29*(1), 3–20.

Carpenter, T. P., Franke, M. L., & Levi, L. (2003). *Thinking mathematically: Integrating arithmetic & algebra in elementary school*. Portsmouth, NH: Heinemann.

Castellano, K. E., Rabe-Hesketh, S., & Skrondal, A. (2014). Composition, context, and endogeneity in school and teacher comparisons. *Journal of Educational and Behavioral Statistics*, *39*(5), 333–367.

Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher, 3*8(3), 181–199.

Dixon, J. K., Larson, M., Leiva, M. A., & Adams, T. L. (2013). *Go math! Florida.* Orlando, FL: Houghton Mifflin Harcourt.

Dunbar, S. B., Hoover, H. D., Frisbie, D. A., Ordman, V. L., Oberley, K. R., Naylor, R. J., & Bray, G. B. (2008). *Iowa Test of Basic Skills.®* Rolling Meadows, IL: Riverside Publishing.

Empson, S. B., and Levi, L. (2011). *Extending children's mathematics: Fractions and decimals.* Portsmouth, NH: Heinemann.

Falkner, K. P., Levi, L., & Carpenter, T. P. (1999). Children's understanding of equality: A foundation for algebra. *Teaching Children Mathematics*, *6*(4), 232–236.

Fennema, E., Carpenter, T. P., Franke, M. L., Levi, L., Jacobs, V. R., & Empson, S. B. (1996). A longitudinal study of learning to use children's thinking in mathematics instruction. *Journal for Research in Mathematics Education,* 27, 458–477.

Fennema, E., Carpenter, T. P., Levi, L., Franke, M. L., & Empson, S. B. (1999). *Children's mathematics: Cognitively guided instruction: A guide for workshop leaders.* Portsmouth, NH: Heinemann.

Fermanich, M. L. (2002). School spending for professional development: A cross-case analysis of seven schools in one urban district. *The Elementary School Journal* 103(1): 27–50.

Franke, M. L., Carpenter, T., Fennema, E., Ansell, E., and Behrend, J. (1998). Understanding teachers' self-sustaining, generative change in the context of professional development. *Teaching and Teacher Education*, *14*(1), 67–80.

Franke, M. L., Carpenter, T. P., Levi, L., & Fennema, E. (2001). Capturing teachers' generative change: A follow-up study of professional development in mathematics. *American Educational Research Journal, 38*(3), 653–689.

Fuson, K. (1992). Research on whole number addition and subtraction. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning.* Reston, VA: National Council of Teachers of Mathematics.

Gage, N. L. (2009). *A Conception of Teaching*. New York: Springer U.S.

Garet, M. S., Heppen, J. B., Walters, K., Parkinson, J., Smith, T. M., Song, M., et al. (2016). *Focusing on mathematical knowledge: The impact of content-intensive teacher professional development* (NCEE 2016-4010). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Retrieved from http://ies.ed.gov/ncee/pubs/20164010/pdf/20164010.pdf

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal, 38*(4), 915–945.

Garet, M., Wayne, A., Stancavage, F., Taylor, J., Eaton, M., Walters, K., Song, M., Brown, S., Hurlburt, S., Zhu, P., Sepanik, S., & Doolittle, F. (2011). *Middle school mathematics professional development impact study: Findings after the second year of implementation* (NCEE 2011-4024). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Boca Raton: CRC press.

Gersten, R., Taylor, M. J., Keys, T. D., Rolfhus, E., & Newman-Gonchar, R. (2014). *Summary of research on the effectiveness of math professional development approaches.* (REL2014–010). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southeast. Retrieved from http://ies.ed.gov/ncee/edlabs.

Hiebert, J., & Carpenter, T. P. (1992). Learning and teaching with understanding. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 65–97). New York: Macmillan.

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematics knowledge for teaching on student achievement. *American Educational Research Journal, 42*(2), 371–406.

Hox, J. J., Maas, C. J. M., & Brinkhuis, M. J. S., (2010). The effect of estimation method and sample size in multilevel structural equation modeling. *Statistica Neerlandica, 64*(2), 157–170.

Ivers, N. M., Halperin, I. J., Barnsley, J., Grimshaw, J. M., Shah, B. R., Tu, K., Upshur, R. & Zwarenstein, M. (2012). Allocation techniques for balance at baseline in cluster randomized trials: A methodological review. *Trials*, *13*(120), 1–9.

Jacob, R., Hill, H., & Corey, D. (2017). The impact of a professional development program on teachers' mathematical knowledge for teaching, instruction, and student achievement. *Journal for Research on Educational Effectiveness, 10*(2), 379–407.

Jacobs, V. R., Franke, M. L., Carpenter, T. P., Levi, L., & Battey, D. (2007). Professional development focused on children's algebraic reasoning in elementary school. *Journal for Research in Mathematics Education, 38*(3), 258–288.

Jaslow, L. & Evans, E. L. (2012). Purposeful pedagogy and discourse instructional model: Student thinking matters most. Little Rock, AR: Arkansas Department of Education.

Kahan, B. C., Jairath, V., Doré, C. J., & Morris, T. P. (2014). The risks and rewards of covariate adjustment in randomized trials: an assessment of 12 outcomes from 8 studies. *Trials*, 15(139), 1–7.

Kaplan, D., & Depaoli, S. (2012). Bayesian structural equation modeling. In R. Hoyle(Ed.), *Handbook of structural equation modeling* (pp. 650– 673). New York, NY: Guilford Press.

Kraft, M. A. (2020). Interpreting effect sizes of educational interventions. *Educational Researcher, 49*(4), 241–253.

Kennedy, M. M. (2016a). How does professional development improve teaching? *Review of Educational Research, 86*(4), 1–36.

Kennedy, M. M. (2016b). Parsing the practice of teaching. *Journal of Teacher Education, 67*(1), 6–17.

Knapp, N. F., & Peterson, P. L. (1995). Teachers' interpretations of "CGI" after four years: Meanings and practices. *Journal for Research in Mathematics Education, 26*(1), 40–65.

Lecoutre, B., Lecoutre, M.-P., & Poitevineau, J. (2001). Uses, abuses and misuses of significance tests in the scientific community: Won't the Bayesian choice be unavoidable? *International Statistical Review*, *69*(3), 399–417.

Levi, L. (2017, October 19). Classroom-embedded work: An alternative to observations lessons. [Blog post]. Retrieved from http://www.teachingisproblemsolving.org/blog/.

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods, 13*(3), 203–229

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 1*(3), 85–91.

Matthews, R., Wasserstein, R., & Spiegelhalter, D. (2017). The ASA's p-value statement, one year on. *Significance*, *14*(2), 1740–9713. doi: 10.1111/j.1740-9713.2017.01021.x

McNeish, D. M. & Stapleton, L. M. (2014). The effect of small sample size on two-level model estimation: A review and illustration. *Educational Psychology Review, 28*, 295–314

Muthén, B. & Asparouhov, T. (2013). *BSEM measurement invariance analysis*. Mplus Web Notes: No. 17. Retrieved from http://www.statmodel.com.

Muthén, B. O., Muthén, L. K., & Asparouhov, T. (2016). *Regression and mediation analysis using Mplus*. Los Angeles, CA: Muthen & Muthen.

Muthén, L. K. and Muthén, B. O. (1998–2012). *Mplus User's Guide* (7th ed.). Los Angeles, CA: Muthen & Muthen.

Nielsen, L., Steinthorsdottir, O. B., & Kent, L. B. (2016). Responding to student thinking: Enhancing mathematics instruction through classroom based professional development. Middle School Journal, 47(3), 17–24. http://www.doi.org/10.1080/00940771.2016.1135096

NGA & CCSSO (National Governors Association Center for Best Practices & Council of Chief State School Officers) (2010). Common core state standards for mathematics. Washington, DC: National Governors Association Center for Best Practices & Council of Chief State School Officers . Retrieved from http://www.corestandards.org/assets/CCSSI_Math%20Standards.pdf.

Odden, A., Archibald, S., Fermanich, M., & Gallagher, H. A. (2002). A cost framework for professional development. *Journal of Education Finance, 28*(1), 51–74.

Peterson, P. L., Fennema, E., Carpenter, T. P., & Loef, M. (1989). Teachers' pedagogical content beliefs in mathematics. *Cognition and Instruction, 6*(1), 1–40.

Raudenbush, S. W. (1997). Statistical analysis and optimal design for group randomized trials. *Psychological Methods*, *2*(2), 173–185.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models.* Thousand Oaks, CA: Sage.

Raudenbush, S. W., Martinez, A., & Spybrook, J. (2007). Strategies for improving precision in group-randomized experiments. *Educational Evaluation and Policy Analysis*, *29*(1), 5–29.

Rothwell, P. M. (2005). Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *Lancet, 365*(9454), 176–186.

Sacerdote, B. (2001). Peer effects with random assignment: Results for Dartmouth roommates. *The Quarterly Journal of Economics*, 116(2), 681–704. https://doi.org/10.1162/00335530151144131.

Schoen, R. C., Kisa,Z. & Tazaz, A. M. (2019, March). *Beyond the horizon: Examining the associations among professional development, teachers' subject-matter knowledge, and student achievement.* Paper presented at the spring conference of the Society for Research in Educational Effectiveness, Washington, DC.

Schoen, R. C., LaVenia, M., Bauduin, C., & Farina, K. (2016). *Elementary mathematics student assessment: Measuring the performance of grade 1 and 2 students in counting, word problems, and computation in fall 2013* (Research Report No. 2016-03). Tallahassee, FL: Learning Systems Institute, Florida State University. https://doi.org/10.17125/fsu.1508170543.

Schoen, R. C., LaVenia, M., Champagne, Z., Farina, K., & Tazaz, A. M. (2016). *Mathematics Performance and Cognition (MPAC) Interview: Measuring first- and second-grade student achievement in number, operations, and equality in spring 2015* (Report No. 2016-02). Tallahassee, FL: Florida State University. https://doi.org/10.17125/fsu.1493238666.

Schoen, R. C., Secada, W., & Tazaz, A. M. (2015, June). *Results after the first year of a randomized controlled trial of CGI.* Presented at the biennial Cognitively Guided Instruction National Conference, Lawndale, CA.

Secada, W. G. & Brendefur, J. L. (2000). CGI student achievement in Region VI: Evaluation findings. *The Newsletter of the Comprehensive Center-Region VI, 5*(2).

Snijders, T. A. B., & Bosker, R. J. (1999). Multilevel analysis: An introduction to basic and advanced multilevel modeling. Thousand Oaks, CA: Sage.

Tanniou, J., van der Tweel, I., Teerenstra, S., & Roes, K. C. B. (2016). Subgroup analyses in confirmatory clinical trials: Time to be specific about their purposes. *BMC Medical Research Methodology*, 16(20), 1–15.

Tazaz, A.M. & Schoen, R. C. (2020). *Measuring Implementation of the First Two Years of the Teacher Development Group Model for Professional Development Based on Cognitively Guided Instruction* (Research Report No. 2020–01). Tallahassee, FL: Learning Systems Institute, Florida State University.

TNTP (The New Teacher Project) (2015). *The mirage: Confronting the hard truth about our quest for teacher development.* Retrieved from https://tntp.org/assets/documents/TNTP-Mirage_2015.pdf.

Turner, E. E., & Celedón-Pattichis, S. (2011). Mathematical problem solving among Latina/o kindergartners: An analysis of opportunities to learn. *Journal of Latinos and Education*, *10*(2), 146–169.

U.S. Department of Education. (2014). *Fiscal year 2014 budget summary and background information*. Washington, DC: U.S. Department of Education . Retrieved from http://www2.ed.gov/about/overview/budget/budget14/summary/14summary.pdf

U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse (2013). *What Works Clearinghouse: Procedures and standards handbook (Version 3.0)*. Retrieved from http://whatworks.ed.gov.

U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse. (2016). *Cluster design standards*. Retrieved from http://whatworks.ed.gov.

Verschaffel, L., Greer, B., & De Corte, E. (2007). Whole numbers concepts and operations. In F.K. Lester, Jr. (Ed.), *Second handbook of research on mathematics teaching and learning.* Reston, VA: National Council of Teachers of Mathematics.

Villaseñor, A., & Kepner, H. S. (1993). Arithmetic from a problem-solving perspective: An urban implementation. *Journal for Research in Mathematics Education, 24*(1), 62–69.

Wasserstein R. L. & Lazar, N. A. (2016). The ASA's Statement on p-Values: Context, Process, and Purpose. *The American Statistician*, *70*(2), 129–133. doi: 10.1080/00031305.2016.1154108.

Wei, R. C., Darling-Hammond, L., & Adamson, F. (2010). *Professional development in the United States: Trends and challenges.* Dallas, TX. National Staff Development Council.

Wilson, S. M. (2013). Professional development for science teachers. *Science, 340*, 310–313.

Yoon, K. S., Duncan, T., Lee, S. W.-Y., Scarloss, B., & Shapley, K. (2007). *Reviewing the evidence on how teacher professional development affects student achievement* (Issues & Answers Report, REL 2007–No. 033). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southwest. Retrieved from http://ies.ed.gov/ncee/edlabs.

Zan, B. S., & Escalada, L. T. (2011). *Ramps and pathways: Evaluation of an inquiry-based approach to engaging young children in physical science.* Paper presented at the meeting of the American Educational Research Association, New Orleans, LA, April 2011.

Appendix A. Descriptive Statistics for Student Demographics and Achievement

Table A.1. Student Demographics for the 2014 MPAC Early-Joiners Analytic Sample, Disaggregated by District

	Treatment		Comparison		Total	
	N	District proportion	*N*	District proportion	*N*	District proportion
	District A (Treatment N = 185; Comparison N = 245; Total N = 430)					
Male	87	.48	120	.49	207	.49
Race/Ethnicity						
Asian	7	.04	10	.04	17	.04
Black	50	.27	54	.22	104	.24
Hispanic	71	.39	127	.52	198	.46
Multiracial	4	.02	2	.01	6	.01
White	51	.28	51	.21	102	.24
FRL	116	.63	192	.79	308	.72
ELL	48	.26	80	.33	128	.30
Exceptionality						
SWD	9	.05	18	.07	27	.06
Gifted	16	.09	3	.01	19	.04
Unknown	2	.01	1	<.01	3	.01
	District B (Treatment N = 120; Comparison N = 72; Total N = 192)					
Male	58	.48	36	.50	94	.49
Race/Ethnicity						
Asian	13	.11	3	.04	16	.08
Black	10	.08	8	.11	18	.09
Hispanic	16	.13	24	.33	40	.21
Multiracial	9	.08	4	.06	13	.07
White	72	.60	33	.46	105	.55
FRL	30	.25	43	.60	73	.38
ELL	1	.01	11	.15	12	.06
Exceptionality						
SWD	10	.08	3	.04	13	.07
Gifted	6	.05	6	.08	12	.06
Unknown	0	.00	0	.00	0	.00

Note. Asian = Asian/Pacific Islander, non-Hispanic; Black = Black/African American, non-Hispanic; Hispanic = Hispanic/Latino ethnicity, any racial group; Multiracial = Multiracial or American Indian/Alaskan Native, non-Hispanic; White = White, non-Hispanic. FRL = Eligible for free/reduced-price lunch. ELL = English Language Learners. SWD = Students with Disabilities. Gifted = Gifted and Talented. Unknown = Missing demographic data.

Table A.2. Student Demographics for the 2014 ITBS Early- and Late-Joiners Analytic Sample, Disaggregated by District

	Treatment		Comparison		Total	
	N	District proportion	*N*	District proportion	*N*	District proportion
	District A (Treatment N = 678; Comparison N = 823; Total N = 1501)					
Male	337	.50	407	.50	744	.50
Race/Ethnicity						
Asian	23	.03	35	.04	58	.04
Black	150	.22	169	.21	319	.21
Hispanic	260	.39	403	.49	663	.45
Multiracial	14	.02	20	.02	34	.02
White	223	.33	191	.23	414	.28
FRL	415	.62	644	.79	1059	.71
ELL	181	.27	266	.33	447	.30
Exceptionality						
SWD	31	.05	68	.08	99	.07
Gifted	36	.05	16	.02	52	.04
Unknown	8	.01	5	.01	13	.01
	District B (Treatment N = 445; Comparison N = 226; Total N = 671)					
Male	232	.52	101	.45	333	.50
Race/Ethnicity						
Asian	45	.10	5	.02	50	.08
Black	37	.08	33	.15	70	.10
Hispanic	67	.15	69	.31	136	.20
Multiracial	16	.04	10	.04	26	.04
White	280	.63	109	.48	389	.58
FRL	126	.28	126	.56	252	.38
ELL	23	.05	22	.10	45	.07
Exceptionality						
SWD	39	.09	20	.09	59	.09
Gifted	25	.06	14	.06	39	.06
Unknown	0	.00	0	.00	0	.00

Note. Abbreviations and designations as in Table A.1.

Table A.3. Student Demographics for the 2014 ITBS Early-Joiners Analytic Sample

	Treatment (*N* = 1,100)		Comparison (*N* = 1,020)		Total (*N* = 2,120)	
	N	Proportion	*N*	Proportion	*N*	Proportion
Male	559	.51	489	.48	1,048	.50
Race/Ethnicity						
Asian	67	.06	40	.04	107	.05
Black	185	.17	195	.19	380	.18
Hispanic	315	.29	460	.45	775	.37
Multiracial	30	.03	28	.03	58	.03
White	495	.45	293	.29	788	.37
FRL	524	.48	749	.74	1,273	.60
ELL	199	.18	280	.28	479	.23
Exceptionality						
SWD	68	.06	86	.09	154	.07
Gifted	61	.06	29	.03	90	.04
Unknown	8	.01	4	<.01	12	.01

Note. Abbreviations and designations as in Table A.1.

Table A.4. Student Demographics for the 2014 ITBS Early-Joiners Analytic Sample, Disaggregated by District

	Treatment		Comparison		Total	
	N	District proportion	*N*	District proportion	*N*	District proportion
	District A (Treatment N = 665; Comparison N = 802; Total N = 1,467)					
Male	329	.50	395	.50	724	.50
Race/Ethnicity						
Asian	23	.04	35	.04	58	.04
Black	148	.23	165	.21	313	.22
Hispanic	252	.38	393	.49	645	.44
Multiracial	14	.02	19	.02	33	.02
White	220	.34	186	.23	206	.28
FRL	406	.62	630	.79	1,036	.71
ELL	177	.27	259	.33	436	.30
Exceptionality						
SWD	30	.05	66	.08	96	.07
Gifted	36	.06	16	.02	52	.04
Unknown	8	.01	4	.01	12	.01
	District B (Treatment N = 435; Comparison N = 218; Total N = 653)					
Male	230	.53	94	.43	324	.50
Race/Ethnicity						
Asian	44	.10	5	.02	49	.08
Black	37	.09	30	.14	67	.10
Hispanic	63	.15	67	.31	130	.20
Multiracial	16	.04	9	.04	25	.04
White	275	.63	107	.49	382	.56
FRL	118	.27	119	.55	237	.36
ELL	22	.05	21	.10	43	.07
Exceptionality						
SWD	38	.09	20	.09	58	.09
Gifted	25	.06	13	.06	38	.06
Unknown	0	.00	0	.00	0	.00

Note. Abbreviations and designations as in Table A.1.

Table A.5. Analytic Sample Summary Statistics for Achievement Measures

	Treatment		Comparison		Total	
	N	*M* (*SD*)	*N*	*M* (*SD*)	*N*	*M* (*SD*)
			MPAC analytic sample			
G2F13 EMSA	161	0.170 (0.691)	175	−0.045 (0.720)	336	0.058 (0.715)
G2F13 EMSA	143	0.111 (0.716)	141	−0.156 (0.802)	284	−0.021 (0.771)
MPAC	305	0.468 (0.905)	317	0.132 (0.872)	622	0.297 (0.904)
			ITBS analytic sample			
G1F13 EMSA	535	0.160 (0.699)	490	−0.052 (0.712)	1,025	0.059 (0.713)
G2F13 EMSA	511	0.065 (0.722)	469	−0.168 (0.760)	980	−0.046 (0.750)
ITBS–MP	1,123	166.625 (20.962)	1,049	161.213 (21.824)	2,172	164.011 (21.553)
ITBS–MC	1,123	160.949 (15.865)	1,049	160.496 (16.111)	2,172	160.730 (15.986)

Note. G1F13 EMSA = Grade 1, fall 2013 baseline mathematics test; G2F13 EMSA = Grade 2 baseline mathematics test, fall 2013; MPAC = *Mathematics Performance and Cognition* test; ITBS–MP = *Iowa Test of Basic Skills Math Problems* test; ITBS–MC = *Iowa Test of Basic Skills Math Computation* test.

Appendix B. Variables and Models

Table B.1. Description of Variables and Models Used in Analyses of Main Effects

Model	Construct/variable	Variable description	Modeling particulars
M0	Student characteristics		
	Grade 2	Binary independent variable indicating student was in grade 2	Modeled at Level 1
	School characteristics		
	Block	Vector of *n*–1 binary independent variables indicating randomization blocks	Modeled at Level 3
M1	Student characteristics		
	Grade 2	Binary independent variable indicating student was in grade 2	Modeled at Level 1
	School characteristics		
	Treatment	Binary independent variable indicating school was assigned to the Treatment group	Modeled at Level 3
	Block	Vector of *n*–1 binary independent variables indicating randomization blocks	Modeled at Level 3
M2	*All variables in M1 plus:* Student characteristics		
	Male	Binary independent variable indicating student was male	Modeled at Level 1. The mean and variance for all student characteristics (including grade 2) are estimated.
	Minority	Binary independent variable indicating student was of a non-White race or Hispanic ethnicity	
	FRL	Binary independent variable indicating student was eligible for free/reduced-price lunch	
	ELL	Binary independent variable indicating student was eligible for English language learner services	
	SWD	Binary independent variable indicating student was identified as having a disability	
M3	*All variables in M1 and M2 plus:* Baseline mathematics		
	Baseline test	Continuous independent variable indicating student mathematics achievement fall 2013	Modeled at all three levels as latent variable covariates at Level 2 and Level 3. Variance of baseline test is estimated at Level 1.

Note. Level 1, Level 2, and Level 3 indicate the within classroom, between classroom, and between school portions of the model, respectively. This same modeling procedure was employed for each dependent variable of student performance on the MPAC Interview, ITBS–MP test, and ITBS–MC test.

Appendix C. Patterns in Missing Data for Confirmatory Analyses

Table C.1. Missing Data Patterns MPAC Analyses

Variable	Pattern 1 (*n* = 336)	Pattern 2 (*n* = 281)	Pattern 3 (*n* = 3)	Pattern 4 (*n* = 2)
MPAC	x	x	x	x
Grade 2	x	x	x	x
Male	x	x	x	x
Minority	x	x		x
FRL	x	x		x
ELL	x	x		x
SWD	x	x		x
G1F13 EMSA	x			
G2F13 EMSA		x	x	
Treatment	x	x	x	x
Block	x	x	x	x

Note. Total *N* = 622. x = Not missing. Abbreviations as in Table A.1.

Table C.2. Missing Data Patterns for ITBS Analyses

Variable	Pattern 1 (*n* = 1,023)	Pattern 2 (*n* = 971)	Pattern 3 (*n* = 165)	Pattern 4 (*n* = 7)	Pattern 5 (*n* = 2)	Pattern 6 (*n* = 2)	Pattern 7 (*n* = 1)	Pattern 8 (*n* = 1)
ITBS	x	x	x	x	x	x	x	x
Grade 2	x	x	x	x	x	x	x	x
Male	x	x	x	x		x		x
Minority	x	x	x					
FRL	x	x	x					
ELL	x	x	x					
SWD	x	x	x					
G1F13 EMSA	x						x	x
G2F13 EMSA		x		x	x			
Treatment	x	x	x	x	x	x	x	x
Block	x	x	x	x	x	x	x	x

Note. Total *N* = 2,172. x = Not missing. Abbreviations as in Table A.1

Appendix D. Model Results

Table D.1. Treatment Effect on MPAC across Different Models with Covariates for Aggregate Sample

	Model 0		Model 1			Model 2			Model 3		
	Estimate (*PSD*)	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI
Fixed effects											
Within classroom											
Grade 2	**0.711 (0.077)**	**[0.560, 0.861]**	**0.708 (0.076)**	—	**[0.560, 0.856]**	**0.668 (0.075)**	—	**[0.522, 0.814]**	**0.703 (0.078)**	—	**[0.542, 0.849]**
Male						0.128 (0.066)	—	[−0.003, 0.258]	**0.123 (0.048)**	—	**[0.029, 0.218]**
Minority						**−0.169 (0.084)**	—	**[−0.334, −0.004]**	−0.069 (0.063)	—	[−0.193, 0.050]
FRL						**−0.226 (0.092)**	—	**[−0.406, −0.045]**	−0.041 (0.071)	—	[−0.178, 0.101]
ELL						**−0.336 (0.091)**	—	**[−0.515, −0.159]**	**−0.173 (0.068)**	—	**[−0.306, −0.042]**
SWD						**−0.574 (0.138)**	—	**[−0.848, −0.304]**	**−0.284 (0.100)**	—	**[−0.484, −0.087]**
G1F13 EMSA									**0.784 (0.048)**	—	**[0.685, 0.874]**
G2F13 EMSA									**0.844 (0.040)**	—	**[0.765, 0.924]**
Between classrooms											
G1F13 EMSA									−0.249 (1.435)	—	[−3.882, 2.209]
G2F13 EMSA									0.758 (1.333)	—	[−2.309, 3.719]
Between schools											
Treatment			0.195 (0.132)	0.20	[−0.074, 0.449]	0.142 (0.129)	0.14	[−0.119, 0.394]	0.083 (0.168)	0.08	[−0.253, 0.417]
Block[a]	—	—	—	—	—	—	—	—	—	—	—
G1F13 EMSA									0.295 (0.702)	—	[−1.071, 1.730]
G2F13 EMSA									0.465 (0.719)	—	[−0.960, 1.897]
Intercept	0.089 (0.224)	[−0.361, 0.534]	−0.010 (0.215)	—	[−0.434, 0.421]	0.302 (0.223)	—	[−0.141, 0.749]	−0.082 (0.275)	—	[−0.637, 0.465]
Variance components											
Within classroom	**0.707 (0.044)**	**[0.626, 0.800]**	**0.708 (0.044)**		**[0.626, 0.800]**	**0.648 (0.041)**		**[0.573, 0.734]**	**0.033 (0.028)**		**[0.001, 0.102]**
Between classrooms	**0.031 (0.023)**	**[0.003, 0.090]**	**0.029 (0.023)**		**[0.002, 0.088]**	**0.030 (0.022)**		**[0.003, 0.084]**	**0.023 (0.018)**		**[0.001, 0.066]**

(Continued)

	Model 0		Model 1			Model 2			Model 3		
	Estimate (*PSD*)	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI
Between schools	**0.049 (0.069)**	**[0.006, 0.245]**	**0.033 (0.070)**		**[0.002, 0.220]**	**0.032 (0.061)**		**[0.002, 0.201]**	**0.029 (0.083)**		**[0.002, 0.244]**
R-Square											
Within classroom	**0.151 (0.029)**	**[0.098, 0.210]**	**0.150 (0.028)**		**[0.097, 0.207]**	**0.219 (0.030)**		**[0.162, 0.162]**	**0.959 (0.036)**		**[0.868, 0.999]**
Between classrooms	—	—	—		—	—		—	**0.497 (0.284)**		**[0.025, 0.970]**
Between schools	0.000 (0.000)	[0.000, 0.000]	0.000 (0.000)		[0.000, 0.000]	0.000 (0.000)		[0.000, 0.000]	**0.713 (0.266)**		**[0.066, 0.985]**
Intraclass correlation											
Between classrooms	.039		.038			.042			.271		
Between schools	.062		.043			.045			.341		

Note. Student *N* = 622; Teacher *N* = 167; School *N* = 22. FRL = Free/reduced-price lunch; ELL = English language learner; SWD = Student with disability. G1F13 EMSA = Grade 1, fall 2013 baseline mathematics test; G2F13 EMSA = Grade 2 baseline mathematics test, fall 2013; 95% CI = 95% credibility intervals of the posterior distribution with equal tail percentages. PSD = the standard deviation of the posterior distribution. Only the effect size for Treatment is presented; it is calculated as Hedges' *g*. Average cluster size for classrooms = 3.725; average cluster size for schools = 28.273. All models used Bayesian estimation. Reported estimates are from the unstandardized solution. Boldface indicates the 95% CI does not include zero.

[a]Block indicates the vector of *n*–1 randomization blocks. Effects for Block are omitted for visual simplicity.

Table D.2. Treatment Effect on ITBS–MP across Different Models with Covariates for Aggregate Sample

	Model 0		Model 1			Model 2			Model 3		
	Estimate (*PSD*)	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI
Fixed effects											
Within classroom											
Grade 2	**1.002 (0.046)**	**[0.913, 1.093]**	**1.003 (0.045)**	—	**[0.914, 1.092]**	**1.002 (0.042)**	—	**[0.919, 1.084]**	**1.006 (0.043)**	—	**[0.922, 1.091]**
Male						0.046 (0.032)	—	[−0.017, 0.108]	0.019 (0.026)	—	[−0.032, 0.070]
Minority						**−0.170 (0.039)**	—	**[−0.246, −0.093]**	**−0.073 (0.032)**	—	**[−0.135, −0.010]**
FRL						**−0.328 (0.042)**	—	**[−0.410, −0.244]**	**−0.165 (0.035)**	—	**[−0.234, −0.096]**
ELL						**−0.279 (0.043)**	—	**[−0.364, −0.195]**	**−0.160 (0.035)**	—	**[−0.229, −0.091]**
SWD						**−0.526 (0.062)**	—	**[−0.647, −0.404]**	**−0.231 (0.051)**	—	**[−0.331, −0.131]**
G1F13 EMSA									**0.670 (0.026)**	—	**[0.618, 0.721]**
G2F13 EMSA									**0.682 (0.025)**	—	**[0.632, 0.731]**
Between classrooms											
G1F13 EMSA									0.397 (0.361)	—	[−0.457, 1.003]
G2F13 EMSA									0.897 (0.815)	—	[−0.462, 3.049]
Between schools											
Treatment			0.103 (0.086)	0.10	[−0.063, 0.279]	0.047 (0.087)	0.05	[−0.119, 0.225]	0.034 (0.104)	0.03	[−0.171, 0.244]
Block[a]	—	—	—	—	—	—	—	—	—	—	—
G1F13									0.382 (0.379)	—	[−0.349, 1.166]
G2F13									0.194 (0.734)	—	[−1.266, 1.687]
Intercept	−0.195 (0.136)	[−0.467, 0.078]	−0.249 (0.142)	—	[−0.533, 0.032]	0.139 (0.147)	—	[−0.153, 0.433]	−0.169 (0.165)	—	[−0.501, 0.150]
Variance components											
Within classroom	**0.587 (0.019)**	**[0.552, 0.625]**	**0.587 (0.019)**		**[0.552, 0.625]**	**0.521 (0.017)**		**[0.490, 0.555]**	**0.133 (0.018)**		**[0.097, 0.169]**
Between classrooms	**0.037 (0.010)**	**[0.021, 0.060]**	**0.037 (0.010)**		**[0.020, 0.060]**	**0.031 (0.009)**		**[0.016, 0.051]**	**0.017 (0.008)**		**[0.002, 0.034]**

(Continued)

	Model 0		Model 1			Model 2			Model 3		
	Estimate (*PSD*)	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI
Between schools	**0.017 (0.026)**	**[0.002, 0.090]**	**0.015 (0.028)**		**[0.002, 0.092]**	**0.018 (0.030)**		**[0.002, 0.097]**	**0.012 (0.035)**		**[0.001, 0.097]**
R-Square											
Within classroom	**0.299 (0.020)**	**[0.260, 0.339]**	**0.300 (0.020)**		**[0.261, 0.339]**	**0.379 (0.019)**		**[0.341, 0.417]**	**0.836 (0.024)**		**[0.788, 0.883]**
Between classrooms	—	—	—		—	—		—	**0.473 (0.254)**		**[0.039, 0.939]**
Between schools	0.000 (0.000)	[0.000, 0.000]	0.000 (0.000)		[0.000, 0.000]	0.000 (0.000)		[0.000, 0.000]	**0.778 (0.240)**		**[0.112, 0.984]**
Intraclass correlation											
Between classrooms	.058		.058			.054			.105		
Between schools	.027		.023			.032			.074		

Note. Student *N* = 2,172; Teacher *N* = 183; School *N* = 22. Average cluster size for classrooms = 11.869; Average cluster size for schools = 98.727. Other abbreviations as in Table D.1.
[a]Block indicates the vector of *n*–1 randomization blocks. Effects for Block are omitted for visual simplicity.

Table D.3. Treatment Effect on ITBS–MC across Different Models with Covariates for Aggregate Sample

	Model 0		Model 1			Model 2			Model 3		
	Estimate (*PSD*)	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI
Fixed effects											
Within classroom											
Grade 2	**1.198 (0.050)**	**[1.100, 1.297]**	**1.197 (0.050)**	—	**[1.100, 1.295]**	**1.201 (0.049)**	—	**[1.104, 1.296]**	**1.207 (0.050)**	—	**[1.109, 1.305]**
Male						**0.089 (0.032)**	—	**[0.027, 0.152]**	**0.061 (0.029)**	—	**[0.006, 0.118]**
Minority						0.001 (0.039)	—	[−0.076, 0.078]	**0.070 (0.035)**	—	**[0.002, 0.139]**
FRL						**−0.247 (0.043)**	—	**[−0.330, −0.162]**	**−0.120 (0.039)**	—	**[−0.196, −0.043]**
ELL						**−0.115 (0.044)**	—	**[−0.201, −0.030]**	−0.026 (0.039)	—	[−0.101, 0.050]
SWD						**−0.519 (0.063)**	—	**[−0.641, −0.395]**	**−0.287 (0.056)**	—	**[−0.397, −0.176]**
G1F13 EMSA									**0.555 (0.029)**	—	**[0.497, 0.611]**
G2F13 EMSA									**0.500 (0.029)**	—	**[0.443, 0.556]**
Between classrooms											
G1F13 EMSA									**0.818 (0.425)**	—	**[0.029, 1.714]**
G2F13 EMSA									0.760 (1.232)	—	[−1.706, 3.631]
Between schools											
Treatment			−0.070 (0.093)	−0.07	[−0.252, 0.121]	−0.105 (0.091)	−0.11	[−0.280, 0.080]	−0.110 (0.113)	−0.11	[−0.338, 0.114]
Block[a]	—	—	—	—	—	—	—	—	—	—	—
G1F13 EMSA									0.348 (0.456)	—	[−0.557, 1.274]
G2F13 EMSA									0.255 (0.719)	—	[−1.153, 1.721]
Intercept	**−0.463 (0.143)**	**[−0.750, −0.177]**	−0.428 (0.093)	—	[−0.741, 0.124]	−0.265 (0.154)	—	[−0.572, 0.042]	**−0.533 (0.186)**	—	**[−0.905, −0.169]**
Variance components											
Within classroom	**0.555 (0.018)**	**[0.522, 0.591]**	**0.555 (0.018)**		**[0.521, 0.591]**	**0.523 (0.017)**		**[0.492, 0.557]**	**0.281 (0.019)**		**[0.245, 0.318]**
Between classrooms	**0.057 (0.012)**	**[0.037, 0.084]**	**0.057 (0.012)**		**[0.037, 0.084]**	**0.057 (0.012)**		**[0.037, 0.083]**	**0.029 (0.014)**		**[0.004, 0.058]**

(Continued)

	Model 0		Model 1			Model 2			Model 3		
	Estimate (*PSD*)	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI
Between schools	**0.018 (0.027)**	**[0.002, 0.095]**	**0.019 (0.034)**		**[0.002, 0.110]**	**0.016 (0.033)**		**[0.002, 0.103]**	**0.014 (0.046)**		**[0.001, 0.120]**
R-Square											
Within classroom	**0.393 (0.021)**	**[0.350, 0.434]**	**0.392 (0.021)**		**[0.350, 0.433]**	**0.433 (0.021)**		**[0.391, 0.474]**	**0.692 (0.023)**		**[0.646, 0.736]**
Between classrooms	—	—	—		—	—		—	**0.497 (0.245)**		**[0.061, 0.940]**
Between schools	0.000 (0.000)	[0.000, 0.000]	0.000 (0.000)		[0.000, 0.000]	0.000 (0.000)		[0.000, 0.000]	**0.768 (0.250)**		**[0.092, 0.985]**
Intraclass correlation											
Between classrooms	.090		.090			.096			.090		
Between schools	.029		.030			.027			.043		

Note. Student *N* = 2,172; Teacher *N* = 183; School *N* = 22. Average cluster size for classrooms = 11.869; Average cluster size for schools = 98.727. Other abbreviations and designations as in Table D.1.
[a]Block indicates the vector of *n*–1 randomization blocks. Effects for Block are omitted for visual simplicity.

Appendix E. Model Results for Early-Joiners Sample

Table E.1. Treatment Effect on ITBS–MP across Different Models with Covariates for Early-Joiners Sample

	Model 0		Model 1			Model 2			Model 3		
	Estimate (*PSD*)	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI
Fixed effects											
Within classroom											
Grade 2	**1.000 (0.046)**	**[0.911, 1.090]**	**1.001 (0.045)**	—	**[0.913, 1.090]**	**1.001 (0.042)**	—	**[0.918, 1.083]**	**1.003 (0.043)**	—	**[0.918, 1.087]**
Male						0.046 (0.032)	—	[−0.017, 0.110]	0.018 (0.026)	—	[−0.033, 0.070]
Minority						**−0.166 (0.039)**	—	**[−0.243, −0.089]**	**−0.069 (0.032)**	—	**[−0.133, −0.006]**
FRL						**−0.328 (0.043)**	—	**[−0.413, −0.243]**	**−0.162 (0.036)**	—	**[−0.232, −0.092]**
ELL						**−0.283 (0.044)**	—	**[−0.368, −0.197]**	**−0.161 (0.036)**	—	**[−0.232, −0.091]**
SWD						**−0.525 (0.063)**	—	**[−0.648, −0.402]**	**−0.225 (0.052)**	—	**[−0.328, −0.124]**
G1F13 EMSA									**0.673 (0.026)**	—	**[0.621, 0.724]**
G2F13 EMSA									**0.683 (0.025)**	—	**[0.633, 0.731]**
Between classrooms											
G1F13 EMSA									0.342 (0.396)	—	[−0.612, 0.906]
G2F13 EMSA									0.892 (0.853)	—	[−0.549, 3.126]
Between schools											
Treatment			0.097 (0.086)	0.10	[−0.071, 0.274]	0.041 (0.086)	0.04	[−0.127, 0.217]	0.028 (0.105)	0.03	[−0.175, 0.239]
Block[a]	—	—	—	—	—	—	—	—	—	—	—
G1F13 EMSA									0.371 (0.389)	—	[−0.381, 1.162]
G2F13 EMSA									0.205 (0.758)	—	[−1.251, 1.737]
Intercept	−0.204 (0.135)	[−0.474, 0.068]	−0.254 (0.143)	—	[−0.540, 0.028]	0.130 (0.146)	—	[−0.161, 0.415]	−0.173 (0.170)	—	[−0.511, 0.162]
Variance components											
Within classroom	**0.589 (0.019)**	**[0.553, 0.627]**	**0.588 (0.019)**		**[0.553, 0.627]**	**0.523 (0.017)**		**[0.491, 0.558]**	**0.130 (0.018)**		**[0.094, 0.166]**
Between classrooms	**0.035 (0.010)**	**[0.018, 0.057]**	**0.035 (0.010)**		**[0.018, 0.057]**	**0.028 (0.009)**		**[0.014, 0.048]**	**0.016 (0.008)**		**[0.002, 0.033]**

(Continued)

	Model 0		Model 1			Model 2			Model 3		
	Estimate (*PSD*)	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI
Between schools	**0.017 (0.026)**	**[0.002, 0.088]**	**0.016 (0.029)**		**[0.002, 0.094]**	**0.018 (0.029)**		**[0.002, 0.097]**	**0.013 (0.041)**		**[0.001, 0.100]**
R-Square											
Within classroom	**0.298 (0.020)**	**[0.259, 0.338]**	**0.298 (0.020)**		**[0.260, 0.338]**	**0.378 (0.019)**		**[0.340, 0.415]**	**0.839 (0.024)**		**[0.792, 0.886]**
Between classrooms	—	—	—		—	—		—	**0.453 (0.251)**		**[0.035, 0.928]**
Between schools	0.000 (0.000)	[0.000, 0.000]	0.000 (0.000)		[0.000, 0.000]	0.000 (0.000)		[0.000, 0.000]	**0.770 (0.242)**		**[0.107, 0.983]**
Intraclass correlation											
Between classrooms	.055		.055			.049			.101		
Between schools	.027		.025			.032			.082		

Note. Student *N* = 2,120; Teacher *N* = 183; School *N* = 22. Average cluster size for classrooms = 11.585; Average cluster size for schools = 96.364. Other abbreviations and designations as in Table D.1.
[a]Block indicates the vector of *n*–1 randomization blocks. Effects for Block are omitted for visual simplicity.

Table E.2. Treatment Effect on ITBS–MC across Different Models with Covariates for Early-Joiners Sample

	Model 0		Model 1			Model 2			Model 3		
	Estimate (*PSD*)	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI
Fixed effects											
Within classroom											
Grade 2	**1.194 (0.050)**	**[1.096, 1.293]**	**1.193 (0.050)**	—	**[1.096, 1.291]**	**1.197 (0.049)**	—	**[1.101, 1.293]**	**1.203 (0.050)**	—	**[1.105, 1.299]**
Male						**0.084 (0.032)**	—	**[0.021, 0.148]**	**0.057 (0.029)**	—	**[0.000, 0.113]**
Minority						0.002 (0.040)	—	[−0.076, 0.080]	**0.072 (0.036)**	—	**[0.002, 0.142]**
FRL						**−0.262 (0.044)**	—	**[−0.348, −0.176]**	**−0.128 (0.040)**	—	**[−0.206, −0.051]**
ELL						**−0.112 (0.044)**	—	**[−0.199, −0.025]**	−0.022 (0.039)	—	[−0.100, 0.055]
SWD						**−0.521 (0.064)**	—	**[−0.645, −0.397]**	**−0.285 (0.057)**	—	**[−0.398, −0.173]**
G1F13 EMSA									**0.558 (0.029)**	—	**[0.499, 0.613]**
G2F13 EMSA									**0.499 (0.029)**	—	**[0.441, 0.555]**
Between classrooms											
G1F13 EMSA									0.769 (0.488)	—	[−0.092, 1.677]
G2F13 EMSA									0.818 (1.216)	—	[−1.269, 4.018]
Between schools											
Treatment			−0.073 (0.095)	−0.07	[−0.257, 0.120]	−0.109 (0.089)	−0.11	[−0.283, 0.072]	−0.116 (0.112)	−0.12	[−0.338, 0.106]
Block[a]	—	—	—	—	—	—	—	—	—	—	—
G1F13 EMSA									0.357 (0.459)	—	[−0.556, 1.282]
G2F13 EMSA									0.255 (0.726)	—	[−1.164, 1.704]
Intercept	**−0.486 (0.146)**	**[−0.778, −0.195]**	−0.449 (0.157)	—	[−0.766, 0.142]	−0.278 (0.152)	—	[−0.582, 0.017]	**−0.542 (0.186)**	—	**[−0.910, −0.175]**
Variance components											
Within classroom	**0.558 (0.018)**	**[0.524, 0.595]**	**0.558 (0.018)**		**[0.524, 0.594]**	**0.525 (0.017)**		**[0.493, 0.560]**	**0.282 (0.019)**		**[0.245, 0.320]**
Between classrooms	**0.055 (0.012)**	**[0.035, 0.082]**	**0.055 (0.012)**		**[0.035, 0.082]**	**0.055 (0.012)**		**[0.036, 0.081]**	**0.028 (0.014)**		**[0.004, 0.056]**

(Continued)

	Model 0		Model 1			Model 2			Model 3		
	Estimate (*PSD*)	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI
Between schools	**0.019 (0.028)**	**[0.002, 0.099]**	**0.019 (0.035)**		**[0.002, 0.114]**	**0.016 (0.031)**		**[0.001, 0.101]**	**0.014 (0.048)**		**[0.001, 0.119]**
R-Square											
Within classroom	**0.390 (0.021)**	**[0.347, 0.431]**	**0.389 (0.021)**		**[0.347, 0.430]**	**0.432 (0.021)**		**[0.390, 0.473]**	**0.691 (0.023)**		**[0.644, 0.736]**
Between classrooms	—	—	—		—	—		—	**0.488 (0.240)**		**[0.056, 0.938]**
Between schools	0.000 (0.000)	[0.000, 0.000]	0.000 (0.000)		[0.000, 0.000]	0.000 (0.000)		[0.000, 0.000]	**0.769 (0.251)**		**[0.093, 0.985]**
Intraclass correlation											
Between classrooms	.087		.087			.092			.086		
Between schools	.030		.030			.027			.043		

Note. Student *N* = 2,120; Teacher *N* = 183; School *N* = 22. Average cluster size for classrooms = 11.585; Average cluster size for schools = 96.364. Other abbreviations and designations as in Table D.1.
[a]Block indicates the vector of *n*–1 randomization blocks. Effects for Block are omitted for visual simplicity.

Appendix F. Model Results with Maximum Likelihood Estimation

Table F.1. Treatment Effect on MPAC across Different Models with Covariates for Aggregate Sample by Maximum Likelihood Estimation

	Model 0		Model 1			Model 2			Model 3		
	Estimate (*SE*)	*p*	Estimate (*SE*)	Effect size	*p*	Estimate (*SE*)	Effect size	*p*	Estimate (*SE*)	Effect size	*p*
Fixed effects											
Within classroom											
Grade 2	**0.697 (0.073)**	**<.001**	**0.689 (0.072)**	**—**	**<.001**	**0.656 (0.069)**	**—**	**<.001**	**0.651 (0.073)**	**—**	**<.001**
Male						0.126 (0.066)	—	.054	**0.111 (0.048)**	**—**	**.021**
Minority						**−0.162 (0.081)**	**—**	**.046**	−0.087 (0.061)	—	.152
FRL						**−0.238 (0.090)**	**—**	**.008**	−0.081 (0.068)	—	.236
ELL						**−0.334 (0.089)**	**—**	**<.001**	**−0.168 (0.067)**	**—**	**.012**
SWD						**−0.580 (0.135)**	**—**	**<.001**	**−0.280 (0.102)**	**—**	**.006**
G1F13 EMSA									**0.803 (0.054)**	**—**	**<.001**
G2F13 EMSA									**0.853 (0.044)**	**—**	**<.001**
Between classrooms											
G1F13 EMSA									−0.883 (1.586)	—	.578
G2F13 EMSA									0.697 (1.144)	—	.543
Between schools											
Treatment			**0.207 (0.078)**	**0.22**	**.008**	**0.152 (0.075)**	**0.16**	**.041**	0.096 (0.067)	0.10	.153
Block[a]	—	—	—	—	—	—	—	—	—	—	—
G1F13 EMSA									0.303 (0.245)	—	.217
G2F13 EMSA									0.454 (0.241)	—	.060
Intercept	0.096 (0.124)	.442	−0.007 (0.128)	—	.956	**0.300 (0.141)**	**—**	**.034**	−0.014 (0.128)	—	.912
Variance components											
Within classroom	**0.707 (0.046)**	**<.001**	**0.704 (0.046)**		**<.001**	**0.639 (0.042)**		**<.001**	0.012 (0.042)		.776
Between classrooms	0.020 (0.027)	.464	0.015 (0.026)		.567	0.016 (0.024)		.500	0.014 (0.036)		.698

(Continued)

	Model 0		Model 1			Model 2			Model 3		
	Estimate (*SE*)	*p*	Estimate (*SE*)	Effect size	*p*	Estimate (*SE*)	Effect size	*p*	Estimate (*SE*)	Effect size	*p*
Between schools	0.001 (0.013)	.955	0.000 (0.006)		.966	0.000 (0.008)		.968	0.000 (0.008)		.990
R-Square											
Within classroom	**0.146 (0.028)**	**<.001**	**0.143 (0.028)**		**<.001**	**0.210 (0.029)**		**<.001**	**0.985 (0.055)**		**<.001**
Between classrooms	—	—	—		—	—		—	0.517 (1.074)		.630
Between schools	**0.994 (0.098)**	**<.001**	**0.998 (0.042)**		**<.001**	**0.995 (0.124)**		**<.001**	**0.997 (0.197)**		**<.001**
Intraclass correlation											
Between classrooms	.027		.021			.024			.538		
Between schools	.001		.000			.000			.000		

Note. Student *N* = 622; Teacher *N* = 167; School *N* = 22. FRL = Free/Reduced-price Lunch; ELL = English Language Learner; SWD = Student with Disability; G1F13 EMSA = Grade 1, fall 2013 baseline mathematics test; G2F13 EMSA = Grade 2 baseline mathematics test, fall 2013. Estimator setting used Mplus ML maximum likelihood parameter estimates with conventional standard errors. Reported estimates are from the unstandardized solution. Only the effect size for Treatment is presented; it is calculated as Hedges' *g*. Boldface indicates *p* <.05. Average cluster size for classrooms = 3.725; Average cluster size for schools = 28.273. Estimator setting used Mplus ML maximum likelihood parameter estimates with conventional standard errors. Reported estimates are from the unstandardized solution.
[a]Block indicates the vector of *n*–1 randomization blocks. Effects for Block are omitted visual simplicity.

Table F.2. Treatment Effect on ITBS–MP across Different Models with Covariates for Aggregate Sample using Maximum Likelihood Estimation

	Model 0		Model 1			Model 2			Model 3		
	Estimate (*SE*)	*p*	Estimate (*SE*)	Effect size	*p*	Estimate (*SE*)	Effect size	*p*	Estimate (*SE*)	Effect size	*p*
Fixed effects											
Within classroom											
Grade 2	**0.992 (0.044)**	**<.001**	**0.991 (0.043)**	**—**	**<.001**	**0.995 (0.041)**	**—**	**<.001**	**0.991 (0.042)**	**—**	**<.001**
Male						0.039 (0.032)	—	.216	0.012 (0.026)	—	.632
Minority						**−0.170 (0.039)**	**—**	**<.001**	**−0.077 (0.032)**	**—**	**.016**
FRL						**−0.337 (0.042)**	**—**	**<.001**	**−0.178 (0.035)**	**—**	**<.001**
ELL						**−0.283 (0.043)**	**—**	**<.001**	**− 0.163 (0.035)**	**—**	**<.001**
SWD						**−0.524 (0.062)**	**—**	**<.001**	**− 0.228 (0.051)**	**—**	**<.001**
G1F13 EMSA									**0.668 (0.026)**	**—**	**<.001**
G2F13 EMSA									**0.679 (0.025)**	**—**	**<.001**
Between classrooms											
G1F13 EMSA									0.463 (0.273)	—	.090
G2F13 EMSA									0.970 (0.498)	—	.051
Between schools											
Treatment			**0.092 (0.046)**	**0.09**	**.043**	0.032 (0.048)	0.03	.509	0.019 (0.047)	0.02	.689
Block[a]	—	—	—	—	—	—	—	—	—	—	—
G1F13 EMSA									**0.336 (0.167)**	**—**	**.044**
G2F13 EMSA									0.257 (0.224)	—	.292
Intercept	**−0.186 (0.078)**	**.017**	**−0.232 (0.081)**	**—**	**.004**	0.162 (0.082)	—	.050	−0.130 (0.082)	—	.111
Variance components											
Within classroom	**0.992 (0.044)**	**<.001**	**0.586 (0.019)**		**<.001**	**0.519 (0.016)**		**<.001**	**0.134 (0.018)**		**<.001**
Between classrooms	**0.033 (0.009)**	**<.001**	**0.032 (0.009)**		**<.001**	**0.027 (0.008)**		**.001**	0.012 (0.010)		.207
Between schools	0.000 (0.003)	.965	0.000 (0.003)		.972	0.000 (0.005)		.973	0.000 (0.003)		.969

(Continued)

	Model 0		Model 1			Model 2			Model 3		
	Estimate (*SE*)	*p*	Estimate (*SE*)	Effect size	*p*	Estimate (*SE*)	Effect size	*p*	Estimate (*SE*)	Effect size	*p*
R-Square											
Within classroom	**0.296 (0.020)**	**<.001**	**0.295 (0.020)**		**<.001**	**0.377 (0.019)**		**<.001**	**0.832 (0.024)**		**<.001**
Between classrooms	—	—	—		—	—		—	0.581 (0.367)		.114
Between schools	**0.999 (0.027)**	**<.001**	**0.999 (0.024)**		**<.001**	**0.997 (0.101)**		**<.001**	**0.996 (0.093)**		**<.001**
Intraclass correlation											
Between classrooms	.032			.052			.049			.082	
Between schools	.000			.000			.000			.000	

Note. Student *N* = 2,172; Teacher *N* = 183; School *N* = 22. Average cluster size for classrooms = 11.869; Average cluster size for schools = 98.727. Abbreviations and other notes as in Table F.1.
[a]Block indicates the vector of *n*–1 randomization blocks. Effects for Block are omitted for visual simplicity.

Table F.3. Treatment Effect on ITBS–MC across Different Models with Covariates for Aggregate Sample using Maximum Likelihood Estimation

	Model 0		Model 1			Model 2			Model 3		
	Estimate (*SE*)	*p*	Estimate (*SE*)	Effect size	*p*	Estimate (*SE*)	Effect size	*p*	Estimate (*SE*)	Effect size	*p*
Fixed effects											
Within classroom											
Grade 2	**1.175 (0.046)**	**<.001**	**1.171 (0.047)**	**—**	**<.001**	**1.166 (0.047)**	**—**	**<.001**	**1.162 (0.048)**	**—**	**<.001**
Male						**0.077 (0.032)**	**—**	**.017**	0.047 (0.028)	—	.102
Minority						−0.014 (0.039)	—	.724	0.051 (0.035)	—	.147
FRL						**−0.270 (0.043)**	**—**	**<.001**	**−0.151 (0.039)**	**—**	**<.001**
ELL						**−0.116 (0.043)**	**—**	**.007**	−0.024 (0.039)	—	.543
SWD						**−0.520 (0.062)**	**—**	**<.001**	**−0.288 (0.056)**	**—**	**<.001**
G1F13 EMSA									**0.551 (0.029)**	**—**	**<.001**
G2F13 EMSA									**0.494 (0.029)**	**—**	**<.001**
Between classrooms											
G1F13 EMSA									**0.857 (0.244)**	**—**	**<.001**
G2F13 EMSA									0.920 (0.847)	—	.278
Between schools											
Treatment			−0.077 (0.052)	−0.08	.135	**−0.113 (0.051)**	**−0.11**	**.026**	**−0.113 (0.049)**	**−0.11**	**.021**
Block[a]	—	—	—	—	—	—	—	—	—	—	—
G1F13 EMSA									**0.364 (0.179)**	**—**	**.042**
G2F13 EMSA									0.219 (0.226)	—	.331
Intercept	**−0.449 (0.085)**	**<.001**	**−0.404 (0.089)**	**—**	**<.001**	**−0.206 (0.093)**	**—**	**.027**	**−0.465 (0.094)**	**—**	**<.001**
Variance components											
Within classroom	**0.554 (0.018)**	**<.001**	**0.554 (0.018)**		**<.001**	**0.521 (0.017)**		**<.001**	**0.282 (0.019)**		**<.001**
Between classrooms	**0.052 (0.010)**	**<.001**	**0.051 (0.010)**		**<.001**	**0.050 (0.010)**		**<.001**	0.022 (0.016)		.181
Between schools	0.000 (0.008)	.987	0.000 (0.003)		.969	0.000 (0.002)		.966	0.000 (0.002)		.950

(Continued)

	Model 0		Model 1			Model 2			Model 3		
	Estimate (*SE*)	*p*	Estimate (*SE*)	Effect size	*p*	Estimate (*SE*)	Effect size	*p*	Estimate (*SE*)	Effect size	*p*
R-Square											
Within classroom	**0.384 (0.021)**	**<.001**	**0.382 (0.022)**		**<.001**	**0.421 (0.021)**		**<.001**	**0.679 (0.024)**		**<.001**
Between classrooms	—	—	—		—	—		—	0.583 (0.332)		.079
Between schools	**0.996 (0.247)**	**<.001**	**0.997 (0.070)**		**<.001**	**0.994 (0.130)**		**<.001**	**0.995 (0.071)**		**<.001**
Intraclass correlation											
Between classrooms	.086			.084			.088			.072	
Between schools	.000			.000			.000			.000	

Note. Student *N* = 2,172; Teacher *N* = 183; School *N* = 22. Average cluster size for classrooms = 11.869; Average cluster size for schools = 98.727. Abbreviations and other notes as in Table F.1.
[a]Block indicates the vector of *n*–1 randomization blocks. Effects for Block are omitted for visual simplicity.

Appendix G. Model Results for Subgroup Analyses

Table G.1. Treatment Effect on MPAC, ITBS–MP, and ITBS–MC for Grade 1 Students

	MPAC[a]			ITBS–MP[b]			ITBS–MC[b]		
	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI
Fixed effects									
Within classroom									
Male	−0.005 (0.073)	—	[−0.147, 0.137]	**−0.135 (0.044)**	—	**[−0.221, −0.048]**	−0.031 (0.051)	—	[−0.131, 0.070]
Minority	−0.073 (0.098)	—	[−0.263, 0.120]	**−0.121 (0.055)**	—	**[−0.228, −0.012]**	−0.012 (0.064)	—	[−0.137, 0.114]
FRL	0.059 (0.102)	—	[−0.142, 0.258]	**−0.088 (0.060)**	—	**[−0.205, −0.030]**	**−0.152 (0.069)**	—	**[−0.288, −0.015]**
ELL	−0.186 (0.103)	—	[−0.389, 0.017]	**−0.156 (0.060)**	—	**[−0.273, −0.040]**	−0.091 (0.070)	—	[−0.227, 0.046]
SWD	**−0.398 (0.156)**	—	**[−0.703, −0.093]**	**−0.262 (0.093)**	—	**[−0.443, −0.081]**	**−0.435 (0.107)**	—	**[−0.644, −0.224]**
G1F13 EMSA	**0.869 (0.063)**	—	**[0.747, 0.992]**	**0.813 (0.037)**	—	**[0.741, 0.884]**	**0.618 (0.043)**	—	**[0.534, 0.701]**
Between classrooms									
G1F13 EMSA	−0.059 (2.530)	—	[−5.346, 5.925]	0.464 (0.438)	—	[−0.481, 1.180]	0.725 (0.575)	—	[−0.238, 1.810]
Between schools									
Treatment	0.245 (0.239)	0.25	[−0.230, 0.723]	0.133 (0.172)	0.14	[−0.198, 0.484]	0.028 (0.198)	0.03	[−0.347, 0.435]
Block[c]	—	—	—	—	—	—	—	—	—
G1F13 EMSA	0.541 (1.017)	—	[−1.495, 2.563]	0.858 (0.647)	—	[−0.412, 2.155]	0.992 (0.770)	—	[−0.522, 2.541]
Intercept	0.106 (0.386)	—	[−0.643, 0.889]	0.214 (0.259)	—	[−0.299, 0.719]	−0.026 (0.300)	—	[−0.623, 0.559]
Variance components									
Within classroom	**0.423 (0.039)**	—	**[0.356, 0.507]**	**0.505 (0.023)**	—	**0.463, 0.554]**	**0.690 (0.031)**	—	**[0.632, 0.755]**
Between classrooms	**0.065 (0.041)**	—	**[0.005, 0.160]**	**0.032 (0.015)**	—	**[0.009, 0.066]**	**0.050 (0.020)**	—	**[0.019, 0.096]**
Between schools	**0.062 (0.192)**	—	**[0.003, 0.536]**	**0.045 (0.117)**	—	**[0.006, 0.315]**	**0.052 (0.168)**	—	**[0.004, 0.430]**
R-Square									
Within classroom	**0.459 (0.044)**	—	**[0.371, 0.542]**	**0.373 (0.024)**	—	**[0.325, 0.421]**	**0.211 (0.023)**	—	**[0.168, 0.256]**

(Continued)

	MPAC[a]			ITBS–MP[b]			ITBS–MC[b]		
	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI
Between classrooms	0.159 (0.268)	—	[0.000, 0.916]	**0.150 (0.167)**	—	**[0.001, 0.582]**	**0.197 (0.179)**	—	**[0.001, 0.636]**
Between schools	**0.373 (0.308)**	—	**[0.001, 0.956]**	**0.529 (0.292)**	—	**[0.004, 0.948]**	**0.570 (0.306)**	—	**[0.004, 0.971]**
Intraclass correlation									
Between classrooms		.118			.055			.063	
Between schools		.113			.077			.066	

Note. FRL = Free/Reduced−price Lunch; ELL = English Language Learner; SWD = Student with Disability. G1F13 EMSA = Grade 1, fall 2013; G2F13 EMSA = Grade 2, fall 2013; PSD = the standard deviation of the posterior distribution. 95% CI = 95% credibility intervals of the posterior distribution with equal tail percentages. Reported estimates are from the unstandardized solution. Only the effect size for Treatment is presented, which is calculated as Hedges' *g*. Boldface indicates the 95% CI does not include zero.

[a]MPAC analysis sample size: Student *N* = 336; Teacher *N* = 88; School *N* = 21. MPAC analysis average cluster size: Teacher *N* = 3.818; School *N* = 16.000.

[b]ITBS analyses sample size: Student *N* = 1103; Teacher *N* = 96; School *N* = 21. ITBS analyses average cluster size: Teacher *N* = 11.490; School *N* = 52.524.

[c]Block indicates the vector of *n*–1 randomization blocks. Effects for Block are omitted for visual simplicity.

Table G.2. Treatment Effect on MPAC, ITBS–MP, and ITBS–MC for Grade 2 Students

	MPAC[a]			ITBS–MP[b]			ITBS–MC[b]		
	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI
Fixed effects									
Within classroom									
Male	**0.272 (0.074)**	**—**	**[0.127, 0.417]**	**0.164 (0.041)**	**—**	**[0.083, 0.245]**	**0.168 (0.050)**	**—**	**[0.070, 0.267]**
Minority	−0.062 (0.095)	—	[−0.248, 0.125]	−0.061 (0.050)	—	[−0.160, 0.037]	**0.144 (0.062)**	**—**	**[0.023, 0.264]**
FRL	−0.166 (0.113)	—	[−0.388, 0.055]	**−0.285 (0.056)**	**—**	**[−0.394, −0.175]**	**−0.171 (0.069)**	**—**	**[−0.306, −0.037]**
ELL	−0.164 (0.107)	—	[−0.374, 0.046]	**−0.220 (0.057)**	**—**	**[−0.332, −0.107]**	0.002 (0.070)	—	[−0.135, 0.139]
SWD	−0.254 (0.169)	—	[−0.584, 0.077]	**−0.293 (0.078)**	**—**	**[−0.446, −0.141]**	**−0.302 (0.095)**	**—**	**[−0.489, −0.115]**
G2F13 EMSA	**0.928 (0.059)**	**—**	**[0.813, 1.043]**	**0.772 (0.032)**	**—**	**[0.708, 0.835]**	**0.679 (0.040)**	**—**	**[0.601, 0.756]**
Between classrooms									
G2F13 EMSA	0.546 (1.778)	—	[−3.462, 4.462]	0.921 (1.344)	—	[−1.921, 4.054]	0.763 (2.218)	—	[−3.965, 5.867]
Between schools									
Treatment	−0.013 (0.168)	−0.01	[−0.352, 0.314]	−0.069 (0.118)	−0.07	[−0.302, 0.168]	−0.287 (0.150)	−0.29	[−0.589, 0.006]
Block[c]	—	—	—	—	—	—	—	—	—
G2F13 EMSA	0.803 (0.704)	—	[−0.630, 2.186]	0.874 (0.779)	—	[−0.699, 2.420]	0.562 (0.976)	—	[−1.371, 2.515]
Intercept	0.385 (0.274)	—	[−0.149, 0.941]	**0.500 (0.193)**	**—**	**[0.114, 0.881]**	0.187 (0.256)	—	[−0.313, 0.701]
Variance components									
Within classroom	**0.354 (0.034)**	**—**	**[0.295, 0.430]**	**0.410 (0.019)**	**—**	**[0.374, 0.450]**	**0.610 (0.029)**	**—**	**[0.557, 0.669]**
Between classrooms	**0.031 (0.025)**	**—**	**[0.003, 0.097]**	**0.023 (0.012)**	**—**	**[0.003, 0.052]**	**0.060 (0.026)**	**—**	**[0.011, 0.116]**
Between schools	**0.031 (0.098)**	**—**	**[0.002, 0.271]**	**0.020 (0.051)**	**—**	**[0.002, 0.151]**	**0.027 (0.081)**	**—**	**[0.002, 0.237]**
R-Square									
Within classroom	**0.562 (0.041)**	**—**	**[0.476, 0.638]**	**0.442 (0.023)**	**—**	**[0.396, 0.488]**	**0.288 (0.025)**	**—**	**[0.239, 0.338]**

(Continued)

	MPAC[a]			ITBS–MP[b]			ITBS–MC[b]		
	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI
Between classrooms	**0.243 (0.281)**	—	**[0.001, 0.912]**	**0.259 (0.259)**	—	**[0.001, 0.887]**	0.133 (0.231)	—	[0.000, 0.835]
Between schools	**0.738 (0.312)**	—	**[0.007, 0.989]**	**0.814 (0.303)**	—	**[0.011, 0.992]**	**0.686 (0.330)**	—	**[0.004, 0.989]**
Intraclass correlation									
Between classrooms		.075			.051			.086	
Between schools		.075			.044			.039	

Note. Abbreviations and notes as in Table G.1.

[a]MPAC analysis sample size: Student *N* = 286; Teacher *N* = 79; School *N* = 22. MPAC analysis average cluster size: Teacher *N* = 3.620; School *N* = 13.000.

[b]ITBS analyses sample size: Student *N* = 1069; Teacher *N* = 88; School *N* = 22. ITBS analyses average cluster size: Teacher *N* = 12.148; School *N* = 48.591.

[c]Block indicates the vector of *n*–1 randomization blocks. Effects for Block are omitted for visual simplicity.

Table G.3. Treatment Effect on MPAC, ITBS–MP, and ITBS–MC for Female Students

	MPAC[a]			ITBS–MP[b]			ITBS–MC[b]		
	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI
Fixed effects									
Within classroom									
Grade 2	**0.587 (0.107)**	**—**	**[0.370, 0.795]**	**0.972 (0.057)**	**—**	**[0.859, 1.084]**	**1.228 (0.061)**	**—**	**[1.109, 1.346]**
Minority	−0.025 (0.096)	—	[−0.207, 0.168]	−0.022 (0.048)	—	[−0.117, 0.072]	**0.133 (0.052)**	**—**	**[0.031, 0.235]**
FRL	0.005 (0.108)	—	[−0.214, 0.212]	**−0.155 (0.054)**	**—**	**[−0.259, −0.049]**	**−0.148 (0.058)**	**—**	**[−0.262, −0.034]**
ELL	**−0.284 (0.109)**	**—**	**[−0.499, −0.073]**	**−0.248 (0.053)**	**—**	**[−0.351, −0.143]**	−0.083 (0.057)	—	[−0.194, 0.030]
SWD	**−0.601 (0.254)**	**—**	**[−1.098, −0.107]**	**−0.221 (0.101)**	**—**	**[−0.417, −0.023]**	**−0.264 (0.109)**	**—**	**[−0.477, −0.050]**
G1F13 EMSA	**0.798 (0.069)**	**—**	**[0.660, 0.928]**	**0.660 (0.038)**	**—**	**[0.584, 0.733]**	**0.579 (0.041)**	**—**	**[0.494, 0.656]**
G2F13 EMSA	**0.847 (0.057)**	**—**	**[0.734, 0.958]**	**0.674 (0.038)**	**—**	**[0.598, 0.747]**	**0.471 (0.043)**	**—**	**[0.386, 0.555]**
Between classrooms									
Grade 2									
G1F13 EMSA	0.208 (1.249)	—	[−2.334, 3.017]	0.597 (0.905)	—	[−1.566, 2.432]	0.576 (1.246)	—	[−2.335, 3.131]
G2F13 EMSA	0.359 (1.292)	—	[−2.654, 2.997]	1.708 (1.318)	—	[−1.128, 4.416]	0.323 (2.025)	—	[−4.160, 4.670]
Between schools									
Treatment	0.105 (0.194)	0.11	[−0.282, 0.488]	0.073 (0.133)	0.07	[−0.187, 0.344]	−0.061 (0.146)	−0.06	[−0.341, 0.242]
Block[c]	—	—	—	—	—	—	—	—	—
G1F13 EMSA	−0.075 (1.452)	—	[−2.729, 2.691]	0.525 (0.462)	—	[−0.353, 1.488]	0.187 (0.674)	—	[−1.185, 1.542]
G2F13 EMSA	0.551 (0.831)	—	[−1.117, 2.157]	0.178 (1.042)	—	[−1.839, 2.286]	0.301 (0.838)	—	[−1.388, 1.956]
Intercept	0.048 (0.366)	—	[−0.652, 0.799]	−0.191 (0.227)	—	[−0.645, 0.249]	−0.440 (0.253)	—	[−0.950, 0.057]
Variance components									
Within classroom	**0.050 (0.040)**	**—**	**[0.004, 0.149]**	**0.155 (0.027)**	**—**	**[0.103, 0.207]**	**0.288 (0.028)**	**—**	**[0.234, 0.343]**
Between classrooms	**0.012 (0.014)**	**—**	**[0.001, 0.053]**	**0.013 (0.011)**	**—**	**[0.001, 0.041]**	**0.031 (0.018)**	**—**	**[0.003, 0.071]**
Between schools	**0.039 (0.137)**	**—**	**[0.002, 0.369]**	**0.024 (0.077)**	**—**	**[0.002, 0.202]**	**0.033 (0.099)**	**—**	**[0.002, 0.276]**

(Continued)

	MPAC[a]			ITBS–MP[b]			ITBS–MC[b]		
	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI
R-Square									
Within classroom	**0.941 (0.049)**	—	**[0.817, 0.996]**	**0.809 (0.035)**	—	**[0.739, 0.877]**	**0.695 (0.032)**	—	**[0.630, 0.756]**
Between classrooms	**0.490 (0.279)**	—	**[0.027, 0.949]**	**0.700 (0.255)**	—	**[0.084, 0.971]**	**0.437 (0.272)**	—	**[0.024, 0.948]**
Between schools	**0.664 (0.282)**	—	**[0.047, 0.983]**	**0.726 (0.254)**	—	**[0.088, 0.984]**	**0.568 (0.283)**	—	**[0.034, 0.973]**
Intraclass correlation									
Between classrooms		.119			.068			.088	
Between schools		.386			.125			.094	

Note. Abbreviations and notes as in Table G.1.

[a]MPAC analysis sample size: Student *N* = 319; Teacher *N* = 167; School *N* = 22. MPAC analysis average cluster size: Teacher *N* = 1.910; School *N* = 14.500.

[b]ITBS analyses sample size: Student *N* = 1086; Teacher *N* = 183; School *N* = 22. Teacher *N* = 183; School *N* = 22. ITBS analyses average cluster size: Teacher *N* = 5.934; School *N* = 49.364.

[c]Block indicates the vector of *n*–1 randomization blocks. Effects for Block are omitted for visual simplicity.

Table G.4. Treatment Effect on MPAC, ITBS–MP, and ITBS–MC for Male Students

	MPAC[a]			ITBS–MP[b]			ITBS–MC[b]		
	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI
Fixed effects									
Within classroom									
Grade 2	**0.825 (0.100)**	**—**	**[0.631, 1.026]**						
Minority	−0.130 (0.095)	—	[−0.317, 0.053]	**−0.125 (0.044)**	**—**	**[−0.211, −0.040]**	−0.002 (0.049)	—	[−0.098, 0.095]
FRL	−0.118 (0.103)	—	[−0.319, 0.083]	**−0.185 (0.048)**	**—**	**[−0.279, −0.091]**	−0.096 (0.054)	—	[−0.202, 0.009]
ELL	−0.104 (0.096)	—	[−0.290, 0.087]	**−0.098 (0.047)**	**—**	**[−0.190, −0.005]**	0.022 (0.054)	—	[−0.084, 0.126]
SWD	**−0.261 (0.119)**	**—**	**[−0.495–0.024]**	**−0.225 (0.059)**	**—**	**[−0.341, −0.109]**	**−0.289 (0.068)**	**—**	**[−0.422, −0.157]**
G1F13 EMSA	**0.703 (0.070)**	**—**	**[0.559, 0.831]**	**0.682 (0.037)**	**—**	**[0.607, 0.753]**	**0.564 (0.043)**	**—**	**[0.476, 0.644]**
G2F13 EMSA	**0.818 (0.064)**	**—**	**[0.692, 0.943]**	**0.715 (0.035)**	**—**	**[0.644, 0.783]**	**0.540 (0.042)**	**—**	**[0.457, 0.622]**
Between classrooms									
Grade 2				**1.040 (0.053)**	**—**	**[0.935, 1.144]**	**1.200 (0.059)**	**—**	**[1.081, 1.315]**
G1F13 EMSA	0.361 (1.510)	—	[−2.932, 3.550]	0.120 (0.557)	—	[−1.554, 1.060]	0.535 (0.961)	—	[−1.777, 2.532]
G2F13 EMSA	0.345 (1.146)	—	[−2.541, 2.406]	0.313 (0.250)	—	[−0.253, 0.676]	0.508 (0.461)	—	[−0.210, 1.589]
Between schools									
Treatment	0.049 (0.212)	0.05	[−0.383, 0.458]	0.010 (0.122)	0.01	[−0.227, 0.257]	−0.155 (0.155)	−0.15	[−0.466, 0.158]
Block[c]	—	—	—	—	—	—	—	—	—
G1F13 EMSA	0.360 (0.782)	—	[−1.178, 1.864]	0.221 (0.378)	—	[−0.524, 0.970]	0.353 (0.397)	—	[−0.421, 1.168]
G2F13 EMSA	0.574 (0.777)	—	[−0.971, 2.061]	0.156 (0.557)	—	[−0.950, 1.282]	0.231 (0.681)	—	[−1.125, 1.578]
Intercept	−0.007 (0.319)	—	[−0.637, 0.603]	−0.094 (0.185)	—	[−0.460, 0.271]	**−0.497 (0.209)**	**—**	**[−0.917, −0.092]**
Variance components									
Within classroom	**0.066 (0.045)**	**—**	**[0.004, 0.170]**	**0.095 (0.025)**	**—**	**[0.045, 0.145**	**0.258 (0.027)**	**—**	**[0.205, 0.312]**
Between classrooms	**0.014 (0.017)**	**—**	**[0.002, 0.062]**	**0.013 (0.009)**	**—**	**[0.002, 0.034]**	**0.026 (0.015)**	**—**	**[0.004, 0.059]**
Between schools	**0.038 (0.153)**	**—**	**[0.002, 0.358]**	**0.012 (0.042)**	**—**	**[0.001, 0.102]**	**0.015 (0.052)**	**—**	**[0.001, 0.135]**

(Continued)

	MPAC[a]			ITBS–MP[b]			ITBS–MC[b]		
	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI
R-Square									
Within classroom	**0.919 (0.056)**	—	**[0.786, 0.995]**	**0.824 (0.049)**	—	**[0.727, 0.919]**	**0.524 (0.053)**	—	**[0.423, 0.629]**
Between classrooms	**0.524 (0.283)**	—	**[0.026, 0.953]**	**0.434 (0.254)**	—	**[0.029, 0.919]**	**0.458 (0.248)**	—	**[0.043, 0.925]**
Between schools	**0.683 (0.273)**	—	**[0.056, 0.982]**	**0.678 (0.270)**	—	**[0.055, 0.975]**	**0.762 (0.257)**	—	**[0.085, 0.985]**
Intraclass correlation									
Between classrooms		.119			.108			.087	
Between schools		.030			.100			.050	

Note. Abbreviations and notes as in Table G.1.

[a]MPAC analysis sample size: Student *N* = 303; Teacher *N* = 166; School *N* = 22. MPAC analysis average cluster size: Teacher *N* = 1.825; School *N* = 13.773.

[a]ITBS analyses sample size: Student *N* = 1083; Teacher *N* = 183; School *N* = 22. ITBS analyses average cluster size: Teacher *N* = 5.918; School *N* = 49.227.

[c]Block indicates the vector of *n*–1 randomization blocks. Effects for Block are omitted for visual simplicity.

Table G.5. Treatment Effect on MPAC, ITBS–MP, and ITBS–MC for Nonminority Students

	MPAC[a]			ITBS–MP[b]			ITBS–MC[b]		
	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI
Fixed effects									
Within classroom									
Grade 2	**0.733 (0.150)**	**—**	**[0.450, 1.027]**	**1.092 (0.067)**	**—**	**[0.958, 1.222]**	**1.201 (0.073)**	**—**	**[1.057, 1.342**
Male	0.145 (0.096)	—	[−0.043, 0.331]	0.079 (0.047)	—	[−0.013, 0.171]	**0.138 (0.050)**	**—**	**[0.040, 0.236]**
FRL	−0.124 (0.128)	—	[−0.376, 0.125]	**−0.183 (0.061)**	**—**	**[−0.304, −0.064]**	−0.057 (0.065)	—	[−0.187, 0.069]
ELL	−0.086 (0.351)	—	[−0.778, 0.596]	−0.207 (0.132)	—	[−0.468, 0.052]	0.188 (0.139)	—	[−0.084, 0.461]
SWD	−0.133 (0.200)	—	[−0.525, 0.262]	−0.167 (0.093)	—	[−0.349, 0.017]	**−0.330 (0.100)**	**—**	**[−0.526, −0.132]**
G1F13 EMSA	**0.906 (0.099)**	**—**	**[0.719, 1.105]**	**0.668 (0.050)**	**—**	**[0.568, 0.764]**	**0.574 (0.053)**	**—**	**[0.463, 0.672]**
G2F13 EMSA	**0.923 (0.098)**	**—**	**[0.733, 1.118]**	**0.677 (0.049)**	**—**	**[0.579, 0.772]**	**0.461 (0.054)**	**—**	**[0.354, 0.567]**
Between classrooms									
G1F13 EMSA	0.981 (1.384)	—	[−1.780, 4.009]	−0.164 (0.775)	—	[−1.846, 1.079]	0.971 (1.327)	—	[−2.574, 3.759]
G2F13 EMSA	0.762 (1.235)	—	[−1.656, 2.796]	0.824 (0.755)	—	[−0.678, 2.545]	0.619 (1.048)	—	[−1.125, 3.096]
Between schools	0.146 (0.322)	0.15	[−0.477, 0.772]	0.053 (0.186)	0.05	[−0.291, 0.434]	−0.031 (0.187)	−0.03	[−0.384, 0.346]
Treatment	—	—	—	—	—	—	—	—	—
Block[c]	−0.017 (2.578)	—	[−4.672, 4.417]	0.575 (1.343)	—	[−1.788, 3.271]	0.586 (1.346)	—	[−1.967, 3.266]
G1F13 EMSA	0.932 (4.002)	—	[−7.025, 8.861]	0.770 (2.246)	—	[−3.476, 5.246]	0.734 (2.031)	—	[−3.177, 4.765]
G2F13 EMSA	−0.527 (1.606)	—	[−3.595, 2.499]	−0.806 (0.697)	—	[−2.247, 0.511]	−1.102 (0.675)	—	[−2.474, 0.194]
Intercept	**0.733 (0.150)**	**—**	**[0.450, 1.027]**	**1.092 (0.067)**	**—**	**[0.958, 1.222]**	**1.201 (0.073)**	**—**	**[1.057, 1.342**
Variance components									
Within classroom	**0.040 (0.043)**	**—**	**[0.001, 0.159]**	**0.207 (0.033)**	**—**	**[0.144, 0.272]**	**0.331 (0.034)**	**—**	**[0.266, 0.399]**
Between classrooms	**0.037 (0.039)**	**—**	**[0.002, 0.144]**	**0.013 (0.011)**	**—**	**[0.002, 0.041]**	**0.022 (0.017)**	**—**	**[0.002, 0.065]**
Between schools	**0.077 (0.659)**	**—**	**[0.003, 0.930]**	**0.035 (0.223)**	**—**	**[0.002, 0.383]**	**0.030 (0.209)**	**—**	**[0.002, 0.346]**

(Continued)

	MPAC[a]			ITBS–MP[b]			ITBS–MC[b]		
	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI
R-Square									
Within classroom	**0.953 (0.053)**	—	**[0.804, 0.999]**	**0.766 (0.039)**	—	**[0.688, 0.842]**	**0.650 (0.039)**	—	**[0.572, 0.723]**
Between classrooms	**0.709 (0.257)**	—	**[0.084, 0.987]**	**0.592 (0.262)**	—	**[0.053, 0.955]**	**0.612 (0.268)**	—	**[0.052, 0.968]**
Between schools	**0.554 (0.288)**	—	**[0.030, 0.977]**	**0.625 (0.276)**	—	**[0.044, 0.974]**	**0.644 (0.276)**	—	**[0.046, 0.976]**
Intraclass correlation									
Between classrooms		.240			.051			.057	
Between schools		.500			.137			.078	

Note. Abbreviations and notes as in Table G.1.

[a]MPAC analysis sample size: Student *N* = 207; Teacher *N* = 113; School *N* = 21. MPAC analysis average cluster size: Teacher *N* = 1.832; School *N* = 9.857.

[b]ITBS analyses sample size: Student *N* = 803; Teacher *N* = 161; School *N* = 21. ITBS analyses average cluster size: Teacher *N* = 4.988; School *N* = 38.238.

[c]Block indicates the vector of *n*–1 randomization blocks. Effects for Block are omitted for visual simplicity.

Table G.6. Treatment Effect on MPAC, ITBS–MP, and ITBS–MC for Minority Students

	MPAC[a]			ITBS–MP[b]			ITBS–MC[b]		
	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI
Fixed effects									
Within classroom									
Grade 2	**0.708 (0.093)**	**—**	**[0.525, 0.890]**	**1.019 (0.051)**	**—**	**[0.919, 1.121]**	**1.229 (0.056)**	**—**	**[1.120, 1.339]**
Male	0.116 (0.064)	—	[−0.008, 0.242]	−0.013 (0.035)	—	[−0.081, 0.055]	0.012 (0.036)	—	[−0.060, 0.083]
FRL	−0.032 (0.098)	—	[−0.224, 0.159]	**−0.172 (0.048)**	**—**	**[−0.268, −0.079]**	**−0.167 (0.051)**	**—**	**[−0.268, −0.067]**
ELL	**−0.182 (0.076)**	**—**	**[−0.331, −0.035]**	**−0.172 (0.039)**	**—**	**[−0.247, −0.096]**	−0.053 (0.041)	—	[−0.134, 0.028]
SWD	**−0.377 (0.135)**	**—**	**[−0.644, −0.113]**	**−0.290 (0.067)**	**—**	**[−0.419, −0.159]**	**−0.269 (0.071)**	**—**	**[−0.407, −0.129]**
G1F13 EMSA	**0.762 (0.065)**	**—**	**[0.629, 0.884]**	**0.717 (0.034)**	**—**	**[0.648, 0.783]**	**0.547 (0.039)**	**—**	**[0.468, 0.619]**
G2F13 EMSA	**0.850 (0.057)**	**—**	**[0.736, 0.961]**	**0.733 (0.032)**	**—**	**[0.670, 0.795]**	**0.533 (0.036)**	**—**	**[0.462, 0.603]**
Between classrooms									
G1F13 EMSA	−0.013 (1.240)	—		0.549 (0.729)	—	[−1.577, 1.646]	0.545 (1.166)	—	[−3.028, 2.202]
G2F13 EMSA	0.289 (1.520)	—	[−2.873, 2.599]	0.010 (0.786)	—	[−1.957, 1.401]	−0.068 (1.494)	—	[−3.612, 2.998]
Between schools									
Treatment	0.072 (0.177)	0.07	[−0.290, 0.413]	0.030 (0.112)	0.03	[−0.189, 0.257]	−0.103 (0.127)	−0.10	[−0.347, 0.158]
Block[c]	—	—	—	—	—	—	—	—	—
G1F13 EMSA	0.360 (1.269)	—	[−1.965, 2.908]	0.520 (0.437)	—	[−0.344, 1.401]	0.319 (0.564)	—	[−0.806, 1.452]
G2F13 EMSA	0.469 (0.694)	—	[−0.936, 1.829]	0.291 (0.665)	—	[−1.031, 1.587]	0.081 (0.672)	—	[−1.261, 1.415]
Intercept	0.090 (0.290)	—	[−0.479, 0.666]	0.021 (0.203)	—	[−0.379, 0.422]	−0.228 (0.200)	—	[−0.630, 0.159]
Variance components									
Within classroom	**0.074 (0.044)**	**—**	**[0.005, 0.172]**	**0.113 (0.025)**	**—**	**[0.064, 0.163]**	**0.266 (0.024)**	**—**	**[0.218, 0.314]**
Between classrooms	**0.027 (0.023)**	**—**	**[0.002, 0.086]**	**0.013 (0.009)**	**—**	**[0.002, 0.036]**	**0.034 (0.017)**	**—**	**[0.004, 0.069]**
Between schools	**0.034 (0.115)**	**—**	**[0.002, 0.302]**	**0.013 (0.040)**	**—**	**[0.001, 0.114]**	**0.016 (0.053)**	**—**	**[0.001, 0.146]**

(Continued)

	MPAC[a]			ITBS–MP[b]			ITBS–MC[b]		
	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI
R-Square									
Within classroom	**0.910 (0.055)**	—	**[0.787, 0.994]**	**0.870 (0.031)**	—	**[0.809, 0.928]**	**0.712 (0.029)**	—	**[0.654, 0.767]**
Between classrooms	**0.416 (0.275)**	—	**[0.021, 0.942]**	**0.468 (0.263)**	—	**[0.030, 0.931]**	**0.364 (0.258)**	—	**[0.018, 0.922]**
Between schools	**0.682 (0.274)**	—	**[0.055, 0.983]**	**0.736 (0.251)**	—	**[0.084, 0.977]**	**0.627 (0.276)**	—	**[0.045, 0.972]**
Intraclass correlation									
Between classrooms		.200			.094			.108	
Between schools		.252			.094			.051	

Note. Abbreviations and notes as in Table G.1.

[a]MPAC analysis sample size: Student *N* = 412; Teacher *N* = 158; School *N* = 22. MPAC analysis average cluster size: Teacher *N* = 2.608; School *N* = 18.727.

[b]ITBS analyses sample size: Student *N* = 1356; Teacher *N* = 183; School *N* = 22. ITBS analyses average cluster size: Teacher *N* = 7.410; School *N* = 61.636.

[c]Block indicates the vector of *n*–1 randomization blocks. Effects for Block are omitted for visual simplicity.

Table G.7. Treatment Effect on MPAC, ITBS–MP, and ITBS–MC for Non-FRL Students

	MPAC[a]			ITBS–MP[b]			ITBS–MC[b]		
	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI
Fixed effects									
Within classroom									
Grade 2	**0.750 (0.136)**	**—**	**[0.483, 1.016]**	**1.178 (0.068)**	**—**	**[1.045, 1.312]**	**1.237 (0.075)**	**—**	**[1.090, 1.383]**
Male	0.173 (0.092)	—	[−0.008, 0.351]	0.044 (0.044)	—	[−0.041, 0.129]	**0.119 (0.047)**	**—**	**[0.027, 0.212]**
Minority	−0.031 (0.099)	—	[−0.228, 0.162]	−0.037 (0.048)	—	[−0.131, 0.057]	**0.113 (0.052)**	**—**	**[0.012, 0.215]**
ELL	−0.299 (0.225)	—	[−0.740, 0.147]	**−0.242 (0.086)**	**—**	**[−0.409, −0.073]**	−0.033 (0.092)	—	[−0.214, 0.149]
SWD	0.098 (0.250)	—	[−0.392, 0.588]	−0.165 (0.099)	—	[−0.359, 0.029]	**−0.243 (0.108)**	**—**	**[−0.453, −0.032]**
G1F13 EMSA	**0.838 (0.101)**	**—**	**[0.630, 1.027]**	**0.691 (0.045)**	**—**	**[0.600, 0.777]**	**0.536 (0.053)**	**—**	**[0.424, 0.633]**
G2F13 EMSA	**0.959 (0.084)**	**—**	**[0.790, 1.120]**	**0.688 (0.047)**	**—**	**[0.593, 0.778]**	**0.487 (0.052)**	**—**	**[0.385, 0.587]**
Between classrooms									
G1F13 EMSA	−0.109 (1.792)	—	[−3.942, 3.824]	−0.208 (0.884)	—	[−2.339, 1.122]	0.890 (1.346)	—	[−2.400, 3.698]
G2F13 EMSA	0.870 (1.926)	—	[−3.331, 5.124]	1.227 (1.015)	—	[−0.854, 3.434]	1.079 (1.254)	—	[−1.263, 3.972]
Between schools									
Treatment	0.233 (0.837)	0.23	[−0.835, 1.332]	0.019 (0.496)	0.02	[−0.759, 0.806]	−0.156 (0.513)	−0.16	[−1.025, 0.688]
Block[c]	—	—	—	—	—	—	—	—	—
G1F13 EMSA	0.584 (14.915)	—	[−14.463, 15.539]	0.551 (4.420)	—	[−5.245, 6.479]	0.429 (4.330)	—	[−5.751, 6.542]
G2F13 EMSA	1.113 (11.590)	—	[−15.323, 17.358]	0.566 (7.974)	—	[−12.640, 13.745]	0.421 (8.767)	—	[−14.375, 15.005]
Intercept	−1.051 (8.331)	—	[−9.726, 7.913]	−0.854 (2.866)	—	[−5.392, 3.755]	−0.946 (3.085)	—	[−5.880, 4.180]
Variance components									
Within classroom	**0.111 (0.067)**	**—**	**[0.011, 0.263]**	**0.175 (0.030)**	**—**	**[0.118, 0.234]**	**0.318 (0.031)**	**—**	**[0.258, 0.381]**
Between classrooms	**0.031 (0.030)**	**—**	**[0.003, 0.113]**	**0.017 (0.013)**	**—**	**[0.002, 0.050]**	**0.026 (0.019)**	**—**	**[0.002, 0.073]**
Between schools	**0.121 (18.146)**	**—**	**[0.005, 4.294]**	**0.074 (4.616)**	**—**	**[0.003, 2.474]**	**0.085 (3.413)**	**—**	**[0.003, 2.683]**

(Continued)

	MPAC[a]			ITBS–MP[b]			ITBS–MC[b]		
	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI
R-Square									
Within classroom	**0.881 (0.075)**	—	**[0.710, 0.988]**	**0.806 (0.035)**	—	**[0.736, 0.873]**	**0.658 (0.037)**	—	**[0.582, 0.729]**
Between classrooms	**0.513 (0.281)**	—	**[0.027, 0.959]**	**0.587 (0.266)**	—	**[0.046, 0.958]**	**0.600 (0.261)**	—	**[0.059, 0.964]**
Between schools	**0.556 (0.287)**	—	**[0.031, 0.977]**	**0.512 (0.284)**	—	**[0.027, 0.969]**	**0.502 (0.286)**	—	**[0.025, 0.970]**
Intraclass correlation									
Between classrooms		.118			.064			.061	
Between schools		.460			.278			.198	

Note. Abbreviations and notes as in Table G.1.

[a]MPAC analysis sample size: Student *N* = 238; Teacher *N* = 103; School *N* = 16. MPAC analysis average cluster size: Teacher *N* = 2.311; School *N* = 14.875.

[b]ITBS analyses sample size: Student *N* = 848; Teacher *N* = 130; School *N* = 17. ITBS analyses average cluster size: Teacher *N* = 6.523; School *N* = 49.882.

[c]Block indicates the vector of *n*–1 randomization blocks. Effects for Block are omitted for visual simplicity.

Table G.8. Treatment Effect on MPAC, ITBS–MP, and ITBS–MC for FRL Students

	MPAC[a]			ITBS–MP[b]			ITBS–MC[b]		
	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI
Fixed effects									
Within classroom									
Grade 2	**0.702 (0.102)**	**—**	**[0.491, 0.897]**	**0.996 (0.055)**	**—**	**[0.887, 1.104]**	**1.229 (0.056)**	**—**	**[1.119, 1.338]**
Male	0.116 (0.066)	—	[−0.015, 0.243]	0.007 (0.037)	—	[−0.066, 0.078]	0.039 (0.038)	—	[−0.037, 0.113]
Minority	−0.155 (0.097)	—	[−0.345, 0.038]	**−0.130 (0.049)**	**—**	**[−0.225, −0.036]**	0.010 (0.051)	—	[−0.090, 0.111]
ELL	**−0.170 (0.077)**	**—**	**[−0.319, −0.018]**	**−0.153 (0.043)**	**—**	**[−0.237, −0.069]**	−0.013 (0.045)	—	[−0.101, 0.076]
SWD	**−0.414 (0.122)**	**—**	**[−0.652, −0.169]**	**−0.281 (0.067)**	**—**	**[−0.413, −0.150]**	**−0.316 (0.070)**	**—**	**[−0.454, −0.178]**
G1F13 EMSA	**0.799 (0.063)**	**—**	**[0.674, 0.919]**	**0.753 (0.036)**	**—**	**[0.681, 0.823]**	**0.575 (0.040)**	**—**	**[0.494, 0.650]**
G2F13 EMSA	**0.870 (0.054)**	**—**	**[0.763, 0.975]**	**0.738 (0.034)**	**—**	**[0.670, 0.804]**	**0.524 (0.038)**	**—**	**[0.449, 0.598]**
Between classrooms									
G1F13 EMSA	0.707 (1.670)	—	[−3.384, 4.089]	0.630 (0.670)	—	[−0.999, 1.817]	0.599 (0.865)	—	[−1.583, 2.116]
G2F13 EMSA	0.119 (1.512)	—	[−3.673, 2.791]	0.329 (1.097)	—	[−2.163, 2.658]	0.220 (1.444)	—	[−2.990, 3.355]
Between schools									
Treatment	−0.032 (0.234)	−0.03	[−0.518, 0.412]	0.049 (0.135)	0.05	[−0.217, 0.324]	−0.080 (0.162)	−0.08	[−0.398, 0.247]
Block[c]	—	—	—	—	—	—	—	—	—
G1F13 EMSA	0.547 (1.860)	—	[−2.953, 4.027]	0.481 (0.713)	—	[−0.920, 1.893]	0.616 (1.030)	—	[−1.438, 2.632]
G2F13 EMSA	0.446 (1.930)	—	[−3.119, 4.171]	−0.063 (0.969)	—	[−1.972, 1.923]	−0.154 (1.089)	—	[−2.239, 2.101]
Intercept	0.360 (0.636)	—	[−0.758, 1.615]	0.035 (0.294)	—	[−0.642, 0.541]	−0.446 (0.323)	—	[−1.112, 0.185]
Variance components									
Within classroom	0.029 (0.029)	—	[0.000, 0.105]	**0.133 (0.028)**	**—**	**[0.077, 0.188]**	**0.294 (0.026)**	**—**	**[0.244, 0.345]**
Between classrooms	**0.038 (0.031)**	**—**	**[0.001, 0.112]**	**0.014 (0.010)**	**—**	**[0.002, 0.038]**	**0.023 (0.014)**	**—**	**[0.003, 0.054]**
Between schools	**0.063 (0.222)**	**—**	**[0.003, 0.555]**	**0.025 (0.075)**	**—**	**[0.002, 0.194]**	**0.032 (0.124)**	**—**	**[0.002, 0.280]**

(Continued)

	MPAC[a]			ITBS–MP[b]			ITBS–MC[b]		
	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI
R-Square									
Within classroom	**0.966 (0.035)**	—	**[0.872, 1.000]**	**0.853 (0.033)**	—	**[0.788, 0.917]**	**0.694 (0.029)**	—	**[0.635, 0.750]**
Between classrooms	**0.501 (0.294)**	—	**[0.026, 0.991]**	**0.536 (0.261)**	—	**[0.041, 0.944]**	**0.441 (0.266)**	—	**[0.028, 0.936]**
Between schools	**0.584 (0.282)**	—	**[0.035, 0.975]**	**0.568 (0.277)**	—	**[0.037, 0.969]**	**0.607 (0.279)**	—	**[0.040, 0.976]**
Intraclass correlation									
Between classrooms		.292			.081			.066	
Between schools		.485			.145			.092	

Note. Abbreviations and notes as in Table G.1.

[a]MPAC analysis sample size: Student *N* = 381; Teacher *N* = 143; School *N* = 22. MPAC analysis average cluster size: Teacher *N* = 2.664; School *N* = 17.318.

[b]ITBS analyses sample size: Student *N* = 1311; Teacher *N* = 182; School *N* = 22. ITBS analyses average cluster size: Teacher *N* = 7.203; School *N* = 59.591.

[c]Block indicates the vector of *n*–1 randomization blocks. Effects for Block are omitted for visual simplicity.

Table G.9. Treatment Effect on MPAC, ITBS–MP, and ITBS–MC for Non-ELL Students

	MPAC[a]			ITBS–MP[b]			ITBS–MC[b]		
	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI
Fixed effects									
Within classroom									
Grade 2	**0.698 (0.089)**	**—**	**[0.521, 0.869]**	**1.034 (0.049)**	**—**	**[0.938, 1.128]**	**1.189 (0.056)**	**—**	**[1.076, 1.296]**
Male	0.101 (0.057)	—	[−0.011, 0.211]	−0.005 (0.030)	—	[−0.064, 0.054]	0.057 (0.033)	—	[−0.006, 0.121]
Minority	−0.058 (0.066)	—	[−0.188, 0.072]	−0.066 (0.034)	—	[−0.133, 0.000]	**0.083 (0.037)**	**—**	**[0.010, 0.155]**
FRL	−0.016 (0.079)	—	[−0.164, 0.143]	**−0.168 (0.039)**	**—**	**[−0.245, −0.092]**	**−0.123 (0.043)**	**—**	**[−0.207, −0.041]**
SWD	**−0.297 (0.124)**	**—**	**[−0.542, −0.054]**	**−0.206 (0.060)**	**—**	**[−0.323, −0.090]**	**−0.311 (0.066)**	**—**	**[−0.440, −0.183]**
G1F13 EMSA	**0.787 (0.052)**	**—**	**[0.681, 0.886]**	**0.687 (0.029)**	**—**	**[0.629, 0.744]**	**0.557 (0.033)**	**—**	**[0.490, 0.619]**
G2F13 EMSA	**0.851 (0.046)**	**—**	**[0.762, 0.942]**	**0.694 (0.030)**	**—**	**[0.634, 0.753]**	**0.499 (0.034)**	**—**	**[0.430, 0.566]**
Between classrooms									
G1F13 EMSA	0.119 (1.109)	—	[−2.322, 2.351]	0.232 (0.532)	—	[−1.063, 0.927]	**0.988 (0.399)**	**—**	**[0.335, 1.876]**
G2F13 EMSA	0.970 (1.142)	—	[−1.165, 3.760]	0.963 (0.692)	—	[−0.169, 2.586]	1.026 (0.831)	—	[−0.323, 3.049]
Between schools									
Treatment	0.072 (0.194)	0.07	[−0.312, 0.459]	0.075 (0.111)	0.08	[−0.139, 0.300]	−0.105 (0.113)	−0.10	[−0.325, 0.124]
Block[c]	—	—	—	—	—	—	—	—	—
G1F13 EMSA	0.429 (0.804)	—	[−1.158, 2.039]	0.363 (0.440)	—	[−0.498, 1.269]	0.472 (0.488)	—	[−0.512, 1.454]
G2F13 EMSA	0.531 (0.887)	—	[−1.241, 2.243]	0.222 (0.818)	—	[−1.360, 1.888]	0.091 (0.742)	—	[−1.367, 1.618]
Intercept	−0.230 (0.338)	—	[−0.915, 0.427]	−0.327 (0.213)	—	[−0.749, 0.095]	**−0.611 (0.229)**	**—**	**[−1.063, −0.152]**
Variance components									
Within classroom	**0.038 (0.030)**	**—**	**[0.004, 0.114]**	**0.135 (0.021)**	**—**	**[0.095, 0.176]**	**0.284 (0.021)**	**—**	**[0.244, 0.327]**
Between classrooms	**0.014 (0.015)**	**—**	**[0.001, 0.054]**	**0.013 (0.008)**	**—**	**[0.002, 0.033]**	**0.016 (0.013)**	**—**	**[0.002, 0.051]**
Between schools	**0.039 (0.134)**	**—**	**[0.002, 0.340]**	**0.018 (0.048)**	**—**	**[0.002, 0.136]**	**0.015 (0.046)**	**—**	**[0.001, 0.131]**

(Continued)

	MPAC[a]			ITBS–MP[b]			ITBS–MC[b]		
	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI
R-Square									
Within classroom	**0.954 (0.038)**	—	**[0.855, 0.995]**	**0.836 (0.027)**	—	**[0.783, 0.887]**	**0.683 (0.027)**	—	**[0.630, 0.734]**
Between classrooms	**0.645 (0.273)**	—	**[0.045, 0.969]**	**0.567 (0.248)**	—	**[0.064, 0.948]**	**0.749 (0.205)**	—	**[0.226, 0.975]**
Between schools	**0.731 (0.265)**	—	**[0.071, 0.987]**	**0.713 (0.257)**	—	**[0.077, 0.982]**	**0.760 (0.252)**	—	**[0.092, 0.984]**
Intraclass correlation									
Between classrooms		.154			.078			.051	
Between schools		.429			.108			.048	

Note. Abbreviations and notes as in Table G.1.

[a]MPAC analysis sample size: Student *N* = 479; Teacher *N* = 162; School *N* = 22. MPAC analysis average cluster size: Teacher *N* = 2.957; School *N* = 21.773.

[b]ITBS analyses sample size: Student *N* = 1667; Teacher *N* = 181; School *N* = 22. ITBS analyses average cluster size: Teacher *N* = 9.210; School *N* = 75.773.

[c]Block indicates the vector of *n*–1 randomization blocks. Effects for Block are omitted for visual simplicity.

Table G.10. Treatment Effect on MPAC, ITBS–MP, and ITBS–MC for ELL Students

	MPAC[a]			ITBS–MP[b]			ITBS–MC[b]		
	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI
Fixed effects									
Within classroom									
Grade 2	**0.847 (0.179)**	**—**	**[0.498, 1.199]**	**1.037 (0.095)**	**—**	**[0.854, 1.226]**	**1.283 (0.091)**	**—**	**[1.105, 1.463]**
Male	0.123 (0.118)	—	[−0.107, 0.355]	0.111 (0.064)	—	[−0.013, 0.237]	0.090 (0.064)	—	[−0.035, 0.216]
Minority	−0.135 (0.432)	—	[−0.968, 0.731]	−0.052 (0.142)	—	[−0.331, 0.225]	−0.071 (0.142)	—	[−0.351, 0.204]
FRL	−0.037 (0.270)	—	[−0.568, 0.492]	−0.126 (0.104)	—	[−0.330, 0.077]	−0.073 (0.104)	—	[−0.278, 0.130]
SWD	−0.266 (0.235)	—	[−0.728, 0.194]	**−0.367 (0.117)**	**—**	**[−0.598, −0.138]**	−0.182 (0.117)	—	[−0.411, 0.049]
G1F13 EMSA	**0.601 (0.206)**	**—**	**[0.156, 0.965]**	**0.692 (0.074)**	**—**	**[0.539, 0.831]**	**0.466 (0.084)**	**—**	**[0.288, 0.618]**
G2F13 EMSA	**0.870 (0.107)**	**—**	**[0.651, 1.073]**	**0.683 (0.057)**	**—**	**[0.569, 0.794]**	**0.495 (0.060)**	**—**	**[0.375, 0.611]**
Between classrooms									
G1F13 EMSA	0.634 (2.290)	—	[−4.860, 5.047]	0.895 (1.444)	—	[−2.205, 4.214]	0.725 (1.470)	—	[−2.076, 4.127]
G2F13 EMSA	−0.128 (2.840)	—	[−5.938, 6.187]	0.694 (1.563)	—	[−2.709, 4.161]	0.493 (1.608)	—	[−3.424, 3.649]
Between schools									
Treatment	0.119 (0.689)	0.12	[−1.046, 1.392]	−0.074 (0.209)	−0.07	[−0.480, 0.355]	−0.005 (0.223)	−0.00	[−0.393, 0.481]
Block[c]	—	—	—	—	—	—	—	—	—
G1F13 EMSA	1.132 (5.662)	—	[−7.738, 10.672]	0.713 (1.316)	—	[−1.589, 3.102]	0.033 (1.728)	—	[−3.313, 3.119]
G2F13 EMSA	0.773 (6.850)	—	[−11.583, 12.592]	0.218 (2.448)	—	[−4.650, 5.117]	−0.119 (2.193)	—	[−4.599, 4.178]
Intercept	0.344 (3.232)	—	[−5.359, 6.089]	0.167 (0.978)	—	[−1.766, 2.119]	−0.445 (0.893)	—	[−2.305, 1.223]
Variance components									
Within classroom	**0.131 (0.077)**	**—**	**[0.012, 0.310]**	**0.186 (0.046)**	**—**	**[0.096, 0.276]**	**0.302 (0.040)**	**—**	**[0.225, 0.381]**
Between classrooms	**0.125 (0.097)**	**—**	**[0.008, 0.364]**	**0.036 (0.028)**	**—**	**[0.003, 0.105]**	**0.046 (0.031)**	**—**	**[0.004, 0.118]**
Between schools	**0.224 (8.670)**	**—**	**[0.007, 4.660]**	**0.035 (0.142)**	**—**	**[0.002, 0.395]**	**0.044 (0.157)**	**—**	**[0.002, 0.444]**

(Continued)

	MPAC[a]			ITBS–MP[b]			ITBS–MC[b]		
	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI
R-Square									
Within classroom	**0.834 (0.097)**	—	**[0.615, 0.986]**	**0.791 (0.055)**	—	**[0.683, 0.897]**	**0.676 (0.047)**	—	**[0.583, 0.765]**
Between classrooms	**0.454 (0.284)**	—	**[0.023, 0.962]**	**0.549 (0.277)**	—	**[0.036, 0.967]**	**0.469 (0.276)**	—	**[0.027, 0.956]**
Between schools	**0.583 (0.287)**	—	**[0.034, 0.981]**	**0.649 (0.280)**	—	**[0.044, 0.979]**	**0.509 (0.285)**	—	**[0.026, 0.965]**
Intraclass correlation									
Between classrooms		.260			.140			.117	
Between schools		.467			.136			.112	

Note. Abbreviations and notes as in Table G.1.

[a]MPAC analysis sample size: Student *N* = 140; Teacher *N* = 89; School *N* = 19. MPAC analysis average cluster size: Teacher *N* = 1.573; School *N* = 7.368.

[b]ITBS analyses sample size: Student *N* = 492; Teacher *N* = 148; School *N* = 22. ITBS analyses average cluster size: Teacher *N* = 3.324; School *N* = 22.364.

[c]Block indicates the vector of *n*–1 randomization blocks. Effects for Block are omitted for visual simplicity.

Table G.11. Treatment Effect on MPAC, ITBS–MP, and ITBS–MC for Non-SWD Students

	MPAC[a]			ITBS–MP[b]			ITBS–MC[b]		
	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI
Fixed effects									
Within classroom									
Grade 2	**0.715 (0.081)**	**—**	**[0.555, 0.870]**	**1.029 (0.045)**	**—**	**[0.941, 1.118]**	**1.228 (0.052)**	**—**	**[1.126, 1.328]**
Male	**0.118 (0.052)**	**—**	**[0.017, 0.220]**	0.017 (0.027)	—	[−0.037, 0.070]	**0.069 (0.029)**	**—**	**[0.011, 0.126]**
Minority	−0.054 (0.066)	—	[−0.186, 0.073]	−0.064 (0.033)	—	[−0.130, 0.001]	0.067 (0.036)	—	[−0.004, 0.137]
FRL	−0.003 (0.073)	—	[−0.145, 0.142]	**−0.161 (0.037)**	**—**	**[−0.234, −0.088]**	**−0.114 (0.040)**	**—**	**[−0.194, −0.035]**
ELL	**−0.210 (0.074)**	**—**	**[−0.357, −0.067]**	**−0.170 (0.038)**	**—**	**[−0.243, −0.096]**	−0.042 (0.041)	—	[−0.122, 0.038]
G1F13 EMSA	**0.795 (0.054)**	**—**	**[0.683, 0.897]**	**0.668 (0.028)**	**—**	**[0.613, 0.721]**	**0.543 (0.031)**	**—**	**[0.481, 0.602]**
G2F13 EMSA	**0.857 (0.045)**	**—**	**[0.767, 0.945]**	**0.685 (0.028)**	**—**	**[0.630, 0.739]**	**0.491 (0.031)**	**—**	**[0.429, 0.553]**
Between classrooms									
G1F13 EMSA	−0.198 (1.392)	—	[−3.496, 2.495]	0.219 (0.577)	—	[−1.273, 1.016]	0.864 (0.719)	—	[−0.542, 2.227]
G2F13 EMSA	0.864 (1.408)	—	[−2.542, 3.730]	1.222 (0.926)	—	[−0.164, 3.688]	0.722 (1.463)	—	[−2.116, 4.417]
Between schools									
Treatment	0.105 (0.165)	0.11	[−0.230, 0.428]	0.028 (0.102)	0.03	[−0.171, 0.232]	−0.114 (0.104)	−0.11	[−0.320, 0.094]
Block[c]	—	—	—	—	—	—	—	—	—
G1F13 EMSA	0.151 (0.649)	—	[−1.133, 1.471]	0.327 (0.426)	—	[−0.493, 1.213]	0.389 (0.451)	—	[−0.511, 1.301]
G2F13 EMSA	0.534 (0.689)	—	[−0.847, 1.898]	0.458 (0.749)	—	[−1.027, 1.964]	0.219 (0.667)	—	[−1.078, 1.584]
Intercept	−0.130 (0.258)	—	[−0.644, 0.385]	−0.194 (0.178)	—	[−0.550, 0.156]	**−0.548 (0.202)**	**—**	**[−0.949, −0.149]**
Variance components									
Within classroom	**0.059 (0.035)**	**—**	**[0.006, 0.137]**	**0.149 (0.019)**	**—**	**[0.112, 0.186]**	**0.291 (0.019)**	**—**	**[0.255, 0.328]**
Between classrooms	**0.024 (0.018)**	**—**	**[0.002, 0.068]**	**0.017 (0.009)**	**—**	**[0.002, 0.037]**	**0.036 (0.016)**	**—**	**[0.006, 0.068]**
Between schools	**0.027 (0.080)**	**—**	**[0.002, 0.235]**	**0.015 (0.041)**	**—**	**[0.001, 0.113]**	**0.014 (0.047)**	**—**	**[0.001, 0.125]**

(Continued)

	MPAC[a]			ITBS–MP[b]			ITBS–MC[b]		
	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI
R-Square									
Within classroom	**0.927 (0.044)**	—	**[0.825, 0.994]**	**0.816 (0.025)**	—	**[0.767, 0.864]**	**0.679 (0.024)**	—	**[0.631, 0.724]**
Between classrooms	**0.522 (0.271)**	—	**[0.034, 0.952]**	**0.538 (0.248)**	—	**[0.053, 0.936]**	**0.427 (0.247)**	—	**[0.034, 0.914]**
Between schools	**0.703 (0.273)**	—	**[0.058, 0.983]**	**0.782 (0.239)**	—	**[0.113, 0.986]**	**0.757 (0.251)**	—	**[0.093, 0.983]**
Intraclass correlation									
Between classrooms		.218			.094			.106	
Between schools		.245			.083			.041	

Note. Abbreviations and notes as in Table G.1.

[a]MPAC analysis sample size: Student *N* = 579; Teacher *N* = 167; School *N* = 22. MPAC analysis average cluster size: Teacher *N* = 3.467; School *N* = 26.318.

[b]ITBS analyses sample size: Student *N* = 2001; Teacher *N* = 183; School *N* = 22. ITBS analyses average cluster size: Teacher *N* = 10.934; School *N* = 90.955.

[c]Block indicates the vector of *n*–1 randomization blocks. Effects for Block are omitted for visual simplicity.

Table G.12. Treatment Effect on MPAC, ITBS–MP, and ITBS–MC for SWD Students

	MPAC[a]			ITBS–MP[b]			ITBS–MC[b]		
	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI
Fixed effects									
Within classroom									
Grade 2	0.751 (0.519)	—	[−0.275, 1.762]	**0.852 (0.157)**	**—**	**[0.551, 1.164]**	**1.109 (0.173)**	**—**	**[0.773, 1.449]**
Male	−0.147 (0.517)	—	[−1.181, 0.879]	−0.006 (0.130)	—	[−0.266, 0.244]	−0.102 (0.152)	—	[−0.405, 0.191]
Minority	−0.249 (0.429)	—	[−1.115, 0.606]	**−0.329 (0.142)**	**—**	**[−0.609, −0.056]**	0.005 (0.159)	—	[−0.309, 0.310]
FRL	−0.722 (0.580)	—	[−1.884, 0.450]	**−0.295 (0.143)**	**—**	**[−0.583, −0.016]**	−0.226 (0.165)	—	[−0.553, 0.095]
ELL	−0.134 (0.545)	—	[−1.209, 0.935]	−0.074 (0.144)	—	[−0.358, 0.204]	0.211 (0.162)	—	[−0.106, 0.528]
G1F13 EMSA	**0.724 (0.312)**	**—**	**[0.126, 1.368]**	**0.781 (0.112)**	**—**	**[0.573, 1.012]**	**0.668 (0.118)**	**—**	**[0.410, 0.879]**
G2F13 EMSA	0.328 (0.726)	—	[−1.130, 1.788]	**0.674 (0.092)**	**—**	**[0.495, 0.852]**	**0.690 (0.113)**	**—**	**[0.468, 0.912]**
Between classrooms									
G1F13 EMSA	0.554 (1.906)	—	[−3.148, 3.887]	0.216 (0.839)	—	[−1.462, 1.796]	0.226 (1.329)	—	[−2.725, 2.736]
G2F13 EMSA	0.361 (2.034)	—	[−3.177, 4.257]	0.452 (1.078)	—	[−1.888, 2.524]	0.490 (1.421)	—	[−2.584, 3.594]
Between schools									
Treatment	0.372 (2.986)	0.36	[−3.085, 3.889]	−0.149 (0.279)	−0.15	[−0.703, 0.401]	−0.267 (0.332)	−0.27	[−0.948, 0.369]
Block[c]	—	—	—	—	—	—	—	—	—
G1F13 EMSA	0.138 (16.999)	—	[−22.421, 22.165]	0.021 (2.607)	—	[−5.005, 4.986]	−0.003 (3.212)	—	[−6.403, 6.366]
G2F13 EMSA	0.207 (8.245)	—	[−10.050, 10.201]	0.583 (1.453)	—	[−1.864, 3.287]	0.298 (2.072)	—	[−3.441, 4.108]
Intercept	1.305 (8.784)	—	[−9.603, 11.509]	0.395 (1.299)	—	[−2.118, 2.941]	−0.322 (1.696)	—	[−3.613, 2.965]
Variance components									
Within classroom	**0.109 (0.187)**	**—**	**[0.004, 0.655]**	**0.046 (0.048)**	**—**	**[0.003, 0.178]**	**0.130 (0.091)**	**—**	**[0.012, 0.348]**
Between classrooms	**0.132 (0.226)**	**—**	**[0.006, 0.793]**	**0.026 (0.030)**	**—**	**[0.002, 0.113]**	**0.048 (0.051)**	**—**	**[0.004, 0.190]**
Between schools	**1.128 (869.040)**	**—**	**[0.029, 40.714]**	**0.064 (0.253)**	**—**	**0.003, 0.684]**	**0.093 (0.357)**	**—**	**[0.003, 0.989]**

(Continued)

	MPAC[a]			ITBS–MP[b]			ITBS–MC[b]		
	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI
R-Square									
Within classroom	**0.898 (0.121)**	—	**[0.551, 0.996]**	**0.946 (0.058)**	—	**[0.784, 0.996]**	**0.875 (0.090)**	—	**[0.655, 0.988]**
Between classrooms	**0.605 (0.284)**	—	**[0.039, 0.980]**	**0.598 (0.274)**	—	**[0.041, 0.966]**	**0.546 (0.277)**	—	**[0.034, 0.964]**
Between schools	**0.532 (0.288)**	—	**[0.029, 0.978]**	**0.600 (0.285)**	—	**[0.037, 0.978]**	**0.537 (0.288)**	—	**[0.029, 0.975]**
Intraclass correlation									
Between classrooms		.096			.191			.117	
Between schools		.824			.471			.343	

Note. Abbreviations and notes as in Table G.1.

[a]MPAC analysis sample size: Student *N* = 40; Teacher *N* = 37; School *N* = 18. MPAC analysis average cluster size: Teacher *N* = 1.081; School *N* = 2.222.

[b]ITBS analyses sample size: Student *N* = 158; Teacher *N* = 103; School *N* = 22. ITBS analyses average cluster size: Teacher *N* = 1.534; School *N* = 7.182.

[c]Block indicates the vector of *n*–1 randomization blocks. Effects for Block are omitted for visual simplicity.

Appendix H. Model Results for Moderation Analyses

Table H.1. Treatment-by-Grade Moderation Effects on MPAC, ITBS–MP, and ITBS–MC

	MPAC[a]			ITBS–MP[b]			ITBS–MC[b]		
	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI
Within classroom									
Male	**0.122 (0.048)**	**—**	**[0.027, 0.216]**	0.019 (0.026)	—	[−0.032, 0.071]	**0.061 (0.029)**	**—**	**[0.005, 0.117]**
Minority	−0.069 (0.062)	—	[−0.190, 0.054]	**−0.073 (0.032)**	**—**	**[−0.135, −0.010]**	0.068 (0.035)	—	[−0.001, 0.137]
FRL	−0.048 (0.070)	—	[−0.185, 0.091]	**−0.161 (0.035)**	**—**	**[−0.231, −0.092]**	**−0.118 (0.039)**	**—**	**[−0.195, −0.042]**
ELL	**−0.165 (0.070)**	**—**	**[−0.305, −0.032]**	**−0.160 (0.035)**	**—**	**[−0.230, −0.091]**	−0.020 (0.039)	—	[−0.097, 0.055]
SWD	**−0.298 (0.105)**	**—**	**[−0.503, −0.093]**	**−0.237 (0.052)**	**—**	**[−0.337, −0.135]**	**−0.291 (0.057)**	**—**	**[−0.402, −0.180]**
G1F13 EMSA	**0.780 (0.047)**	**—**	**[0.684, 0.867]**	**0.670 (0.026)**	**—**	**[0.617, 0.719]**	**0.555 (0.029)**	**—**	**[0.495, 0.609]**
G2F13 EMSA	**0.845 (0.040)**	**—**	**[0.766, 0.925]**	**0.683 (0.025)**	**—**	**[0.633, 0.732]**	**0.501 (0.029)**	**—**	**[0.445, 0.558]**
Between classrooms									
G1F13 EMSA	−0.284 (1.208)	—	[−3.096, 2.129]	0.233 (0.339)	—	[−0.556, 0.755]	0.306 (0.391)	—	[−0.532, 0.944]
G2F13 EMSA	0.727 (1.251)	—	[−2.262, 3.267]	0.738 (0.727)	—	[−0.846, 2.331]	0.491 (0.763)	—	[−1.215, 2.101]
Between schools									
Grade 2[c]	**0.835 (0.119)**	**—**	**[0.603, 1.073]**	**1.084 (0.086)**	**—**	**[0.920, 1.258]**	**1.268 (0.106)**	**—**	**[1.056, 1.476]**
Treatment	0.230 (0.189)	0.23	[−0.154, 0.591]	0.111 (0.120)	0.11	[−0.119, 0.355]	−0.001 (0.118)	−0.00	[−0.232, 0.234]
Treatment by Grade 2	−0.263 (0.167)	−0.27	[−0.593, 0.065]	−0.136 (0.123)	−0.14	[−0.385, 0.104]	−0.180 (0.154)	−0.18	[−0.494, 0.118]
Block[d]	—	—	—	—	—	—	—	—	—
G1F13 EMSA	0.355 (0.718)	—	[−1.036, 1.836]	0.453 (0.460)	—	[−0.436, 1.393]	0.451 (0.475)	—	[−0.472, 1.428]
G2F13 EMSA	0.375 (0.770)	—	[−1.161, 1.890]	−0.038 (0.835)	—	[−1.731, 1.601]	0.216 (0.754)	—	[−1.316, 1.698]
Intercept	−0.172 (0.293)	—	[−0.755, 0.889]	−0.263 (0.192)	—	[−0.649, 0.117]	**−0.654 (0.192)**	**—**	**[−1.039, −0.276]**
Variance components									
Within classroom	**0.034 (0.027)**	**—**	**[0.002, 0.102]**	**0.132 (0.019)**	**—**	**[0.095, 0.168]**	**0.279 (0.019)**	**—**	**[0.244, 0.316]**
Between classrooms	**0.020 (0.017)**	**—**	**[0.002, 0.062]**	**0.010 (0.007)**	**—**	**[0.001, 0.026]**	**0.008 (0.007)**	**—**	**[0.001, 0.026]**
Between schools	**0.032 (0.128)**	**—**	**[0.002, 0.274]**	**0.018 (0.048)**	**—**	**[0.002, 0.134]**	**0.016 (0.044)**	**—**	**[0.001, 0.133]**

(Continued)

	MPAC[a]			ITBS–MP[b]			ITBS–MC[b]		
	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI
Intraclass correlation									
Between classrooms		.233			.063			.026	
Between schools		.372			.113			.053	

Note. FRL = Free/Reduced−price Lunch; ELL = English Language Learner; SWD = Student with Disability. G1F13 EMSA = Grade 1, fall 2013; G2F13 EMSA = Grade 2, fall 2013; PSD = the standard deviation of the posterior distribution. 95% CI = 95% credibility intervals of the posterior distribution with equal tail percentages. Reported estimates are from the unstandardized solution. Only the effect size for Treatment is presented; it is calculated as Hedges' *g*. Boldface indicates the 95% CI does not include zero.

[a]MPAC analysis sample size: Student *N* = 622; Teacher *N* = 167; School *N* = 22. MPAC analysis average cluster size: Teacher *N* = 3.725; School *N* = 28.273.

[b]ITBS analyses sample size: Student *N* = 2,172; Teacher *N* = 183; School *N* = 22. ITBS analyses average cluster size: Teacher *N* = 11.869; School *N* = 98.727.

[c]Grade 2 slope is specified to vary randomly across clusters; therefore, the value reported is the intercept for the Grade 2 between-school slope (i.e., the Grade 2 slope, holding all school-level covariates constant at zero).

[d]Block indicates the vector of *n*–1 randomization blocks. Effects for Block are omitted for visual simplicity.

Table H.2. Treatment-by-Male Moderation Effects on MPAC, ITBS–MP, and ITBS–MC

	MPAC[a]			ITBS–MP[b]			ITBS–MC[b]		
	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI
Within classroom									
Grade 2	**0.701 (0.086)**	**—**	**[0.495, 0.854]**	**0.994 (0.044)**	**—**	**[0.908, 1.080]**	0.788 (0.422)	—	[−0.043, 1.619]
Minority	−0.070 (0.062)	—	[−0.192, 0.051]	**−0.073 (0.032)**	**—**	**[−0.137, −0.010]**	0.065 (0.036)	—	[−0.005, 0.134]
FRL	−0.040 (0.071)	—	[−0.179, 0.098]	**−0.164 (0.036)**	**—**	**[−0.234, −0.094]**	**−0.108 (0.040)**	**—**	**[−0.186, −0.030]**
ELL	**−0.175 (0.070)**	**—**	**[−0.314, −0.037]**	**−0.161 (0.035)**	**—**	**[−0.231, −0.092]**	−0.034 (0.040)	—	[−0.112, 0.044]
SWD	**−0.299 (0.105)**	**—**	**[−0.500, −0.088]**	**−0.232 (0.052)**	**—**	**[−0.334, −0.130]**	**−0.291 (0.058)**	**—**	**[−0.404, −0.177]**
G1F13 EMSA	**0.785 (0.046)**	**—**	**[0.688, 0.871]**	**0.671 (0.026)**	**—**	**[0.619, 0.722]**	**0.558 (0.029)**	**—**	**[0.498, 0.613]**
G2F13 EMSA	**0.846 (0.040)**	**—**	**[0.766, 0.922]**	**0.684 (0.025)**	**—**	**[0.634, 0.733]**	**0.505 (0.029)**	**—**	**[0.448, 0.562]**
Between classrooms									
G1F13 EMSA	−0.228 (1.226)	—	[−3.184, 1.940]	0.360 (0.396)	—	[−0.554, 0.977]	**1.975 (0.623)**	**—**	**[0.726, 3.318]**
G2F13 EMSA	0.749 (1.586)	—	[−3.222, 3.934]	0.930 (0.878)	—	[−0.729, 3.115]	2.350 (1.483)	—	[−0.302, 5.619]
Between schools									
Male[ac]	0.132 (0.082)	—	[−0.020, 0.303]	0.047 (0.052)	—	[−0.055, 0.148]	0.086 (0.058)	—	[−0.026, 0.203]
Treatment	0.089 (0.180)	0.09	[−0.274, 0.446]	0.059 (0.116)	0.06	[−0.170, 0.293]	−0.095 (0.188)	−0.09	[−0.457, 0.283]
Treatment by Male	−0.017 (0.116)	−0.02	[−0.259, 0.199]	−0.052 (0.073)	−0.05	[−0.195, 0.092]	−0.046 (0.083)	−0.05	[−0.213, 0.113]
Block[c]	—	—	—	—	—	—	—	—	—
G1F13 EMSA	0.299 (0.704)	—	[−1.068, 1.732]	0.447 (0.411)	—	[−0.342, 1.302]	0.181 (0.785)	—	[−1.380, 1.688]
G2F13 EMSA	0.478 (0.733)	—	[−0.976, 1.937]	0.224 (0.850)	—	[−1.445, 1.948]	0.350 (1.176)	—	[−1.950, 2.715]
Intercept	−0.083 (0.284)	—	[−0.654, 0.472]	−0.193 (0.177)	—	[−0.551, 0.152]	0.117 (0.401)	—	[−0.666, 0.918]
Variance components									
Within classroom	0.028 (0.028)	—	[0.000, 0.101]	**0.127 (0.018)**	—	**[0.093, 0.163]**	**0.274 (0.019)**	—	**[0.237, 0.311]**
Between classrooms	**0.021 (0.018)**	—	**[0.001, 0.065]**	**0.015 (0.008)**	—	**[0.002, 0.032]**	**0.057 (0.066)**	—	**[0.004, 0.249]**
Between schools	**0.030 (0.110)**	—	**[0.002, 0.270]**	**0.013 (0.038)**	—	**[0.001, 0.107]**	**0.029 (0.114)**	—	**[0.002, 0.312]**

(Continued)

	MPAC[a]			ITBS–MP[b]			ITBS–MC[b]		
	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI
Intraclass correlation									
Between classrooms		.266			.097			.158	
Between schools		.380			.084			.081	

Note. Abbreviations and notes as in Table H.1.

[a]MPAC analysis sample size: Student *N* = 622; Teacher *N* = 167; School *N* = 22. MPAC analysis average cluster size: Teacher *N* = 3.725; School *N* = 28.273.

[b]ITBS analyses sample size: Student *N* = 2,169; Teacher *N* = 183; School *N* = 22. ITBS analyses average cluster size: Teacher *N* = 11.852; School *N* = 98.591.

[c]Male slope is specified to vary randomly across clusters; therefore, the value reported is the intercept for the Male between-school slope (i.e., the Male slope, holding all school-level covariates constant at zero).

[d]Block indicates the vector of *n*–1 randomization blocks. Effects for Block are omitted for visual simplicity.

Table H.3. Treatment-by-Minority Moderation Effects on MPAC, ITBS–MP, and ITBS–MC

	MPAC[a]			ITBS–MP[b]			ITBS–MC[b]		
	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI
Within classroom									
Grade 2	**0.703 (0.081)**	**—**	**[0.539, 0.858]**	**1.010 (0.044)**	**—**	**[0.924, 1.095]**	**1.208 (0.051)**	**—**	**[1.107, 1.306]**
Male	**0.127 (0.048)**	**—**	**[0.029, 0.220]**	0.019 (0.026)	—	[−0.033, 0.071]	**0.060 (0.029)**	**—**	**[0.004, 0.117]**
FRL	−0.050 (0.070)	—	[−0.183, 0.087]	−0.167 (0.036)	—	[−0.237, −0.096]	**−0.118 (0.039)**	**—**	**[−0.195, −0.041]**
ELL	**−0.170 (0.070)**	**—**	**[−0.306, −0.031]**	**−0.163 (0.036)**	**—**	**[−0.233, −0.092]**	−0.028 (0.039)	—	[−0.105, 0.048]
SWD	**−0.295 (0.104)**	**—**	**[−0.503, −0.094]**	**−0.235 (0.052)**	**—**	**[−0.336, −0.135]**	**−0.288 (0.057)**	**—**	**[−0.400, −0.178]**
G1F13 EMSA	**0.778 (0.048)**	**—**	**[0.682, 0.871]**	**0.666 (0.026)**	**—**	**[0.615, 0.717]**	**0.554 (0.029)**	**—**	**[0.495, 0.609]**
G2F13 EMSA	**0.835 (0.042)**	**—**	**[0.753, 0.919]**	**0.678 (0.025)**	**—**	**[0.627, 0.727]**	**0.497 (0.029)**	**—**	**[0.439, 0.554]**
Between classrooms									
G1F13 EMSA	−0.083 (1.244)	—	[−3.422, 2.198]	0.362 (0.377)	—	[−0.501, 0.918]	**0.830 (0.357)**	**—**	**[0.104, 1.554]**
G2F13 EMSA	0.718 (1.125)	—	[−1.863, 3.061]	0.830 (0.860)	—	[−0.914, 2.865]	0.739 (1.332)	—	[−2.022, 3.916]
Between schools									
Minority[c]	−0.111 (0.106)	—	[−0.319, 0.099]	−0.089 (0.060)	—	[−0.204, 0.031]	0.102 (0.059)	—	[−0.013, 0.219]
Treatment	0.055 (0.186)	0.06	[−0.313, 0.422]	0.030 (0.120)	0.03	[−0.201, 0.272]	−0.064 (0.126)	−0.06	[−0.311, 0.184]
Treatment by minority	0.055 (0.150)	0.06	[−0.248, 0.340]	0.013 (0.082)	0.01	[−0.157, 0.168]	−0.059 (0.080)	−0.06	[−0.215, 0.100]
Block[d]	—	—	—	—	—	—	—	—	—
G1F13 EMSA	0.281 (0.714)	—	[−1.159, 1.688]	0.370 (0.427)	—	[−0.452, 1.242]	0.431 (0.496)	—	[−0.539, 1.445]
G2F13 EMSA	0.533 (0.696)	—	[−0.863, 1.920]	0.192 (0.825)	—	[−1.417, 1.877]	0.244 (0.770)	—	[−1.272, 1.791]
Intercept	−0.071 (0.297)	—	[−0.654, 0.522]	−0.167 (0.184)	—	[−0.538, 0.187]	−0.602 (0.200)	—	[−1.002, 0.214]
Variance components									
Within classroom	**0.035 (0.027)**	**—**	**[0.002, 0.101]**	**0.134 (0.018)**	**—**	**[0.098, 0.170]**	**0.280 (0.019)**	**—**	**[0.243, 0.317]**
Between classrooms	**0.015 (0.014)**	**—**	**[0.001, 0.052]**	**0.015 (0.008)**	**—**	**[0.002, 0.032]**	**0.026 (0.014)**	**—**	**[0.003, 0.057]**
Between schools	**0.028 (0.095)**	**—**	**[0.002, 0.264]**	**0.014 (0.047)**	**—**	**[0.001, 0.119]**	**0.014 (0.057)**	**—**	**[0.001, 0.033]**

(Continued)

	MPAC[a]			ITBS–MP[b]			ITBS–MC[b]		
	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI
Intraclass correlation									
Between classrooms		.192			.092			.081	
Between schools		.359			.086			.044	

Note. Abbreviations and notes as in Table H.1.

[a]MPAC analysis sample size: Student *N* = 619; Teacher *N* = 167; School *N* = 22. MPAC analysis average cluster size: Teacher *N* = 3.707; School *N* = 28.136.

[b]ITBS analyses sample size: Student *N* = 2,159; Teacher *N* = 183; School *N* = 22. ITBS analyses average cluster size: Teacher *N* = 11.798; School *N* = 98.136.

[c]Minority slope is specified to vary randomly across clusters; therefore, the value reported is the intercept for the Minority between-school slope (i.e., the Minority slope, holding all school-level covariates constant at zero).

[d]Block indicates the vector of *n*–1 randomization blocks. Effects for Block are omitted for visual simplicity.

Table H.4. Treatment-by-FRL Moderation Effects on MPAC, ITBS–MP, and ITBS–MC

	MPAC[a]			ITBS–MP[b]			ITBS–MC[b]		
	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI
Within classroom									
Grade 2	**0.704 (0.081)**	**—**	**[0.544, 0.858]**	**1.015 (0.044)**	**—**	**[0.928, 1.101]**	**1.210 (0.051)**	**—**	**[1.110, 1.308]**
Male	**0.128 (0.048)**	**—**	**[0.032, 0.221]**	0.021 (0.026)	—	[−0.031, 0.073]	**0.066 (0.029)**	**—**	**[0.009, 0.122]**
Minority	−0.077 (0.061)	—	[−0.194, 0.044]	**−0.073 (0.032)**	**—**	**[−0.135, −0.010]**	**0.069 (0.035)**	**—**	**[0.000, 0.138]**
ELL	**−0.181 (0.070)**	**—**	**[−0.316, −0.042]**	**−0.164 (0.036)**	**—**	**[−0.234, −0.094]**	−0.029 (0.039)	—	[−0.106, 0.048]
SWD	**−0.279 (0.104)**	**—**	**[−0.487, −0.078]**	**−0.234 (0.052)**	**—**	**[−0.336, −0.134]**	**−0.290 (0.057)**	**—**	**[−0.402, −0.179]**
G1F13 EMSA	**0.768 (0.048)**	**—**	**[0.674, 0.861]**	**0.670 (0.026)**	**—**	**[0.618, 0.720]**	**0.552 (0.029)**	**—**	**[0.493, 0.607]**
G2F13 EMSA	**0.838 (0.041)**	**—**	**[0.757, 0.918]**	**0.679 (0.025)**	**—**	**[0.628, 0.728]**	**0.495 (0.029)**	**—**	**[0.437, 0.552]**
Between classrooms									
G1F13 EMSA	−0.075 (1.210)	—	[−3.233, 2.147]	0.344 (0.380)	—	[−0.522, 0.890]	**0.802 (0.375)**	**—**	**[0.029, 1.543]**
G2F13 EMSA	0.771 (1.091)	—	[−1.786, 2.989]	0.872 (0.827)	—	[−0.776, 2.832]	0.756 (1.351)	—	[−2.002, 3.987]
Between schools									
FRL[c]	0.059 (0.134)	—	[−0.203, 0.324]	−0.141 (0.077)	—	[−0.290, 0.015]	−0.129 (0.076)	—	[−0.280, 0.019]
Treatment	0.178 (0.200)	0.18	[−0.208, 0.586]	0.032 (0.133)	0.03	[−0.224, 0.303]	−0.132 (0.138)	−0.13	[−0.406, 0.140]
Treatment by FRL	−0.207 (0.180)	−0.21	[−0.571, 0.140]	−0.026 (0.102)	−0.03	[−0.227, 0.176]	0.031 (0.101)	0.03	[−0.159, 0.239]
Block[d]	—	—	—	—	—	—	—	—	—
G1F13 EMSA	0.414 (0.801)	—	[−1.208, 2.008]	0.451 (0.488)	—	[−0.474, 1.455]	0.340 (0.549)	—	[−0.753, 1.443]
G2F13 EMSA	0.617 (0.787)	—	[−0.950, 2.179]	0.318 (0.926)	—	[−1.527, 2.174]	0.214 (0.852)	—	[−1.467, 1.941]
Intercept	−0.190 (0.325)	—	[−0.837, 0.452]	−0.188 (0.203)	—	[−0.603, 0.201]	**−0.519 (0.215)**	**—**	**[−0.943, −0.091]**
Variance components									
Within classroom	**0.031 (0.025)**	**—**	**[0.002, 0.094]**	**0.131 (0.018)**	**—**	**[0.095, 0.167]**	**0.281 (0.019)**	**—**	**[0.245, 0.318]**
Between classrooms	**0.015 (0.014)**	**—**	**[0.001, 0.052]**	**0.014 (0.008)**	**—**	**[0.002, 0.031]**	**0.028 (0.014)**	**—**	**[0.003, 0.058]**
Between schools	**0.033 (0.116)**	**—**	**[0.002, 0.318]**	**0.016 (0.060)**	**—**	**[0.001, 0.148]**	**0.017 (0.065)**	**—**	**[0.001, 0.159]**

(Continued)

	MPAC[a]			ITBS–MP[b]			ITBS–MC[b]		
	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI
Intraclass correlation									
Between classrooms		.190			.087			.086	
Between schools		.418			.099			.052	

Note. Abbreviations and notes as in Table H.1.

[a]MPAC analysis sample size: Student *N* = 619; Teacher *N* = 167; School *N* = 22. MPAC analysis average cluster size: Teacher *N* = 3.707; School *N* = 28.136.

[b]ITBS analyses sample size: Student *N* = 2,159; Teacher *N* = 183; School *N* = 22. ITBS analyses average cluster size: Teacher *N* = 11.798; School *N* = 98.136.

[c]FRL slope is specified to vary randomly across clusters; therefore, the value reported is the intercept for the FRL between-school slope (i.e., the FRL slope, holding all school-level covariates constant at zero).

[d]Block indicates the vector of *n*–1 randomization blocks. Effects for Block are omitted for visual simplicity.

Table H.5. Treatment-by-ELL Moderation Effects on MPAC, ITBS–MP, and ITBS–MC

	MPAC[a]			ITBS–MP[b]			ITBS–MC[b]		
	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI
Within classroom									
Grade 2	**0.705 (0.080)**	**—**	**[0.546, 0.857]**	**1.013 (0.044)**	**—**	**[0.927, 1.098]**	**1.209 (0.051)**	**—**	**[1.108, 1.307]**
Male	**0.119 (0.048)**	**—**	**[0.023, 0.212]**	0.016 (0.026)	—	[−0.036, 0.068]	**0.064 (0.029)**	**—**	**[0.008, 0.120]**
Minority	−0.065 (0.061)	—	[−0.185, 0.055]	**−0.069 (0.032)**	**—**	**[−0.131, −0.006]**	0.068 (0.035)	—	[−0.001, 0.137]
FRL	−0.030 (0.071)	—	[−0.171, 0.108]	**−0.164 (0.036)**	**—**	**[−0.234, −0.094]**	**−0.119 (0.039)**	**—**	**[−0.196, −0.043]**
SWD	**−0.285 (0.105)**	**—**	**[−0.496, −0.082]**	**−0.233 (0.052)**	**—**	**[−0.334, −0.132]**	**−0.288 (0.057)**	**—**	**[−0.400, −0.177]**
G1F13 EMSA	**0.767 (0.048)**	**—**	**[0.671, 0.861]**	**0.668 (0.026)**	**—**	**[0.616, 0.718]**	**0.552 (0.029)**	**—**	**[0.493, 0.608]**
G2F13 EMSA	**0.838 (0.041)**	**—**	**[0.756, 0.919]**	**0.679 (0.025)**	**—**	**[0.628, 0.728]**	**0.496 (0.029)**	**—**	**[0.438, 0.553]**
Between classrooms									
G1F13 EMSA	−0.122 (1.268)	—	[−3.474, 2.097]	0.352 (0.380)	—	[−0.524, 0.905]	—	—	—
G2F13 EMSA	0.806 (1.119)	—	[−1.755, 3.250]	0.852 (0.893)	—	[−0.915, 2.991]	0.789 (1.354)	—	[−1.886, 4.111]
Between schools									
ELL[c]	−0.148 (0.114)	—	[−0.374, 0.074]	−0.098 (0.066)	—	[−0.229, 0.030]	−0.072 (0.076)	—	[−0.230, 0.069]
Treatment	0.099 (0.177)	0.10	[−0.255, 0.454]	0.082 (0.113)	0.08	[−0.139, 0.308]	−0.118 (0.118)	−0.12	[−0.352, 0.114]
Treatment by ELL	−0.112 (0.183)	−0.11	[−0.466, 0.258]	−0.151 (0.097)	−0.15	[−0.337, 0.046]	0.095 (0.112)	0.09	[−0.108, 0.337]
Block[d]	—	—	—	—	—	—	—	—	—
G1F13 EMSA	0.362 (0.739)	—	[−1.085, 1.875]	0.407 (0.419)	—	[−0.389, 1.275]	0.402 (0.491)	—	[−0.562, 1.397]
G2F13 EMSA	0.443 (0.741)	—	[−1.045, 1.917]	0.116 (0.810)	—	[−1.472, 1.768]	0.231 (0.766)	—	[−1.275, 1.765]
Intercept	−0.093 (0.293)	—	[−0.672, 0.488]	−0.201 (0.179)	—	[−0.564, 0.146]	**−0.543 (0.193)**	**—**	**[−0.929, −0.164]**
Variance components									
Within classroom	**0.034 (0.026)**	**—**	**[0.002, 0.099]**	**0.130 (0.018)**	**—**	**[0.094, 0.166]**	**0.279 (0.019)**	**—**	**[0.242, 0.316]**
Between classrooms	**0.017 (0.015)**	**—**	**[0.001, 0.056]**	**0.016 (0.008)**	**—**	**[0.002, 0.034]**	**0.028 (0.015)**	**—**	**[0.003, 0.059]**
Between schools	**0.035 (0.101)**	**—**	**[0.002, 0.289]**	**0.015 (0.043)**	**—**	**[0.002, 0.115]**	**0.015 (0.052)**	**—**	**[0.001, 0.131]**

(Continued)

	MPAC[a]			ITBS–MP[b]			ITBS–MC[b]		
	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI
Intraclass correlation									
Between classrooms		.198			.099			.087	
Between schools		.407			.093			.047	

Note. Abbreviations and notes as in Table H.1.

[c]ELL slope is specified to vary randomly across clusters; therefore, the value reported is the intercept for the ELL between-school slope (i.e., the ELL slope, holding all school-level covariates constant at zero).

[d]Block indicates the vector of *n*–1 randomization blocks. Effects for Block are omitted for visual simplicity.

[a]MPAC analysis sample size: Student *N* = 619; Teacher *N* = 167; School *N* = 22. MPAC analysis average cluster size: Teacher *N* = 3.707; School *N* = 28.136.

[b]ITBS analyses sample size: Student *N* = 2,159; Teacher *N* = 183; School *N* = 22. ITBS analyses average cluster size: Teacher *N* = 11.798; School *N* = 98.136.

Table H.6. Treatment-by-SWD Moderation Effects on MPAC, ITBS–MP, and ITBS–MC

	MPAC[a]			ITBS–MP[b]			ITBS–MC[b]		
	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI
Within classroom									
Grade 2	**0.698 (0.079)**	**—**	**[0.541, 0.850]**	**1.006 (0.044)**	**—**	**[0.920, 1.091]**	**1.209 (0.051)**	**—**	**[1.108, 1.307]**
Male	**0.119 (0.049)**	**—**	**[0.022, 0.213]**	0.017 (0.026)	—	[−0.034, 0.069]	**0.063 (0.029)**	**—**	**[0.007, 0.119]**
Minority	−0.068 (0.063)	—	[−0.191, 0.054]	**−0.073 (0.032)**	**—**	**[−0.135, −0.010]**	**0.069 (0.035)**	**—**	**[0.000, 0.137]**
FRL	−0.037 (0.071)	—	[−0.180, 0.102]	**−0.166 (0.036)**	**—**	**[−0.236, −0.095]**	**−0.121 (0.039)**	**—**	**[−0.198, −0.045]**
ELL	**−0.181(0.069)**	**—**	**[−0.318, −0.047]**	**−0.161 (0.036)**	**—**	**[−0.232, −0.092]**	−0.027 (0.039)	—	[−0.104, 0.049]
G1F13 EMSA	**0.779 (0.049)**	**—**	**[0.680, 0.872]**	**0.669 (0.026)**	**—**	**[0.618, 0.720]**	**0.553 (0.029)**	**—**	**[0.494, 0.608]**
G2F13 EMSA	**0.838 (0.042)**	**—**	**[0.755, 0.921]**	**0.679 (0.026)**	**—**	**[0.628, 0.728]**	**0.494 (0.029)**	**—**	**[0.436, 0.551]**
Between classrooms									
G1F13 EMSA	−0.189 (1.347)	—	[−3.766, 2.104]	0.376 (0.384)	—	[−0.512, 0.929]	**0.811 (0.372)**	**—**	**[0.045, 1.549]**
G2F13 EMSA	0.748 (1.285)	—	[−2.295, 3.492]	0.912 (0.897)	—	[−0.831, 3.086]	0.802 (1.399)	—	[−1.990, 4.239]
Between schools									
SWD[c]	−0.283 (0.196)	—	[−0.662, 0.107]	−0.212 (0.088)	—	[−0.381, 0.036]	**−0.262 (0.119)**	**—**	**[−0.487, −0.015]**
Treatment	0.089 (0.164)	0.09	[−0.242, 0.413]	0.036 (0.104)	0.04	[−0.170, 0.243]	−0.111 (0.111)	−0.11	[−0.331, 0.109]
Treatment by SWD	−0.019 (0.286)	−0.02	[−0.596, 0.538]	−0.034 (0.131)	−0.03	[−0.293, 0.221]	−0.008 (0.175)	−0.01	[−0.363, 0.329]
Block[d]	—	—	—	—	—	—	—	—	—
G1F13 EMSA	0.300 (0.681)	—	[−1.033, 1.697]	0.381 (0.394)	—	[−0.374, 1.190]	0.383 (0.467)	—	[−0.532, 1.335]
G2F13 EMSA	0.429 (0.681)	—	[−0.932, 1.790]	0.212 (0.752)	—	[−1.260, 1.754]	0.232 (0.727)	—	[−1.207, 1.685]
Intercept	−0.078 (0.272)	—	[−0.619, 0.460]	−0.160 (0.167)	—	[−0.498, 0.163]	**−0.524 (0.185)**	**—**	**[−0.894, −0.164]**
Variance components									
Within classroom	**0.039 (0.029)**	—	**[0.002, 0.108]**	**0.133 (0.018)**	—	**[0.097, 0.169]**	**0.277 (0.019)**	—	**[0.241, 0.313]**
Between classrooms	**0.021 (0.017)**	—	**[0.002, 0.063]**	**0.017 (0.008)**	—	**[0.002, 0.034]**	**0.030 (0.015)**	—	**[0.004, 0.061]**
Between schools	**0.028 (0.086)**	—	**[0.002, 0.240]**	**0.012 (0.036)**	—	**[0.001, 0.096]**	**0.013 (0.047)**	—	**[0.001, 0.118]**

(Continued)

	MPAC[a]			ITBS–MP[b]			ITBS–MC[b]		
	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI
Intraclass correlation									
Between classrooms		.239			.105			.094	
Between schools		.318			.074			.041	

Note. Abbreviations and notes as in Table H.1.

[a]MPAC analysis sample size: Student *N* = 619; Teacher *N* = 167; School *N* = 22. MPAC analysis average cluster size: Teacher *N* = 3.707; School *N* = 28.136.

[b]ITBS analyses sample size: Student *N* = 2,159; Teacher *N* = 183; School *N* = 22. ITBS analyses average cluster size: Teacher *N* = 11.798; School *N* = 98.136.

[c]SWD slope is specified to vary randomly across clusters; therefore, the value reported is the intercept for the SWD between-school slope (i.e., the SWD slope, holding all school-level covariates constant at zero).

[d]Block indicates the vector of *n*–1 randomization blocks. Effects for Block are omitted for visual simplicity.

Table H.7. Treatment-by-Pretest Moderation Effects on Grade 1 MPAC, ITBS–MP, and ITBS–MC

	MPAC[a]			ITBS–MP[b]			ITBS–MC[b]		
	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI
Within classroom									
Male	−0.012 (0.070)	—	[−0.149, 0.125]	**−0.135 (0.045)**	**—**	**[−0.223, −0.046]**	−0.032 (0.053)	—	[−0.136, 0.072]
Minority	−0.073 (0.097)	—	[−0.262, 0.118]	**−0.109 (0.055)**	**—**	**[−0.217, −0.001]**	−0.012 (0.065)	—	[−0.139, 0.117]
FRL	0.137 (0.104)	—	[−0.069, 0.341]	−0.075 (0.061)	—	[−0.195, 0.044]	**−0.157 (0.072)**	**—**	**[−0.298, −0.015]**
ELL	−0.163 (0.103)	—	[−0.365, 0.040]	**−0.137 (0.061)**	**—**	**[−0.256, −0.018]**	−0.077 (0.072)	—	[−0.217, 0.064]
SWD	**−0.415 (0.154)**	**—**	**[−0.716, −0.113]**	**−0.239 (0.095)**	**—**	**[−0.425, −0.053]**	**−0.426 (0.113)**	**—**	**[−0.647, −0.205]**
Between classrooms									
Between schools									
G1F13 EMSA[c]	**0.784 (0.139)**	**—**	**[0.503, 1.054]**	**0.801 (0.078)**	**—**	**[0.650, 0.959]**	**0.587 (0.075)**	**—**	**[0.438, 0.734]**
Treatment	0.191 (0.196)	0.20	[−0.209, 0.568]	0.143 (0.147)	0.15	[−0.137, 0.443]	0.050 (0.155)	0.05	[−0.239, 0.376]
Treatment by EMSA	0.192 (0.203)	0.20	[−0.198, 0.605]	0.022 (0.110)	0.02	[−0.202, 0.236]	0.071 (0.105)	0.07	[−0.137, 0.280]
Block[d]	—	—	—	—	—	—	—	—	—
Intercept	0.039 (0.324)	—	[−0.600, 0.693]	0.185 (0.237)	—	[−0.290, 0.652]	0.044 (0.250)	—	[−0.454, 0.538]
Variance components									
Within classroom	**0.376 (0.038)**	—	**[0.311, 0.459]**	**0.487 (0.023)**	—	**[0.445, 0.535]**	**0.678 (0.032)**	—	**[0.619, 0.745]**
Between classrooms	**0.108 (0.041)**	—	**[0.044, 0.203]**	**0.037 (0.015)**	—	**[0.013, 0.073]**	**0.054 (0.020)**	—	**[0.022, 0.102]**
Between schools	**0.061 (0.153)**	—	**[0.003, 0.456]**	**0.046 (0.085)**	—	**[0.007, 0.261]**	**0.043 (0.095)**	—	**[0.003, 0.289]**
Intraclass correlation									
Between classrooms		.198			.065			.070	
Between schools		.112			.081			.055	

Note. Abbreviations and notes as in Table H.1.

[a]MPAC analysis sample size: Student *N* = 336; Teacher *N* = 88; School *N* = 21. MPAC analysis average cluster size: Teacher *N* = 3.818; School *N* = 16.000.

[b]ITBS analyses sample size: Student *N* = 1,025; Teacher *N* = 94; School *N* = 21. ITBS analyses average cluster size: Teacher *N* = 10.904; School *N* = 48.810.

[c]G1F13 EMSA baseline test slope is specified to vary randomly across clusters; therefore, the value reported is the intercept for the G1F13 EMSA test between-school slope (i.e., the baseline test slope, holding all school-level covariates constant at zero).

[d]Block indicates the vector of *n*–1 randomization blocks. Effects for Block are omitted for visual simplicity.

Table H.8. Treatment-by-Pretest Moderation Effects on Grade 2 MPAC, ITBS–MP, and ITBS–MC

	MPAC[a]			ITBS–MP[b]			ITBS–MC[b]		
	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI	Estimate (*PSD*)	Effect size	95% CI
Within classroom									
Male	**0.277 (0.073)**	**—**	**[0.134, 0.422]**	**0.166 (0.042)**	**—**	**[0.083, 0.248]**	**0.157 (0.052)**	**—**	**[0.055, 0.259]**
Minority	−0.061 (0.096)	—	[−0.248, 0.128]	−0.054 (0.052)	—	[−0.155, 0.047]	**0.140 (0.064)**	**—**	**[0.014, 0.267]**
FRL	−0.130 (0.113)	—	[−0.351, 0.091]	**−0.251 (0.058)**	**—**	**[−0.364, −0.139]**	−0.135 (0.072)	—	[−0.275, 0.006]
ELL	−0.138 (0.107)	—	[−0.350, 0.072]	**−0.196 (0.060)**	**—**	**[−0.314, −0.080]**	0.014 (0.074)	—	[−0.130, 0.159]
SWD	−0.246 (0.173)	—	[−0.585, 0.097]	**−0.291 (0.080)**	**—**	**[−0.448, −0.135]**	**−0.274 (0.099)**	**—**	**[−0.468, −0.079]**
Between classrooms									
Between schools									
G2F13 EMSA[c]	**0.843 (0.109)**	**—**	**[0.632, 1.061]**	**0.732 (0.058)**	**—**	**[0.619, 0.845]**	**0.690 (0.079)**	**—**	**[0.541, 0.840]**
Treatment	−0.008 (0.147)	−0.01	[−0.313, 0.271]	−0.064 (0.102)	−0.06	[−0.264, 0.141]	**−0.310 (0.132)**	**−0.31**	**[−0.577, −0.056]**
Treatment by G2F13EMSA	0.165 (0.155)	0.17	[−0.150, 0.465]	0.084 (0.081)	0.08	[−0.075, 0.242]	−0.031 (0.106)	−0.03	[−0.251, 0.168]
Block[d]	—	—	—	—	—	—	—	—	—
Intercept	0.340 (0.233)	—	[−0.109, 0.814]	**0.473 (0.165)**	**—**	**[0.144, 0.796]**	0.126 (0.208)	—	[−0.282, 0.538]
Variance components									
Within classroom	**0.342 (0.035)**	—	**[0.283, 0.417]**	**0.400 (0.020)**	—	**[0.364, 0.441]**	**0.600 (0.029)**	**—**	**[0.546, 0.662]**
Between classrooms	**0.034 (0.026)**	—	**[0.003, 0.098]**	**0.031 (0.013)**	**—**	**[0.011, 0.062]**	**0.074 (0.024)**	—	**[0.038, 0.130]**
Between schools	**0.032 (0.176)**	—	**[0.002, 0.239]**	**0.018 (0.038)**	**—**	**[0.002, 0.118]**	**0.023 (0.057)**	—	**[0.002, 0.183]**
Intraclass correlation									
Between classrooms		.083			.069			.106	
Between schools		.078			.040			.033	

Note. Abbreviations and notes as in Table H.1.

[a]MPAC analysis sample size: Student *N* = 284; Teacher *N* = 79; School *N* = 22. MPAC analysis average cluster size: Teacher *N* = 3.595; School *N* = 16.000.

[b]ITBS analyses sample size: Student *N* = 980; Teacher *N* = 88; School *N* = 22. ITBS analyses average cluster size: Teacher *N* = 11.490; School *N* = 44.545.

[c]G2F13 EMSA baseline test slope is specified to vary randomly across clusters; therefore, the value reported is the intercept for the [c]G2F13 EMSA test between-school slope (i.e., the Pretest slope, holding all school-level covariates constant at zero).

[d]Block indicates the vector of *n*–1 randomization blocks. Effects for Block are omitted for visual simplicity.