U.S. Department of Education
September 2016

# Focusing on Mathematical Knowledge: The Impact of Content-Intensive Teacher Professional Development

**Michael S. Garet**
**Jessica B. Heppen**
**Kirk Walters**
**Julia Parkinson**
**Toni M. Smith**
**Mengli Song**
**Rachel Garrett**
**Rui Yang**
American Institutes for Research

**Geoffrey D. Borman**
University of Wisconsin-Madison

**Thomas E. Wei**
*Project Officer*
Institute of Education Sciences

ies INSTITUTE OF EDUCATION SCIENCES

NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE

This page has been left blank for double-sided copying.

# Focusing on Mathematical Knowledge: The Impact of Content-Intensive Teacher Professional Development

September 2016

**Michael S. Garet**
**Jessica B. Heppen**
**Kirk Walters**
**Julia Parkinson**
**Toni M. Smith**
**Mengli Song**
**Rachel Garrett**
**Rui Yang**
American Institutes for Research

**Geoffrey D. Borman**
University of Wisconsin-Madison

**Thomas E. Wei**
*Project Officer*
Institute of Education Sciences

**NCEE 2016-4010**
**U.S. DEPARTMENT OF EDUCATION**

ies NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE
Institute of Education Sciences

This page has been left blank for double-sided copying.

**U.S. Department of Education**
John King
*Secretary*

**Institute of Education Sciences**
Ruth Curran Neild
*Deputy Director for Policy and Research*
*Delegated Duties of the Director*

**National Center for Education Evaluation and Regional Assistance**
Joy Lesnick
*Acting Commissioner*

**September 2016**

This report was prepared for the Institute of Education Sciences under Contract ED-IES-12-C-0080. The project officer is Thomas E. Wei in the National Center for Education Evaluation and Regional Assistance.

IES evaluation reports present objective information on the conditions of implementation and impacts of the programs being evaluated. IES evaluation reports do not include conclusions or recommendations or views with regard to actions policymakers or practitioners should take in light of the findings in the report.

Garet, M. S., Heppen, J. B., Walters, K., Parkinson, J., Smith, T. M., Song, M., Garrett, R., Yang, R., & Borman, G. D. (2016). *Focusing on mathematical knowledge: The impact of content-intensive teacher professional development* (NCEE 2016-4010). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

This report is available on the IES website at http://ies.ed.gov/ncee.

**Alternate Formats:** Upon request, this report is available in alternate formats, such as Braille, large print, audiotape, or computer diskette. For more information, please contact the Department's Alternate Format Center at 202-260-9895 or 202-205-8113.

This page has been left blank for double-sided copying.

# Acknowledgments

This study represents a collaborative effort among school districts, schools, teachers, researchers, and professional development providers. We appreciate the willingness and commitment of the school districts, schools, and teachers to volunteer for the study, participate in the professional development activities, and respond to requests for data, feedback, and access to classrooms. We benefited from advice provided by the study's expert technical working group. Members included Sybilla Beckmann, University of Georgia; Linda Davenport, Boston Public Schools; Skip Fennell, McDaniel University; Drew Gitomer, Rutgers University; Patrick Meyer, University of Virginia; Christopher Rhoads, University of Connecticut; Jeffrey Smith, University of Michigan; Jon Star, Harvard University; and Suzanne Wilson, University of Connecticut. We also received important feedback related to classroom video analysis from Catherine McClellan, Clowder Consulting.

We appreciate the dedication and close collaboration of our partners at Harvard University: Heather Hill, Daniel McGinn, Corinne Herlihy, and their team of classroom observation raters. We would also like to thank the many individuals who supported or provided the professional development during the study, including the Intel Math instructors and the Intel Math project manager, the Mathematics Learning Community trainers, the Harvard team, and the district-based facilitators. We are also grateful to the AIR staff who closely coordinated the study professional development activities in participating districts: Rachel Croninger, Jennifer Ford, Arayle Freels, Jeanette Moses, Cheryl Pruce, and Abigail Stein; the local videographers who participated in training and collected the classroom videos used in the study; AIR senior staff, Anja Kurki and Nicholas Sorensen, who led the study's recruitment efforts, and Seth Brown, who served as a senior recruiter; and AIR staff, Edith Tuazon and Zipporah Sanders, who provided administrative and communications assistance throughout the project. Finally, we would like to thank our report editor, Linda Schuch, who helped make the report usable and understandable.

# Disclosure of Potential Conflicts of Interest

The research team for this study consisted of a prime contractor, American Institutes for Research (AIR), and two subcontractors, Harvard University and Measured Decisions Inc. None of these organizations or their key staff has financial interests that could be affected by findings from the Impact Evaluation of Math Professional Development. No one on the 9-member Expert Advisory Panel, convened by the research team twice to provide advice and guidance, has financial interests that could be affected by findings from the evaluation.

# Contents

# List of Exhibits

# Executive Summary

Improving math achievement among U.S. students remains a high priority as results from recent math assessments continue to show room for improvement. For example, 60 percent of fourth-graders scored below the proficient level on the 2015 National Assessment of Educational Progress. On the most recent Program for International Student Assessment's math problem-solving test, U.S. 15-year-olds outperformed students in only 6 of the 34 participating countries.

In an era of increasingly rigorous state standards, teachers at all grade levels face heightened expectations to deepen their students' understanding of mathematical concepts. Teachers may thus benefit from professional development (PD) that deepens their own conceptual understanding of math. Elementary school teachers may especially benefit from content-focused PD because they are less likely to formally study math in college than secondary teachers, who tend to specialize in the subject matter they teach. Unfortunately, there is limited convincing evidence to date on the effectiveness of content-focused PD.

This report examines the impact of content-intensive PD on teachers' math content knowledge, their instructional practice, and their students' achievement. The study's PD had three components, totaling 93 hours. The core of the PD was *Intel Math*, an intensive 80-hour workshop delivered in summer 2013 that focused on deepening teachers' knowledge of grades K–8 mathematics. Two additional PD components totaling 13 hours were delivered during the 2013–14 school year: the *Mathematics Learning Community*, a series of five 2-hour collaborative meetings focused on analyzing student work; and *Video Feedback Cycles*, a series of three one-on-one coaching sessions where teachers' lessons were observed and critiqued. The purpose of these two components was to reinforce the math content in Intel Math and help teachers apply the content to improve their instruction.

Grade 4 teachers from 94 schools in six districts and five states participated in the study and were randomly assigned within schools to either a treatment group that received the study PD or a control group that did not receive the study PD. The key findings on the impact of the study PD on teacher knowledge, practice, and student achievement include:

- **The PD had a positive impact on teacher knowledge.** On average, treatment teachers' math knowledge scores on a study-administered math assessment were 21 percentile points higher than control teachers' scores in spring 2014, after the PD was completed.
- **The PD had a positive impact on some aspects of instructional practice, particularly *Richness of Mathematics.*** We assessed teachers on three dimensions of practice: *Richness of Mathematics*, which emphasizes the conceptual aspects of math, such as the use and quality of mathematical explanations; *Student Participation in Mathematics*, which focuses on student mathematical contributions, explanations, and reasoning; and *Errors and Imprecision*, which focuses on incorrect, unclear, and imprecise use of math. On average, treatment teachers had *Richness of Mathematics* scores that were 23 percentile points higher than the scores of control teachers in spring 2014, after the PD was completed. Although treatment teachers also had better average scores for *Student*

*Participation in Mathematics* and *Errors and Imprecision* than did control group teachers, these differences were not statistically significant.

- **Despite the PD's generally positive impact on teacher outcomes, the PD did not have a positive impact on student achievement.** On average, treatment teachers' students scored 2 percentile points lower than control teachers' students in spring 2014 on both a study-administered math assessment aligned with the content of the PD and the state math assessment. This difference was statistically significant for the state math assessment but not for the study-administered assessment. However, the state math assessment difference was not statistically significant in any of our sensitivity analyses.

## Study overview

The study addressed the following research questions:

1. Was the study PD implemented with fidelity? What were the features of the PD as implemented? To what extent did teachers participate in the PD?
2. What was the impact on teachers' content knowledge, teachers' classroom practices, and student achievement, of offering content-focused PD relative to business-as-usual PD?

### *Study design and samples*

The study sample included 221 grade 4 teachers from 94 schools who agreed to participate in the study. The schools were diverse, situated in urban, suburban, and rural settings across six districts and five states, and served students from a range of racial and socioeconomic backgrounds. The schools had self-contained classes in which the teachers taught multiple subjects, including math.

Random assignment of grade 4 teachers occurred separately within each school, generating a treatment group of 104 teachers and a control group of 117 teachers. As expected, there were no statistically significant differences between teachers and students in the two groups on any measured baseline characteristics. The final analysis sample included 165 of these 221 randomly assigned teachers from 73 schools, who remained in their schools teaching grade 4 through the study year, provided all outcome data described in the section on data sources, and taught in a school where at least one grade 4 teacher in the opposite treatment condition also had no missing outcome data. The percentage of randomly assigned teachers who were in the analysis sample was similar by condition (76 percent of treatment teachers, 74 percent of control teachers), and there were no statistically significant differences between teachers in the final analysis sample and teachers not in the final analysis sample on various baseline characteristics, including baseline teacher knowledge, years of teaching experience, level of education, certification status, and number of math courses taken. The grade 4 students in the classes of teachers in the analysis sample were the basis for the student samples.

### *Description of the PD program*

The 93-hour PD program had three interrelated components:

- *Intel Math* (Intel Foundation, 2009), a widely used, 80-hour professional development workshop designed to promote deep understanding of the conceptual foundations and interconnectedness of grades K–8 mathematics topics through solving and discussing math problems. Intel Math is often used by school districts in federally funded Math-Science Partnership programs.
- *Mathematics Learning Community* (Regional Science Resource Center at the University of Massachusetts Medical School, 2011), a series of five collaborative meetings (10 hours total) in which teachers analyzed student work on topics covered in Intel Math.
- *Video Feedback Cycles,* three rounds of individualized, video-based coaching (3 hours total) that provided feedback to teachers on the quality and clarity of their mathematical explanations.

Intel Math was delivered to treatment teachers by experienced instructors (one mathematician and one math educator in each district) in summer 2013. In each district, the grade 4 treatment teachers were joined by about 10 teachers in grades K–3 and 5–8, to emulate typical implementation of Intel Math in which participants span across grades K–8. The five Mathematics Learning Community meetings were led by two trained, district-based facilitators in each district during the 2013–14 school year. Participants in the Mathematics Learning Community included the grade 4 treatment teachers and the grades 3 and 5 teachers who took part in Intel Math. Approximately 5 of the 10 grades K–3 and 5–8 teachers who participated in Intel Math taught grades 3 and 5. The district-based facilitators also delivered the Video Feedback Cycles, three one-on-one feedback sessions with grade 4 treatment teachers during the 2013–14 school year. The feedback was based on video excerpts of teachers' lessons on topics covered in Intel Math and the Mathematics Learning Community, coded with the Mathematical Quality of Instruction (MQI) instrument. The MQI focuses on three dimensions of instructional practice: *Richness of Mathematics* emphasizes the conceptual aspects of math, such as the use and quality of mathematical explanations; *Student Participation in Mathematics* focuses on student mathematical contributions, explanations, and reasoning; and *Errors and Imprecision* focuses on incorrect, unclear, and imprecise use of math.

### *Data sources*

**Data on the implementation of the PD program.** We documented the delivery of each PD component with activity logs completed by the study team or PD facilitators, and we documented teacher participation in PD sessions with detailed attendance records. To describe the features of the PD components as delivered, we video-recorded all group sessions, including the 80 hours of Intel Math and the 10 hours of Mathematics Learning Community meetings in each district. We coded the videos to describe the activities in which instructors/facilitators and teachers engaged, and we coded a subsample of the videos using the MQI to assess the mathematical quality of the discussions. To describe the features of the individually delivered Video Feedback Cycles, we used the feedback forms completed jointly by the MQI raters and the district-based facilitators.

**Data on outcome measures.** We measured teacher knowledge at three time points: in the summer (at baseline), in the fall (after Intel Math), and in the spring (after the full PD program). We measured teacher knowledge using an adaptive assessment provided by the Northwest Evaluation Association (NWEA). Because the NWEA assessment was customizable, we were able to ensure that the content of the assessment aligned with the content of the PD. In particular, the assessment covered five mathematical domains emphasized by the PD: whole numbers; fractions; rational numbers; ratio, proportion, and rate; and linear equations and functions.

We measured instructional practice at two time points: in the fall (after Intel Math) and in the spring (after the full PD program). We measured instructional practice by video-recording participating teachers' lessons and using established procedures to score them on the three MQI dimensions of instructional practice. The first two dimensions, *Richness of Mathematics* and *Student Participation in Mathematics*, were scored on a four-point scale, with a score of 1 indicating no evidence of the practice. The other three possible scores ranged from 2 (low) if the practice was evident and had at least a basic level of quality to 3 (mid) and 4 (high) if the practice occurred with greater intensity and/or at a higher level of quality. The third dimension, *Errors and Imprecision*, was reverse coded on the same four-point scale, with the lowest score (1) being the most desirable because it indicated no errors and imprecision.

We measured student achievement in spring 2014 using an adaptive assessment provided by NWEA. We were able to customize the assessment to ensure that it focused on the mathematical domains covered in grade 4 and emphasized by the PD: namely, whole numbers, decimals, and fractions. All students took the same NWEA assessment, thus ensuring comparability in student outcomes. However, because the state mathematics assessment may be a policy-relevant outcome, we also collected and analyzed these scores for students in spring 2014.

### *Methods*

We conducted descriptive analyses to assess the fidelity of PD implementation, examine the features of the PD components as implemented, and document treatment teachers' participation. We also compared treatment teachers and control teachers in their self-reported, math-related PD experiences during the study year, to determine the contrast in the amount and type of math-related PD experienced by the two groups of teachers.

We assessed the impact of the PD by comparing teacher and student outcomes between the treatment and control groups. Because the study used random assignment, any differences in teacher or student outcomes between the treatment and control groups can be attributed to the study PD and not some other characteristic of the districts, schools, or teachers.

## Detailed summary of findings

**The PD was well implemented with mathematical instructional quality evident most of the time, based on MQI scores.** All three components of the PD were implemented with high fidelity. On average, 96 percent of the expected 80 hours of Intel Math, and 100 percent of the planned Mathematics Learning Community and Video Feedback Cycle hours were delivered. Mathematical instructional quality was evident during most of the whole-group discussion of math content and solution strategies. For example, MQI scores indicated that *Richness of Mathematics* was evident at a low, mid, or high level in 94 percent of the whole-group discussion in Intel Math (53 percent at a mid or high level), and in 93 percent of the whole-group discussion in the Mathematics Learning Community (45 percent at a mid or high level).

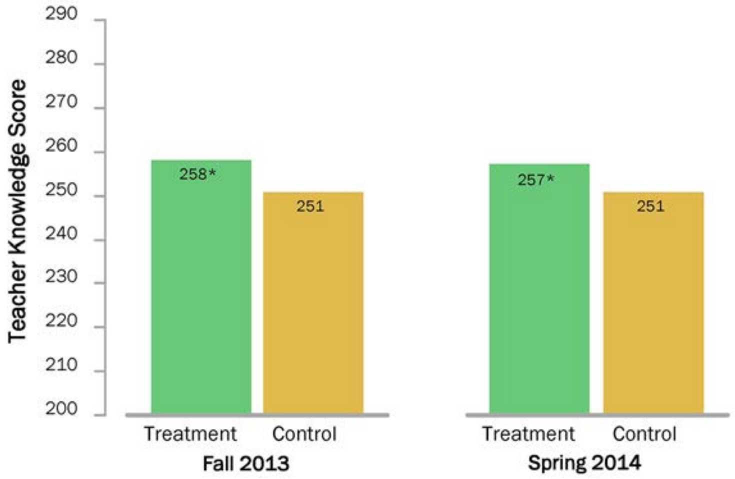
**The PD provided extended time for teachers to solve math problems, analyze student work, explain their solutions to math problems, share their analyses of student work, and receive feedback.** For example, 88 percent of the time in Intel Math and 84 percent of the time in the Mathematics Learning Community were spent in individual or small-group table work and whole-group discussions of the table work, where the above activities would occur. The video-based feedback provided to teachers emphasized the richness of mathematical presentations and discussions, with 82 percent of the feedback focused on this area. The feedback also focused to a lesser extent on identifying and addressing errors and instances of imprecision and lack of clarity, with the remaining 18 percent of feedback focused on this area.

**Treatment teachers' participation in the PD was high, and the contrast between treatment and control teachers' math-related PD was considerable.** On average, treatment teachers participated in more than 90 percent of the implemented hours for each component of the PD program (98 percent of Intel Math, 90 percent of the Mathematics Learning Community, and 97 percent of the Video Feedback Cycles). Treatment and control teachers differed substantially in both the amount and the type of math-related PD in which they participated during the year of the study. Overall, treatment teachers participated in 95 more hours of math-related PD than did control teachers, which is close to the approximately 93 hours of math PD provided by the study. Treatment teachers reported a greater focus on K–8 math content and student thinking in their workshop, study group, and feedback-related PD than did control teachers who reported participating in these types of math-related PD (average differences in reported focus were statistically significant, ranging from 0.7 to 1.3 on a 4-point qualitative survey scale).

**The PD had a positive impact on teacher knowledge.** The PD had a statistically significant impact on teachers' content knowledge in the fall, after the 80 hours of Intel Math were delivered but before the 13 hours of supports for enactment were delivered. This impact was largely sustained into the spring, after all 93 hours of the PD were delivered. Treatment teachers' average knowledge score was 7 points higher than control teachers' average score in the fall, and 6 points higher than control teachers' average score in the spring (see Exhibit ES.1). These differences correspond to an improvement of 24 percentile points in the fall and 21 percentile points in the spring.[1]

[1] More specifically, the math knowledge score for a typical control teacher would have increased from the 50th to the 71st percentile had the teacher received the study PD. This is referred to as the "improvement index," which is based on the outcome distribution within the control group (What Works Clearinghouse, 2014).

**Exhibit ES.1. Teacher Knowledge Scores in Fall and Spring**



Note: Sample size = 73 schools; 79 treatment teachers and 86 control teachers.

The teacher knowledge score is reported on the scale used by the test developer (Northwest Evaluation Association), which takes into account the difficulty of individual test questions in measuring teacher knowledge. The assessment is not typically given to adults; 11th graders are the oldest students for whom norming data are available. The scale shown ranges from 200, the score that corresponds approximately to the 1st percentile for 11th graders, to 290, the score that corresponds approximately to the 99th percentile for 11th graders. In the fall, the average scores correspond to the 84th percentile for treatment teachers and 74th percentile for control teachers. In the spring, the average scores correspond to the 82nd percentile for treatment teachers and 74th percentile for control teachers.

* Difference between the average treatment teacher score and the average control teacher score is statistically significant at the 0.05 level, two-tailed test.

Source: Fall 2013 and Spring 2014 Teacher Knowledge Tests.

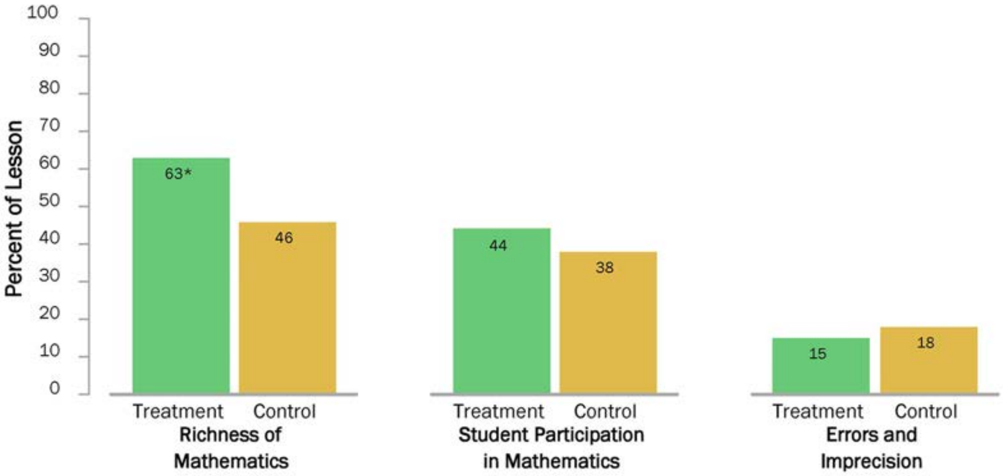
**The PD's impact on teacher knowledge in the spring was larger for teachers with higher baseline knowledge.** The impact of the PD on teachers' knowledge in the fall did not differ for teachers with different levels of prior math content knowledge. The estimated impact of the PD on fall teacher knowledge was positive and statistically significant for teachers with a baseline knowledge score 1 standard deviation above average (improvement index of 25 percentile points) as well as for teachers 1 standard deviation below average (improvement index of 23 percentile points).[2] However, the impact of the PD on teachers' knowledge in the spring was larger for teachers with higher baseline knowledge scores than for teachers with lower baseline scores. As in the fall, the estimated impact of the PD on spring teacher knowledge was positive and statistically significant for teachers with a baseline knowledge score 1 standard deviation above average (improvement index of 34 percentile points) but was not statistically significant for teachers with a baseline knowledge score 1 standard deviation below average (improvement index of 8 percentile points). This finding indicates that while Intel Math on average provided an initial boost to all teachers' content knowledge, the initial boost was not sustained for teachers who began the PD with lower levels of knowledge, even with the additional PD supports over the course of the school year.

**The PD had a positive impact on some aspects of instructional practice, particularly *Richness of Mathematics*.** The PD's effect on *Richness of Mathematics* in the spring was statistically significant and

[2] Teachers' baseline knowledge scores were standardized using the control group mean and standard deviation within the teacher analysis sample.

positive. An average treatment teacher demonstrated *Richness of Mathematics* at a mid or high level during 63 percent of a typical lesson, compared with 46 percent for an average control teacher (see Exhibit ES.2).[3] This 17 percent difference corresponds to an improvement of 23 percentile points. The impact on *Student Participation in Mathematics* and *Errors and Imprecision* in the spring was in the expected direction but not statistically significant (improvement of 6 and −9 percentile points; note that a negative impact on *Errors* is expected because it corresponds to a decrease in *Errors and Imprecision*). We also observed a statistically significant impact on *Student Participation in Mathematics* in the fall (after Intel Math but before the other two PD components) corresponding to an improvement index of 11 percentile points, but it was not sustained into the spring. Impacts on the other two dimensions of practice were in the expected direction in the fall but not statistically significant. In contrast to the teacher knowledge impacts, the impacts on instructional practice did not vary based on teachers' prior math knowledge.

**Exhibit ES.2. Percentage of an Average Teacher's Lesson Demonstrating Three Dimensions of Mathematical Quality of Instruction in Spring**



Note: Sample size = 73 schools; 79 teachers, 158 lessons, and 1,277 7.5-minute lesson segments for the treatment group; 86 teachers, 172 lessons, and 1,352 7.5-minute lesson segments for the control group.

The graph shows the percent of a typical lesson in which an average treatment or control teacher demonstrated each of the three MQI dimensions of instructional quality in spring 2014. Demonstrating *Richness* or *Student Participation* is defined as scoring mid or high on one or more of the elements that comprise the dimension. Demonstrating *Errors* is defined as scoring present (low, mid, or high) on one or more of the elements that comprise the dimension. *Richness of Mathematics* emphasizes the conceptual aspects of math, such as the use and quality of mathematical explanations; *Student Participation in Mathematics* focuses on student mathematical contributions, explanations, and reasoning; and *Errors and Imprecision* focuses on incorrect, unclear, and imprecise use of math. Lower error and imprecision scores are desirable and indicate fewer content errors and less imprecision than higher scores.

* Difference between the average treatment teacher percentage and the average control teacher percentage is statistically significant at the 0.05 level, two-tailed test.

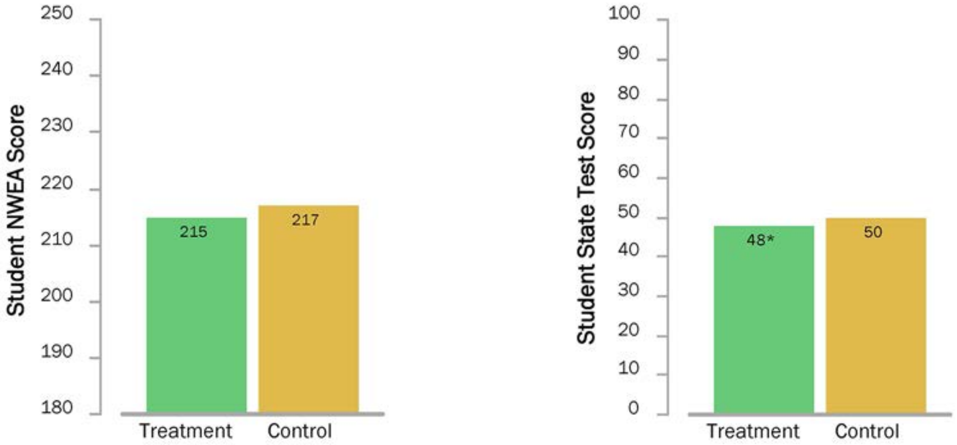
Source: Mathematical Quality of Instruction (MQI) scores of video-recorded lessons from spring 2014.

**Despite the PD's generally positive impact on teacher outcomes, the PD did not have a positive impact on student achievement.** On average, treatment teachers' students scored 2 percentile points lower than

[3] By "demonstrated *Richness of Mathematics*," we mean that the teacher was rated mid or high on one or more of the seven elements that comprise the *Richness of Mathematics* dimension.

control teachers' students on both spring 2014 student achievement measures, including the study-administered math assessment aligned with the content of the PD and the state-specific math assessment (see Exhibit ES.3). The difference between treatment and control group students was statistically significant for the state math assessment but not the study-administered assessment.[4] The results were similar for students who had higher or lower prior achievement, and were also similar for students whose teachers had higher or lower baseline knowledge or more or less teaching experience.

**Exhibit ES.3. Student Math Scores in Spring**



Note: Sample size for analysis of NWEA scores = 73 schools; 79 teachers and 806 students in the treatment group; 86 teachers and 891 students in the control group. Sample size for analysis of state test scores = 73 schools; 79 teachers and 1,760 students in the treatment group; 86 teachers and 1,917 students in the control group. Student sample sizes are smaller for the NWEA assessment because we administered the assessment to a random subsample of students in each class.

The NWEA score is reported on the scale used by the test developer, Northwest Evaluation Association, which takes into account the difficulty of individual test items in measuring student achievement. The scale shown ranges from 180, the score that corresponds approximately to the 1st percentile for fourth graders, to 250, the score that corresponds approximately to the 99th percentile for fourth graders.

The state score is reported using Normal Curve Equivalent (NCE) scores. NCE scores measure a student's position on the normal curve, relative to other students in their state. NCE values run from 0 to 100. They are similar to percentile ranks, but on an equal-interval scale.

* Difference between the average treatment student score and the average control student score is statistically significant at the 0.05 level, two-tailed test.

Source: District administrative records; Spring 2014 NWEA Test.

**Teacher knowledge and instructional practices were generally not correlated with student achievement.** The conceptual framework underlying the study PD assumed that teachers' content knowledge is related to instructional practice, which in turn is related to student achievement. Contrary to these assumptions, both knowledge and instructional practice as measured in the study were not statistically significantly associated with student achievement (estimates of association between 0.00 and −0.05). The only teacher measure associated with student achievement was the *Errors and Imprecision* dimension, which was statistically significantly related to student achievement in the expected direction (estimate of association −0.20).

[4] However, the statistically significant impact on state math assessment scores was sensitive to sample definition and the inclusion of covariates—it was not statistically significant in any of our sensitivity analyses.

### Concluding thoughts

Together these results show that the study PD did change some aspects of teachers' knowledge and classroom practice, but not in a way that led to improved student achievement. This may be partially explained by our finding that the math content knowledge and dimensions of instructional practice targeted by the study PD were generally not correlated with student math achievement. The one exception was *Errors and Imprecision*, on which the study PD did not have a statistically significantly impact. Thus, future research might focus on identifying PD that will improve this aspect of practice. Future research might also seek to identify other aspects of knowledge and practice to target with PD that are more strongly related to improved student achievement.

# I. Introduction

Improving student achievement in mathematics has been a policy priority in the United States for many years (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010; National Mathematics Advisory Panel, 2008; National Research Council, 2001), and results from recent domestic and international mathematics assessments show continued room for improvement. Although fourth-grade student performance on the National Assessment of Educational Progress increased from 1990 to 2013, scores decreased in 2015, and only 40 percent of fourth-graders scored at or above the proficient level (National Center for Education Statistics, 2014; National Center for Education Statistics, 2015). U.S. students in grades 4 and 8 scored above the international average on the math portion of the most recent Third International Math and Science Study, but U.S. 15-year-olds outperformed only 6 of 34 countries on the Program for International Student Assessment's math problem-solving test (Mullis, Martin, Foy, & Arora, 2012; Organization for Economic Cooperation and Development, 2014).

Professional development (PD) for teachers is viewed as an important strategy to improve math achievement. Federal and local governments invest billions of dollars each year in PD programs designed to improve teaching and learning (Birman et al., 2009; U.S. Department of Education, 2014). Despite these investments, there is little rigorous evidence on the effectiveness of PD programs. For example, Yoon, Duncan, Lee, Scarloss, and Shapely (2007) reviewed more than 1,300 studies of PD in mathematics, science, and English language arts and found only nine that examined the impact of PD on student achievement and met the review's criteria based on What Works Clearinghouse (WWC) design standards (What Works Clearinghouse, 2014).

Although the evidence on effective PD is limited, there is growing consensus among mathematicians and math educators that deepening teachers' content knowledge is an essential component of effective math PD in particular (Ball, Thames, & Phelps, 2008; Conference Board of the Mathematical Sciences, 2012; Martin & Umland, 2008; Wu, 2011). They argue that a deep understanding of the content is foundational to delivering the types of instructional practices that may lead to improved student achievement. Yet many teachers—especially at the elementary level—lack formal training in mathematics, including in-depth study of the topics they teach (Conference Board of the Mathematical Sciences, 2012; Greenberg & Walsh, 2008). Because elementary teachers typically teach multiple subjects, it is especially difficult for them to develop math content expertise on the job. In addition, the demands on teachers' content knowledge may be increasing as many states have adopted more demanding content standards and assessments (Glancy et al., 2014; U.S. Department of Education, 2010).

Despite the plausibility of improving students' math achievement by improving teachers' content knowledge, few rigorous studies have focused explicitly on testing this strategy. Only three randomized trials published prior to the launch of the current study had a focus on improving teachers' mathematical content knowledge. Two were studies of the Cognitively Guided Instruction program (Carpenter, Fennema, Peterson, Chiang, & Loef, 1989; Jacobs, Franke, Carpenter, Levi, & Battey, 2007). Both found the program had positive effects on student achievement, although most of the results were not statistically significant. The third was a U.S. Department of Education-commissioned, large-scale experimental study of a two-year math PD program (Garet et al., 2010; Garet et al., 2011). The study, which focused on improving seventh-grade teachers' knowledge of seventh-grade rational number content as well as their pedagogical content knowledge (Shulman, 1986), found no statistically significant impact on student achievement after the first or second year of PD, even though the PD program showed some impact on teachers' instructional practices.[5]

The current study was designed to build off of Garet and colleagues' (2010; 2011) work primarily by testing a PD program with a much more intensive and explicit focus on improving teachers' conceptual understanding of mathematics, not only for the specific grade level content they teach but also more generally across the K-8 spectrum. In addition, building on recent research on the potential value of teacher collaboration (Perry & Lewis, 2011) and video-feedback (Allen, Pianta, Gregory, Mikami, & Lun, 2011), the PD was augmented by collaborative mathematics learning communities and video-based coaching, to help teachers enact their mathematical knowledge in the classroom. The study's 1-year PD program totaled 93 hours and included the following three interrelated components:

- *Intel Math* (Intel Foundation, 2009): A widely used, 80-hour professional development workshop designed to promote deep understanding of the conceptual foundations and interconnectedness of grades K–8 mathematics topics through solving and discussing mathematics problems. Intel Math is often used by school districts in federally funded Math-Science Partnership programs. In the study, Intel Math was delivered to study teachers by experienced instructors (one mathematician and one math educator per site) in the summer of 2013.
- *Mathematics Learning Community* (Regional Science Resource Center at the University of Massachusetts Medical School, 2011): A series of collaborative meetings in which teachers analyze student work on topics covered in Intel Math. The program is designed to complement Intel Math and utilizes trained district-based facilitators. Two facilitators per district delivered five of these 2-hour collaborative meetings (10 hours total) to study teachers during the 2013–14 school year.
- *Video Feedback Cycles:* An individualized, video-based coaching program that provides teachers feedback on the quality and clarity of their mathematical explanations. District-based facilitators delivered three individual feedback sessions per study teacher during the 2013–14 school year. The facilitators used video excerpts of teachers' lessons, coded with a structured rubric, the Mathematical Quality of Instruction (MQI) instrument,[6] as the basis for the feedback. Each lesson emphasized topics covered in Intel Math and the Mathematics Learning Community.

[5] To identify studies of content-focused PD in math, we examined experimental studies included in the Yoon et al. (2007) review and those that were published prior to the current study but included in a subsequently-published synthesis by Gersten, Taylor, Keys, Rolfhus, and Newman-Gonchar (2014). These reviews included studies meeting standards similar to those established by the WWC. For other recent syntheses of impact studies of PD in math, as well as in other subjects, see Blank and de las Alas (2009), Desimone and Stuckey (2014), Kennedy (2016), and Scher and O'Reilly (2009). In addition to the studies cited in the text, these syntheses include some quasi-experimental studies of PD programs that do not appear to meet WWC standards. Two experimental studies of other math content-focused PD programs (Developing Mathematical Ideas and Math Solutions) were underway at the same time as the present study.

[6] The MQI instrument itself is described more fully in chapter II, and how it was used as the basis for feedback in chapter III.

## Conceptual framework

In designing the study, we articulated some potential mechanisms through which the study's PD program might lead to improved student outcomes, drawing on the available evidence and expert opinion (see Exhibit 1.1).[7] The main hypothesis underlying the PD is that by boosting teachers' content knowledge, teachers' instructional quality and their students' achievement can be improved. In defining the aspects of content knowledge of focus in the study, we drew on Ball, Thames, and Phelps' (2008) conceptualization of Mathematical Knowledge for Teaching, which identifies three dimensions of content knowledge: knowledge of how topics build across grade levels (horizon knowledge); knowledge of the math used in general settings (common content knowledge); and knowledge of math particular to teaching, for example, alternative algorithms for fraction operations (specialized content knowledge). Ball and colleagues also have defined three dimensions of pedagogical content knowledge: knowledge of typical student misconceptions and errors (knowledge of content and students), knowledge of examples and concrete materials that facilitate learning (knowledge of content and teaching), and knowledge of the specific materials used in instruction (knowledge of content and curriculum).

The study's PD program focused primarily on teachers' content knowledge, with an emphasis on deepening teachers' common content knowledge, specialized content knowledge, and horizon knowledge. For example, the core of the study's PD, Intel Math, covered math content spanning grades K–8 and included grades K–8 teachers as participants. One purpose of this structure was to help build teachers' horizon knowledge by emphasizing the connections between math content across grade levels and how to mitigate common student misconceptions about interrelated concepts that appear in earlier and later grades. The study's PD program had a secondary focus on deepening knowledge of content and students, one aspect of teachers' pedagogical content knowledge.

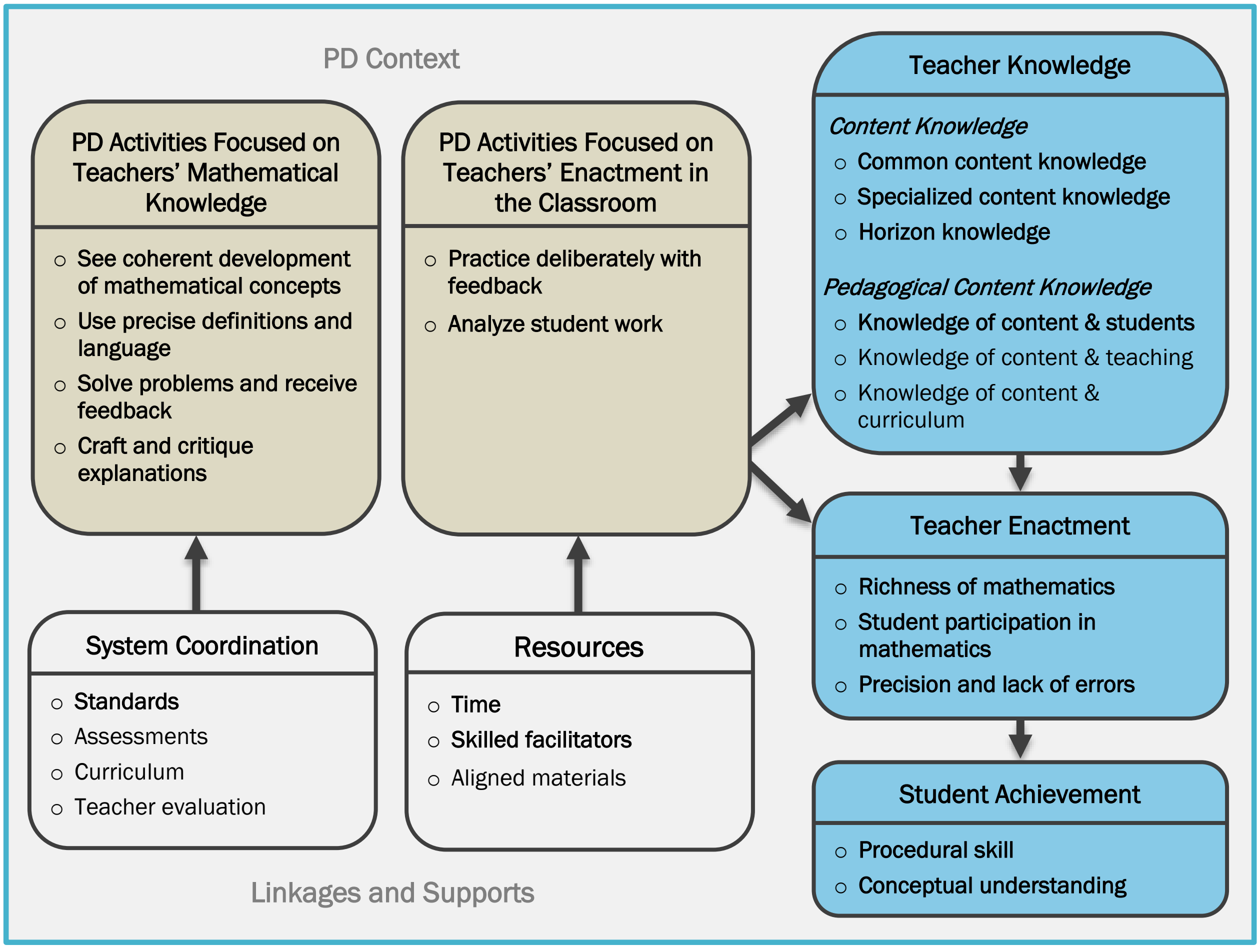
Improvements in student achievement depend not only on teachers' knowledge of math but also on their capacity to draw on their knowledge in their instruction—what we term teacher enactment in the classroom. For purposes of this study, we focused on three ways teachers' mathematical knowledge may be enacted in their teaching: in the richness, coherence, and depth of teachers' mathematics instruction; in their capacity to promote student participation in doing and making sense of mathematics; and in the precision and clarity of their mathematical language (see Hill et al., 2008). When their teaching reflects these qualities, student achievement is expected to improve—both students' procedural fluency and conceptual understanding (Ball, Thames, & Phelps, 2008; Hill, Rowan, & Ball, 2005). (For related frameworks, see Colby, Boston, and Smith [2011] and Ottmar, Rimm-Kaufman, Larsen, and Merritt [2011]).

As shown in the top left boxes of Exhibit 1.1, several features of PD are expected to support the development of teachers' knowledge and enactment in the classroom. In particular, the available research and expert opinion of mathematicians and mathematics educators suggests that PD experiences are likely to promote teachers' learning of mathematics if they are based on a coherent development of the mathematical ideas; if they involve precise definitions and language; if they offer teachers an opportunity

[7] For a related effort to conceptualize mechanisms by which PD might affect teacher and student learning, see Borko (2004).

to solve problems and receive feedback on their work (including homework); and if they provide teachers an opportunity to craft and critique explanations (Burmester & Wu, 2004; Conference Board of the Mathematical Sciences, 2012). The PD program tested in this study had all four of these features.

**Exhibit 1.1. Study Conceptual Framework**


PD Context
PD Activities Focused on Teachers' Mathematical Knowledge
See coherent development of mathematical concepts
Use precise definitions and language
Solve problems and receive feedback
Craft and critique explanations
PD Activities Focused on Teachers' Enactment in the Classroom
Practice deliberately with feedback
Analyze student work
Teacher Knowledge
Content Knowledge
Common content knowledge
Specialized content knowledge
Horizon knowledge
Pedagogical Content Knowledge
Knowledge of content & students
Knowledge of content & teaching
Knowledge of content & curriculum
Teacher Enactment
Richness of mathematics
Student participation in mathematics
Precision and lack of errors
Student Achievement
Procedural skill
Conceptual understanding
System Coordination
Standards
Assessments
Curriculum
Teacher evaluation
Resources
Time
Skilled facilitators
Aligned materials
Linkages and Supports

Note*:* The bolded text refers to elements of the conceptual framework that were emphasized in the study.

In addition, several types of PD experiences are thought to support teachers' enactment of content knowledge in the classroom. In particular, PD experiences may promote enactment if they provide teachers opportunities to analyze student approaches to math problems and opportunities for deliberate practice with feedback, based on a structured rubric (Allen, Hafen, Gregory, Mikami, & Pianta, 2015; Allen et al., 2011; Kraft & Blazar, 2016; Supovitz, 2013). The feedback hypothesis is bolstered by Allen and colleagues' recent experimental studies of My Teaching Partner, which found that the program had positive impacts on student achievement (effect sizes 0.22 to 0.48). The current study's PD program included opportunities for teachers to analyze student approaches and practice with feedback.

The potential impact of any PD strategy is partially dependent on contextual factors: in particular, system coordination and available resources. The study's PD program was designed to attend to some of these contextual factors. In terms of system coordination, there was strong alignment between the math content emphasized in the PD program and the content standards used in the study districts. The study's resources included significant time and incentives to support teacher participation, and the PD facilitators were trained and skilled. Though the study's PD program was not directly and explicitly aligned with the curricular materials used by students, the school-year PD materials corresponded with when topics were taught in each district.

## Study design

This study was a randomized field trial designed to rigorously test the impact of the previously described three-part PD program (Intel Math, Math Learning Community, and Video Feedback Cycles) with fourth-grade teachers. Grade 4 was chosen as the target grade level for three main reasons. First, elementary teachers deliver challenging math content (Wu, 2009), but they typically have limited preservice training formally focused on mathematics. Second, grade 4 falls in the middle of the K–8 content addressed by Intel Math, and thus the PD provided many opportunities to grade 4 teachers to better understand where their students are coming from and where they are heading in terms of the math content. In addition, grade 4 math content (e.g., division and fractions) are especially foundational for later mathematics, including algebra. The introduction of these topics in grade 4 can be challenging for teachers and students alike.

The study was designed to answer the following research questions:

1. Was the study PD implemented with fidelity? What were the features of the PD as implemented? To what extent did teachers participate in the PD?
2. What was the impact on teachers' content knowledge, teachers' classroom practices, and student achievement, of offering content-focused PD relative to business-as-usual PD?

## Organization of this report

The report includes four chapters beyond this introduction. Chapter II describes the design, samples, and analytic methods used in the experimental evaluation. Chapter III provides details about each component of the study's three-part PD program, describes how each component was implemented in the study, and contrasts the amount and type of PD received by treatment and control teachers. Chapter IV presents results of analyses assessing the impact of the study PD on teachers' content knowledge, teachers' classroom practice, and student achievement. Chapter V concludes the report with an examination of the associations among teacher and student outcomes and a discussion of the findings.

This page has been left blank for double-sided copying.

# II. Study Design

The study employed an experimental design, in which volunteer grade 4 teachers from six districts were randomly assigned within schools to treatment and control conditions. The three-part PD program consisted of an 80-hour summer workshop, five 2-hour collaborative meetings, and 3 hours of individualized feedback. It was offered to treatment teachers from summer 2013 through spring 2014. Control teachers did not receive the study PD but could continue to participate in whatever "business-as-usual" PD they would have received in the absence of the study (as could treatment teachers).

This chapter describes the study design, beginning with the process for recruiting districts, schools, and teachers; the definition and characteristics of study samples; and the equivalence of treatment and control groups. This is followed by a description of the data collected for the study, including the measures used to determine the impact of the study PD on three types of outcomes: teacher knowledge, classroom instructional practice, and student achievement. Finally, the chapter describes the methods used for the implementation, impact, and correlational analyses presented in this report. The chapter focuses on describing the most essential aspects of the study design. Readers who are interested in the technical details related to the study design should consult Appendix A.

## Recruitment and study samples

The study team aimed to recruit about 200 volunteer teachers from six districts to meet the study's goals for statistical power. With 200 teachers, the study was powered to detect a minimum effect of 0.30 to 0.40 standard deviations on teacher outcomes and 0.12 standard deviations on student outcomes. The recruitment process began with widespread district outreach followed by district, school, and teacher screening for eligibility and interest. We sought *districts* with at least 16 elementary schools that each had at least two grade 4 teachers, and no conflicting initiatives planned for the 2013–14 school year. We sought *schools* with at least two grade 4 teachers willing to participate in the study because random assignment of teachers to condition was within school. (With two volunteer teachers, one would be assigned to the treatment group and the other to control). In addition, we restricted the sample to schools with non-departmentalized math instruction in order to ensure that the teacher sample would reflect the target population of grade 4 teachers who teach multiple subjects and do not specialize in teaching math. Schools also were ineligible if they sorted students into math classes by ability, as this would mean that the grade 4 classes within schools would not be comparable. At the *teacher* level, the recruitment efforts focused on volunteers, to ensure that teachers were willing to commit the required time. Recruitment efforts also sought teachers whose principals approved their participation. This approach reflected the way intensive PD might be typically rolled out.

In total, 94 eligible schools in six districts located in five states were recruited for the study. These schools were located in urban, suburban, and rural settings and were ethnically diverse. Compared with elementary schools in the national population, study schools were larger and more urban, with a larger proportion of students eligible for free or reduced-price lunch, a lower proportion of White students, and a higher proportion of Hispanic students, as shown in Exhibit 2.1.

**Exhibit 2.1. Characteristics of Schools in the Study Sample and Schools Serving Grade 4 in the National Population**

| Characteristics | Schools in Study Sample | National Population of Schools Serving Grade 4 | Difference | P value |
|---|---|---|---|---|
| Urbanicity (percent) | | | | |
| Urban | 50.5 | 29.0 | 21.5* | <0.001 |
| Suburban | 26.9 | 29.6 | -2.7 | 0.566 |
| Rural | 22.6 | 41.4 | -18.8* | <0.001 |
| Students eligible for free or reduced-price lunch (percent) | 65.6 | 53.7 | 11.9* | <0.001 |
| Racial-Ethnic composition (percent) | | | | |
| White, non-Hispanic | 40.6 | 52.4 | -11.8* | <0.001 |
| Black, non-Hispanic | 20.2 | 15.3 | 4.9 | 0.095 |
| Asian/Pacific Islander, non-Hispanic | 5.1 | 4.3 | 0.8 | 0.168 |
| Hispanic | 30.3 | 23.4 | 6.9* | 0.032 |
| Other, non-Hispanic | 3.8 | 4.6 | -0.8* | 0.002 |
| Total school enrollment (mean) | 534.3 | 450.8 | 83.5* | <0.001 |
| Number of grade 4 students (mean) | 83.8 | 69.6 | 14.2* | <0.001 |
| Number of full-time-equivalent teachers, all grades (mean) | 29.9 | 27.9 | 2.0* | 0.017 |

Note: Sample size = 94 schools in study sample and 53,213 schools serving grade 4 in the national population.

* Difference between schools in the study sample and the national population of schools serving grade 4 is statistically significant at the 0.05 level, two-tailed test.

Source: *Common Core of Data* (CCD), 2011–12 school year.

*Teacher sample at random assignment.* Across the 94 schools recruited for the study, 221 eligible grade 4 teachers volunteered to participate. They were randomly assigned to the treatment or control condition within schools for a total of 104 treatment and 117 control teachers. The treatment and control groups are slightly unbalanced because in most schools with three volunteer teachers, one teacher was randomly assigned to the treatment group and two to the control group, to minimize the chances of "spillover" from treatment teachers to control teachers. We compared background characteristics of teachers in the treatment and control groups to determine whether random assignment produced two groups that were equivalent at baseline. We observed no statistically significant differences on any measured characteristics for teachers in the randomized sample.

*Teacher analysis sample.* To facilitate the interpretation of findings for different teacher outcomes, all analyses of the impact of the study PD on teacher outcomes were based on a common analysis sample, which included teachers who met three criteria: (1) remained in grade 4 over the school year; (2) had data on all outcomes; and (3) were in a school with at least one teacher per condition who also had remained in grade 4 and had all outcome data. This last criterion was necessary because the within-school random assignment design required that schools in the impact analyses have at least one treatment teacher and one control teacher. Of the 221 teachers in the full study sample, 165 teachers (75 percent) in 73 schools met all three criteria and formed the sample for teacher impact analyses. The percentage of randomly assigned teachers who were in the analysis sample was similar by condition (76 percent of treatment teachers, 74 percent of control teachers). Teachers who met the criteria for inclusion in the analysis sample and teachers who did not meet the criteria had no statistically significant differences on baseline characteristics (i.e., teacher knowledge, years of teaching experience, level of education, certification status, and number of math courses taken).

Exhibit 2.2 shows the characteristics of teachers in the teacher analysis sample, as well as grade 4 teachers in the national population. Teachers in the analysis sample differed in two respects from grade 4 teachers in the national population. First, while almost all teachers in the study's analysis sample had standard certification (94 percent), this was statistically significantly lower than the proportion of teachers in the national population with standard certification (99 percent). Second, while teachers in the study's analysis sample were fairly experienced (half had 11 or more years of experience), a higher proportion of teachers in the analysis sample had 3 or fewer years of experience than grade 4 teachers in the national population (16 percent versus 10 percent), and a lower proportion of teachers in the analysis sample had more than 20 years of experience than grade 4 teachers in the national population (17 percent versus 25 percent). Teachers in the analysis sample otherwise had similar characteristics as grade 4 teachers in the national population.

**Exhibit 2.2. Background Characteristics of Grade 4 Teachers in the Analysis Sample and Grade 4 Teachers in the National Population**

| Characteristics | Grade 4 Study Teachers | National Population of Grade 4 Teachers | Difference | P value |
|---|---|---|---|---|
| Standard certification (percent) | 94.0 | 98.9 | −4.9* | 0.008 |
| Years of teaching experience (percent) | | | | |
| 3 years or fewer | 16.4 | 9.9 | 6.5* | 0.032 |
| 4–10 years | 33.3 | 32.5 | 0.8 | 0.840 |
| 11–20 years | 33.3 | 33.1 | 0.2 | 0.956 |
| More than 20 years | 17.0 | 24.5 | −7.5* | 0.020 |
| Master's degree or higher (percent) | 63.6 | 58.1 | 5.5 | 0.171 |

Note: Sample size = 165 grade 4 teachers in the analysis sample and 3,300 grade 4 teachers in schools serving grade 4 students in the national population.

The SASS-based estimates for the national population of grade 4 teachers are approximate, as the teacher weights in SASS are not specifically designed to generate estimates representative of this population.

* Difference between grade 4 study teachers and teachers in the national population is statistically significant at the 0.05 level, two-tailed test.

Source: Spring 2014 Teacher Survey; U.S. Department of Education, National Center for Education Statistics, *Schools and Staffing Survey* (SASS), 2011–12 school year.

The 79 treatment teachers and 86 control teachers in the analysis sample had similar background characteristics, as shown in Exhibit 2.3. In particular, we observed no statistically significant differences between treatment and control teachers' baseline scores on the teacher knowledge assessment administered as part of the study.[8]

[8] As described in the Measures section below, we assessed teacher knowledge at baseline in summer 2013 using a study-administered Northwest Evaluation Association (NWEA) math assessment.

**Exhibit 2.3. Background Characteristics of Treatment and Control Teachers in the Analysis Sample**

| Characteristics | Treatment Group | Control Group | Estimated Difference | P value |
|---|---|---|---|---|
| Teacher knowledge at baseline[a] | 252.8 | 255.3 | −2.5 | 0.186 |
| Standard certification (percent) | 96.2 | 91.3 | 4.9 | 0.189 |
| Years of teaching experience (percent) | | | | |
| 3 years or fewer | 16.5 | 16.5 | 0.0 | 1.000 |
| 4–10 years | 34.2 | 33.0 | 1.2 | 0.874 |
| 11–20 years | 34.2 | 33.4 | 0.8 | 0.918 |
| More than 20 years | 15.2 | 17.3 | −2.1 | 0.732 |
| Master's degree or higher (percent) | 64.6 | 63.8 | 0.8 | 0.916 |
| Calculus course (percent) | 28.0 | 19.6 | 8.4 | 0.227 |
| Number of mathematics courses | 4.4 | 3.5 | 0.9 | 0.088 |

Note: Sample size = 73 schools; 79 treatment teachers and 86 control teachers.

[a] Teacher knowledge scores at baseline are based on a computer-adaptive math assessment provided by the Northwest Evaluation Association, administered in summer 2014. The assessment is typically given to students, not adults. Grade 11 students are the oldest students for whom norming data are available; scores for 11th graders range from 197 (1st percentile) to 286 (99th percentile). On this scale, the average baseline score for treatment teachers corresponds to the 76th percentile, and the average score for control teachers corresponds to the 79th percentile.

The analyses are based on a teacher-level regression controlling for school fixed effects. Treatment group means are unadjusted means. For continuous measures of teacher characteristics, the control group mean was computed by subtracting the estimated difference from the treatment group mean.

None of the differences between treatment teachers and control teachers is statistically significant at the 0.05 level, two-tailed test.

A likelihood ratio test of overall baseline equivalence confirmed that there was not a statistically significant difference between the treatment and control groups across the full set of teacher baseline characteristics (p = 0.29).

Source: Baseline Teacher Knowledge Test; Spring 2014 Teacher Survey.

*Student samples.* We assessed the impact of the study PD on student achievement using two student outcome measures: (1) the state math assessment and (2) a study-administered NWEA assessment. The student sample for the state math assessment outcome analysis included all students in the teacher analysis sample classrooms, based on the first available roster of the 2013–14 school year. A total of 3,677 students (94 percent of the combined 1,865 treatment group students and 2,067 control group students on the first available rosters) composed the student sample for the state math assessment. State math assessment scores were not available for students who had moved from study schools between the first available rosters and the spring state assessment.

The student sample for the study-administered NWEA assessment outcome analysis included approximately 10 students randomly selected from the classrooms of teachers in the teacher analysis sample, based on the first available roster of the 2013–14 school year. We sampled students instead of testing whole classes to reduce time and burden. A total of 1,697 students (85 percent of students randomly selected from the first available rosters that were eligible for testing) composed the student sample for the study-administered NWEA assessment. The final student sample for the NWEA assessment excluded students who had moved from study schools between the first available rosters and the spring, were opted out by parents, or were not tested for other reasons (e.g., absent on the day of testing). (See Appendix A, page A–4 for information about the sample and data collection procedures for the student NWEA assessment.)

Students in the state assessment analysis sample were ethnically diverse: 46 percent White, 14 percent Black, and 30 percent Hispanic. Twelve percent were identified as English language learners, and 14 percent had an individualized education plan indicating special education status. The study-administered assessment analysis sample had similar characteristics.

We observed baseline equivalence on all measured student characteristics for the state assessment analysis sample, as shown in Exhibit 2.4. For the NWEA analysis sample (shown in Exhibit A.5 in Appendix A), baseline equivalence was achieved on all measured characteristics with the exception of gender and Asian/Pacific Islander students: There were more female students in the treatment group than the control group (53.2 percent versus 45.8 percent) and more Asian/Pacific Islanders in the control group than the treatment group (7.4 percent versus 4.9 percent). We thus included both of these measures (gender and race/ethnicity indicators) as covariates in the student impact analyses.

**Exhibit 2.4. Background Characteristics of Treatment and Control Students in the Student State Assessment Analysis Sample**

| Characteristics | Treatment Group | Control Group | Estimated Difference | P value |
|---|---|---|---|---|
| Third-grade math standardized score on spring 2013 state assessment | 0.2 | 0.2 | 0.0 | 0.682 |
| Age (years) | 10.5 | 10.5 | 0.0 | 0.084 |
| Female (percent) | 51.0 | 48.0 | 3.0 | 0.071 |
| Race/Ethnicity (percent) | | | | |
| White, non-Hispanic | 45.8 | 46.6 | −0.8 | 0.560 |
| Black, non-Hispanic | 14.7 | 13.7 | 1.0 | 0.295 |
| Asian/Pacific Islander, non-Hispanic | 5.6 | 6.9 | −1.3 | 0.135 |
| Hispanic | 30.4 | 29.5 | 0.9 | 0.546 |
| Other, non-Hispanic | 3.6 | 3.2 | 0.4 | 0.575 |
| Eligibility for free or reduced-price lunch (percent)[a] | 58.3 | 57.3 | 1.0 | 0.607 |
| English language learner (percent) | 12.6 | 11.4 | 1.2 | 0.424 |
| Special education status (percent) | 14.3 | 13.6 | 0.7 | 0.645 |

Note: Sample size = 73 schools; 79 teachers and 1,760 students in the treatment group; 86 teachers and 1,917 students in the control group.

[a] Estimates for free or reduced-price lunch status were unavailable for two of the six study districts. Sample size = 57 teachers and 1,279 students in the treatment group; 63 teachers and 1,423 students in the control group.

The analyses are based on a two-level model controlling for school fixed effects. Treatment group means are unadjusted means; control group means were computed by subtracting the estimated difference from the treatment group means. Reported means and differences were rounded to the nearest tenth. However, reported p-values were calculated based on the exact (unrounded) differences.

None of the differences between treatment students and control students is statistically significant at the 0.05 level, two-tailed test.

A likelihood ratio test of the overall baseline equivalence confirmed that there was not a statistically significant difference between treatment and control groups across the full set of student baseline characteristics (excluding free or reduced-price lunch, which was missing in two districts) (p = 0.17).

Source: District administrative records.

*Statistical power.* Based on the size of the teacher and student analysis samples, the realized minimum detectable effect sizes were 0.22 to 0.46 standard deviations for teacher outcomes, and 0.08 to 0.10 standard deviations for student outcomes.

## Measures and data collection

The study team gathered data on program implementation, teachers' PD experiences, teacher and student background characteristics, teacher outcomes, and student outcomes during the course of the study, from summer 2013 through summer 2014.

*Measures of implementation.* Throughout the period of implementation (summer 2013 to spring 2014), we collected data to document the fidelity of implementation and teacher participation in the study PD, and to describe the features of each of the three PD components as implemented. To document fidelity, we used forms specific to each PD component (e.g., activity logs), completed by the study team or PD facilitators. To document teacher participation in the PD sessions, we kept detailed attendance records.

To describe the features of the study PD components, we video-recorded all group sessions, including the 80 hours of Intel Math and the 10 hours of Mathematics Learning Community sessions in each district. We coded the video to describe the structural activities in which instructors/facilitators and teachers engaged. For those activities that constituted whole-group discussions of mathematics, we further coded the video to assess the mathematical quality of instruction during those discussions using the Mathematical Quality of Instruction (MQI) rubric, described further below. To describe the features of the individually delivered video feedback cycles, we used the feedback forms completed by the facilitators.

*Measures of teachers' PD experiences.* To assess whether the study PD provided a meaningful contrast between the treatment and control groups in teachers' PD experiences, we administered a teacher survey at the end of the 2013–14 school year. The survey collected information on the types and amount of math-related PD activities in which teachers participated from summer 2013 through spring 2014. The survey also gathered information on aspects of each type of PD, as experienced by the treatment and control teachers. Multiple survey items were averaged to create indices related to (1) mathematical and student thinking activities; (2) math topic focus of the PD; and (3) coherence of the PD with goals, materials, and expectations.

*Measures of background characteristics.* We gathered teacher and student background characteristics to describe the study sample, to test baseline equivalence, and to include as covariates in the impact models. Teacher characteristics were collected as part of the teacher survey administered at the end of the 2013–2014 school year. Student background characteristics were collected via district administrative records, which we requested in summer 2014. These records included student demographic characteristics and student scores on the state math assessment from spring of 2013 (when students were in grade 3).

*Measure of teacher knowledge.* Teachers' mathematical content knowledge was measured using a 35-item, computer-adaptive assessment provided by NWEA. The NWEA assessment was selected as the measure of teacher knowledge for the study because the PD was primarily focused on building teachers' knowledge of math content, rather than other aspects of teacher knowledge, including pedagogical content knowledge. The test was based on an item bank of more than 3,000 items from a wide range of item difficulties within the following mathematical domains covered in Intel Math: whole numbers; fractions; rational numbers; ratio, proportion, and rate; and linear equations and functions. Although the NWEA items had not been administered to adults prior to their use in the study, using a teacher knowledge assessment that provides scores on a stable equal-interval vertical scale for students allowed some degree of comparison between teacher scores and national norms for students in grade 11 (the oldest group of student norms available for the NWEA).

Trained study staff administered the teacher knowledge assessment in person at three time points: baseline (summer 2013), after the summer PD (fall 2013), and after the full PD program (spring 2014). NWEA provided vertically aligned and equal distance RIT (**R**asch Un**it**) scaled scores, which took into account the difficulty of individual test questions in measuring teacher knowledge, for use in analyses.

*Measures of classroom practice.* To measure instructional practice, trained videographers recorded three lessons for each teacher in both the treatment and control conditions. One lesson was recorded after Intel Math (fall 2013), and two additional lessons were recorded after the full PD program (spring 2014). The study team worked with each teacher to schedule observations of introductory lessons on math topics covered in Intel Math and the Mathematics Learning Communities.

The video-recorded lessons were scored with the Mathematical Quality of Instruction (MQI) observation rubric developed by Heather Hill and colleagues at Harvard University (Mathematics Instrument Development Group, 2013). The MQI was selected because it is widely used in studies of classroom instruction in mathematics, and there is some existing evidence for positive associations between MQI scores and student achievement (Blazar, 2015; Hill, Kapitula, & Umland, 2011). The MQI used a 4-point scale to rate 16 elements within three dimensions of instructional practice (see Exhibit 2.5 for details):

- *Richness of Mathematics* emphasizes the conceptual aspects of mathematics, including the use and quality of mathematical explanations, linking between representations, mathematical language, and multiple procedures or solution methods (7 elements).
- *Student Participation in Mathematics* focuses on teachers' use of student mathematical contributions, student explanations, and student mathematical questioning and reasoning (6 elements).
- *Errors and Imprecision* focuses on incorrect, unclear, and imprecise use of mathematics (3 elements).

**Exhibit 2.5. Descriptions of MQI Elements and Dimensions**

| MQI Dimension/Element | Description |
|---|---|
| ***Richness of Mathematics*** | **Focus on the meaning of mathematics and/or mathematical practices** |
| Linking Between Representations | Explicit connections made between mathematical representations, for instance, a written computation and manipulatives both showing long division |
| Explanations | Focus on the "why" associated with problems, procedures, or ideas |
| Mathematical Sense-Making | Focus on meaning of numbers, problems, answers, quantities, procedures |
| Multiple Procedures or Solution Methods | Discussion of different mathematical approaches and how they compare |
| Patterns and Generalizations | Use of examples to derive a mathematical property, extend a mathematical pattern, or build or test a mathematical definition |
| Mathematical Language | Frequent and precise use of mathematical language and encouraging students to do the same |
| Remediation of Student Errors and Difficulties | Targeting the conceptual base/source for mathematical errors and/or calling student attention to errors commonly made by students |
| ***Student Participation in Mathematics*** | **Student contribution to meaning-making and engagement in doing mathematics** |

| | |
|---|---|
| Teacher Uses Student Mathematical Contribution | Teacher use of student contributions (e.g., questions, work, explanations, representations) to develop mathematical ideas and content |
| Students Provide Explanations | Student presentation of a mathematical explanation (as defined above) for an idea, procedure, or solution |
| Student Mathematical Questioning and Reasoning | Student engagement in mathematical thinking that has features of important mathematical practices (e.g., make conjectures, provide counter-claims, ask questions, form conclusions based on patterns) |
| Students Communicate about the Mathematics of the Segment | Student communication of mathematical ideas (e.g., explanations, solutions, questions, methods), either in whole-group or small-group settings |
| Task Cognitive Demand | Student engagement in tasks that require students to think deeply and reason about mathematics (e.g., determine the meaning of mathematical concept, draw connections among representations, make or test conjectures) |
| Students Work with Contextualized Problems | Student work with contextualized problems (e.g., story problems, real-world applications, experiments that generate data) |
| *Errors and Imprecision* | **Incorrect, unclear, and/or imprecise use of mathematics** |
| Mathematical Content Errors | Major mathematical errors (incorrect solution, incorrect definition, etc.) and/or allowing student errors to go uncorrected (except in cases where it is intentional) |
| Imprecision in Language or Notation | Incorrect or imprecise use of mathematical symbols and mathematical terms |
| Lack of Clarity in Presentation of Mathematical Content | Unclear presentation of mathematical content, including unclear launch of mathematical tasks and unclear discussions or presentation of mathematical content |

Following standard procedures for MQI coding, each video-recorded lesson was divided into 7.5-minute segments and coded by two trained MQI raters per lesson. Videos were systematically assigned to raters to ensure a balanced mix of treatment and control teachers' videos for each rater, and raters were blind to condition. Raters scored each segment using the 16 MQI elements plus a holistic score for each overall dimension on a 1–4 scale (1 = not present, 2 = low, 3 = mid, and 4 = high).[9] For elements within the *Richness of Mathematics* dimension and the *Student Participation in Mathematics* dimension, "not present" indicates that a given element of practice was not present in the classroom during the segment.[10] Scores of "low," "mid," and "high" are determined by the intensity and level of the given element of practice. A score of "low" indicates the element of practice was minimally present (low intensity and basic level), a "mid" indicates the element of practice was present but did not characterize the entire segment (moderate intensity and level), and a "high" indicates the element of practice was present in the whole segment or was present during only a portion of the segment but affected the quality of the math over the segment as a whole. For example, within *Richness of Mathematics*, a segment would have been rated as high on linking between representations if links and connections between two mathematical representations were explicit and occurred throughout the segment, or if the explicit connections offered significant insight into the main mathematics of the segment, even though the connections did not occur throughout the segment. For items within the *Errors and Imprecision* dimension, a "not present" indicates no errors or imprecision, a "low" indicates a brief error or imprecision, a "mid" indicates an occasional error or imprecision that obscured the mathematics for part of a segment, and a "high" indicates pervasive errors or imprecisions that obscured the math during the entire segment.

[9] The overall dimension scores for each segment were not included in the construction of instructional practice measures for impact analyses. However, we did use the overall dimension scores for implementation analyses—that is, the analysis of the mathematical quality of the discussions in Intel Math and the Mathematics Learning Community.

[10] Not present could reflect either that there was no opportunity for the element of practice to occur, or there was an opportunity but the element practice was not observed.

We used Rasch scaling to generate a dimension score for each segment for the *Richness of Mathematics* and *Student Participation in Mathematics* dimensions, based on the scores for the individual elements within each dimension. For the *Errors and Imprecision* dimension, we used Rasch scaling to generate a dimension score for each lesson rather than for each segment, because of the large percentage of "not present" scores on the three *Errors and Imprecision* elements. (See Appendix A, page A–15 for more information about the Rasch scoring.)

We then used the Rasch scores to determine the probability that a teacher demonstrated *Richness*, *Student Participation*, or *Errors* during a typical segment. We defined demonstrating *Richness* or demonstrating *Student Participation* in a segment as scoring mid or high on one or more elements of the dimension during the segment. We defined demonstrating *Errors* in a segment as scoring low, mid, or high on one or more elements during the segment. We defined demonstrating *Errors* as scoring low, mid, or high on one or more elements, rather than only mid or high, because the *Errors* dimension taps a negative aspect of instruction, and even low-level errors in the presentation of mathematical content might reduce student learning.

For example, for the *Richness* dimension, we used the Rasch scores to calculate teachers' probability of scoring mid or high on one or more of the seven elements within the *Richness* dimension during a typical 7.5–minute lesson segment. Because teachers tended to receive higher ratings on some MQI elements of *Richness* than others, the probabilities are based on Rasch estimates averaged across all seven elements within the *Richness* dimension.

A typical lesson lasted about 60 minutes, or eight segments. If, for example, an average teacher had a 25 percent chance of demonstrating *Richness of Mathematics* during a typical segment, the teacher would have been expected to demonstrate *Richness* in two segments over the course of a typical lesson, or 25 percent of the lesson.

***Measures of student achievement.*** We used two measures to assess the impact of the PD program on student mathematics achievement in spring 2014: scores on a study-administered assessment provided by NWEA and state mathematics assessment scores provided by study districts.

The study-administered assessment was a 30-item, computer-adaptive test provided by NWEA, administered by trained study staff in spring 2014. The assessment aimed to assess knowledge in the mathematical domains that were both addressed by the study PD program and are typically covered in grade 4 math: whole numbers and decimals, and fractions. Similar to the teacher knowledge assessment, the scores for this test were also on the RIT scale, which took into account the difficulty of individual test questions in measuring student achievement. Although both the teacher assessment and student assessment were provided by NWEA, the teacher knowledge assessment covered a broader range of K–8 math topics, and the teacher knowledge scores should not be compared directly to student assessment scores.

We collected the state assessment scores from study districts. Because the state scores for students in different study districts were from different assessments, we first standardized the state assessment scores within each district using the state mean and standard deviation (May, Perez-Johnson, Haimson, Sattar, & Gleason, 2009). Normal curve equivalent (NCE) scores were then calculated from the standardized scores.

NCE values run from 0 to 100, similar to percentile ranks, but are on an equal-interval scale and are a way of measuring where a student falls along the normal curve.

## Analysis methods

This section summarizes the analysis methods used to estimate impacts. Details on the analytic approaches for outcome analyses, including all estimating equations, and other related issues can be found in Appendix A in the "Analyses" section.

*Teachers' PD experiences.* To examine whether treatment teachers and control teachers differed in the types of PD attended, the hours teachers spent in each type of PD, and the aspects of the math PD they experienced, we estimated teacher-level regressions that modeled measures of teachers' PD experience as a function of treatment status. The model controlled for school fixed effects. Thus, the treatment-control comparisons were estimated within schools, reflecting the within-school random assignment design.

*Impact of the study PD.* We specified different statistical models to assess the impact of the study PD on each type of outcome. Impacts on teacher knowledge were estimated using a teacher-level regression model. Impacts on teacher practice were estimated using multilevel models, one for the fall observation (with lesson segments nested within teachers) and one for the two spring observations (with lesson segments nested within lessons nested within teachers). Impacts on student achievement also were estimated using multilevel models, with students nested within teachers. In addition to school fixed effects, all impact analyses also incorporated a set of covariates (e.g., student and teacher background characteristics) to improve the precision of the impact estimates and adjust for any baseline differences between the study groups.

Before performing the impact analyses, we standardized the teacher knowledge and student achievement measures using the control group means and standard deviations; estimates of the PD's impact on these outcomes therefore can be interpreted as effect sizes. Impact analyses for classroom practice were based on unstandardized Rasch scores at the segment level (for *Richness of Mathematics* and *Student Participation in Mathematics*) or lesson level (for *Errors and Imprecision*). We calculated the effect sizes for the estimated impacts on the classroom practice measures by dividing the impact estimate by the teacher-level standard deviation for the control group.[11]

All teachers and students in the analysis samples had full data on the outcome measures but some had missing baseline data (not including baseline teacher knowledge). In particular, 4 percent of teachers were missing data on whether or not they took a calculus course, 1 percent of students were missing race/ethnicity data, and 6 percent of students were missing the prior year math achievement score. To retain all cases in the sample, missing covariate data were handled using the dummy variable adjustment

[11] Teacher-level practice scores were not directly observed. To compute the effect sizes for classroom practice based on teacher-level standard deviations, we estimated the standard deviation of "observed" teacher-level practice scores based on the variance in the Rasch scores of practice at the segment, lesson, and teacher levels (lesson and teacher levels for *Errors and Imprecision*), as well as the average number of segments per lesson and the average number of lessons per teacher, as appropriate.

approach (Puma, Olsen, Bell, & Price, 2009). Sensitivity analyses conducted for each outcome examined the robustness of the main impact results to alternative definitions of the analysis sample and model specification.

In addition to the main impact analyses, we also examined whether the PD had a larger or smaller impact on teacher knowledge, instructional practice, or student achievement for some types of teachers or students than others. In particular, we tested whether the PD impacts were different for (1) teachers with higher versus lower baseline math knowledge; (2) teachers with more or less teaching experience; (3) teachers with higher versus lower classroom average prior student achievement (i.e., their students' prior-year achievement from grade 3); or (4) students with higher versus lower prior (grade 3) achievement. The models used for these analyses were variants of the main impact models.

*Correlational analyses.* In addition to the main impact analyses and the tests for differential impacts, we examined the correlations between teacher knowledge, instructional practice, and student achievement.

This page has been left blank for double-sided copying.

## III. Design and Implementation of the Professional Development Program

This chapter describes the design and implementation of the PD program examined in the study. The chapter begins by describing the design of the PD program and its three components. We then examine the fidelity of implementation and the features of the components as implemented (e.g., the types of learning activities made available to teachers and the mathematical quality of the discussions). The chapter concludes by examining whether treatment teachers received the intended dosage of the PD and whether the amount of PD received by treatment teachers differed from the amount received by control teachers.

### The PD program was designed to enhance teachers' mathematical knowledge and to support teachers in enacting their knowledge in the classroom

The three components of the PD program totaled 93 hours and included Intel Math, the Mathematics Learning Community, and Video Feedback Cycles. Intel Math was an 80-hour summer workshop focused primarily on enhancing teachers' understanding of mathematics. The Mathematics Learning Community component was a series of five 2-hour collaborative meetings that gave teachers an opportunity to revisit the content that was addressed in Intel Math and develop their understanding of student thinking about that content. The Video Feedback Cycles entailed three hours of individualized, video-based feedback on the math-specific aspects of instruction (e.g., the quality of mathematical presentation and discussions); these provided teachers the opportunity to examine their enactment of the knowledge gained in Intel Math and the Mathematics Learning Community. Intel Math and the Mathematics Learning Community were existing programs that have been implemented together in some Math-Science Partnership programs. The Video Feedback Cycle component was developed for the study based on emerging research on individualized feedback for teachers. All three components shared the same mathematical focus, primarily on number and operations, and algebra. All three components intended to build off of each other to form a coherent package of PD activities.

The next three sections describe the three components, how each component was meant to build off of the other components, and how each component was *intended* to be implemented in the study. Later on in the chapter, we focus on how the three components were *actually* implemented in the study. Readers who are interested in seeing detailed examples of the PD materials should consult Appendix B.

***Design of Intel Math.*** Intel Math is a widely used PD workshop designed to promote deep understanding of the conceptual foundations and interconnectedness of grades K–8 mathematics topics, primarily by engaging participants in solving mathematics problems. Since 2007, more than 6,000 teachers have participated in Intel Math, representing more than 200 cohorts delivered in 16 states. The program was initially developed by Ken Gross as part of the Vermont Mathematics Initiative, a master's degree program for in-service teachers, and is currently managed by the Institute for Mathematics Education at the University of Arizona. Intel Math is ordinarily delivered to groups of teachers who span grades K–8, making it possible to compare the solution strategies for the same problems used by teachers in different grades (e.g., solutions based on arithmetic or algebra).

In the study, treatment teachers participated in an Intel Math workshop held in their district over 10 or 13 days in the summer. In addition to the grade 4 treatment teachers, participants included a total of approximately 10 teachers from each district who did not teach grade 4. Approximately half of these 10 teachers taught grades K–3, and the other half taught grades 5–8. In that way, implementation for the study mimicked that of typical Intel Math implementation.

The content of Intel Math is organized into eight units. Three of the eight units focus on key grade 4 topics: multiplication, division, and fractions. Two units focus on topics introduced before grade 4 (addition and subtraction), and the remaining three units focus on topics typically introduced after grade 4 (rational numbers, linear relations, and functions). Exhibit 3.1 lists the eight units of Intel Math and the content within each unit.

**Exhibit 3.1. Intel Math Unit Titles and Content**

| Intel Math Unit | Content |
|---|---|
| Unit 1: Addition | Different methods of solution; interconnectedness of arithmetic, algebra, and geometry; meaning of "equals"; adjective-noun theme for addition; different ways of solving problems; solving simple equations; student understanding of addition |
| Unit 2: Subtraction | Properties of number systems; meaning of subtraction; adjective-noun theme for subtraction; alternative algorithms for subtraction; processes and inverse processes; addition and subtraction of signed numbers; student understanding of subtraction |
| Unit 3: Multiplication | Meaning of multiplication; adjective-noun theme for multiplication; distributive property; area model for multiplication; multiplication of signed numbers; primes and composites; Least Common Multiple (LCM) and Greatest Common Factor (GCF); problem-solving with LCM and GCF; student understanding of multiplication |
| Unit 4: Division | Meaning of division; models for division; adjective-noun theme for division; types of division—partitive and quotative; introduction to rates; student understanding of division |
| Unit 5: Operations with Fractions | Meaning of fractions; models for fractions; representations for fractions; equivalent fractions; addition and subtraction of fractions; multiplication of fractions; division of fractions; fractions in context; making sense of fractions; student understanding of fractions |
| Unit 6: Rational Numbers | Exponents; decimals; algebraic fractions; rates revisited; rates in context; student understanding of place value |
| Unit 7: Linear Relations | Relationship between feet and inches; everyday examples of linear relations; slope of a line; slope-intercept form; point-slope form; standard form; parallel and perpendicular lines; linear equations in context; student understanding of linear equations |
| Unit 8: Functions | Transitioning from processes to functions; features of functions (domain and range; inverse functions; onto; one-to-one; function notation); linear functions; functions in context; composition of functions; distance function; course capstone |

Source: Intel Math materials (Intel Foundation, 2009).

Units consist of three to five sessions, each covering 1 or 2 hours of material, for a total of 42 sessions across the eight units. Thirty-five of the sessions (and 90 percent of the time) focus on math content and emphasize the conceptual foundations for common mathematical procedures; connections between mathematics concepts, including connections between arithmetic and algebra; and multiple ways of solving the same problem. Seven of the sessions (one per unit, except the last unit, covering 10 percent of the session time) are dedicated to examination of student learning trajectories and analysis of student approaches to math problems.

Teachers participate in several types of activities throughout Intel Math. During each session, teachers engage in instructor-led, whole-group discussion of new material, which is presented in information sheets that describe the key mathematical concepts that are the focus of the session. In addition, teachers work individually or in small groups at their tables to solve mathematics problems or analyze examples of student work to identify what the student does not understand and suggest potential next steps for that student. Finally, teachers present and discuss the work done at their tables. Intel Math instructors also assign homework, consisting of reading and additional mathematics problem sets, typically discussed at the start of the next day. In addition, the 80-hour Intel Math workshop includes discussion of daily course evaluations completed by participating teachers, and completion of an Intel Math-developed pre- and post-test.

To facilitate conversations of mathematics content and student thinking, each Intel Math workshop is co-led by one mathematician and one mathematics educator. For the study, staff at the Institute for Mathematics Education at the University of Arizona chose the instructors and assigned them to study districts. The instructors were quite experienced. The 12 instructors (two per district) had, on average, 4.8 years of experience leading Intel Math, and five of the six pairs had previously co-led Intel Math together. All 12 instructors had a master's degree or higher in mathematics or mathematics education.

*Design of the Mathematics Learning Community.* The full Mathematics Learning Community program is a series of 15 2-hour collaborative meetings for teachers developed by the Regional Science Resource Center at the University of Massachusetts Medical School in collaboration with the Massachusetts Department of Elementary and Secondary Education. The program addresses the same topics as Intel Math and was designed as a form of job-embedded PD to provide teachers with an opportunity to revisit the content addressed in Intel Math, continue to deepen their knowledge of grades K–8 mathematics and student thinking, and make connections to their instruction. The 15 meetings are typically implemented over a 2-year time period, and program participants typically include teachers from different grade levels, providing opportunities for teachers to discuss how concepts and skills develop across grades. Exhibit 3.2 lists the 15 Mathematics Learning Community sessions that compose the full program.

**Exhibit 3.2. Mathematics Learning Community Sessions**

| Mathematics Learning Community Session Number | Title/Content Addressed |
|---|---|
| Session 1 | Getting Started (orientation, greatest common factor, and least common multiple) |
| Session 2 | Understanding Counting (skip counting and time conversion) |
| Session 3 | Working with Addition |
| Session 4 | The Relationship Between Addition and Subtraction |
| Session 5 | Subtraction Strategies |
| Session 6 | Multiplication Strategies |
| Session 7 | The Distributive Property |
| Session 8 | Dealing with Division (division strategies) |
| Session 9 | Partitive and Quotative Division |
| Session 10 | Interpreting Remainders in Division Contexts |
| Session 11 | Representing and Interpreting Fractions |
| Session 12 | Adding and Subtracting Fractions |
| Session 13 | Multiplying Fractions |
| Session 14 | Dividing Fractions |
| Session 15 | Working with Fractions, Decimals, and Percents |

Source: Mathematics Learning Community materials (Regional Science Resource Center at the University of Massachusetts Medical School, 2011).

In collaboration with the participating districts, the study team determined that 5 of the 15 Mathematics Learning Community sessions could feasibly be held during the school year, in an afterschool setting. The five meetings were selected by each district and the study team from among the six that focused on the key grade 4 topics addressed in Intel Math (multiplication, divisions, and fractions): Multiplication Strategies, The Distributive Property, Dealing with Division, Partitive and Quotative Division, Representing and Interpreting Fractions, and Adding and Subtracting Fractions. Exhibit 3.3 displays the order of the five Mathematics Learning Community sessions implemented by each of the six study districts.

**Exhibit 3.3. Mathematics Learning Community Meetings Implemented in Each Study District**

| | Mathematics Learning Community Session | | | | |
|---|---|---|---|---|---|
| District | 1 | 2 | 3 | 4 | 5 |
| A | Multiplication Strategies | Distributive Property | Dealing with Division | Representing and Interpreting Fractions | Adding and Subtracting Fractions |
| B | Distributive Property | Dealing with Division | Partitive and Quotative Division | Representing and Interpreting Fractions | Adding and Subtracting Fractions |
| C | Distributive Property | Dealing with Division | Partitive and Quotative Division | Representing and Interpreting Fractions | Adding and Subtracting Fractions |
| D | Multiplication Strategies | Representing and Interpreting Fractions | Distributive Property | Dealing with Division | Adding and Subtracting Fractions |
| E | Multiplication Strategies | Dealing with Division | Partitive and Quotative Division | Representing and Interpreting Fractions | Adding and Subtracting Fractions |
| F | Dealing with Division | Representing and Interpreting Fractions | Adding and Subtracting Fractions | Multiplication Strategies | Distributive Property |

Note: The districts are shown in no particular order.

Source: Study records.

Each Mathematics Learning Community meeting was delivered prior to the time at which the topic covered was scheduled to be taught by the grade 4 treatment teachers in each district. This schedule enabled treatment teachers to enact what they learned in their instruction. To support cross-grade conversations that typically occur in a Mathematics Learning Community, participants in the five meetings included the grades 3 and 5 teachers who had participated in Intel Math, along with the grade 4 treatment teachers. In each district, approximately five of the Intel Math participants taught grades 3 or 5.

Each 2-hour meeting is organized in four parts, with activities similar to those in Intel Math. The first two parts, which take up about 40 percent of the meeting, focus on mathematics content. Facilitators lead a discussion of a mathematical topic and connections to other topics, and teachers solve a problem addressing that topic and discuss solution methods. The second two parts, which take up about 60 percent of the session, focus on student thinking and classroom connections. Teachers analyze and discuss student work for a problem similar to the one they solved and then reflect on their own learning and make connections to their instruction. In addition to the four main parts of each Mathematics Learning Community session, time is dedicated to completing and discussing teachers' evaluations of the prior session. Unlike Intel Math, there is no homework.

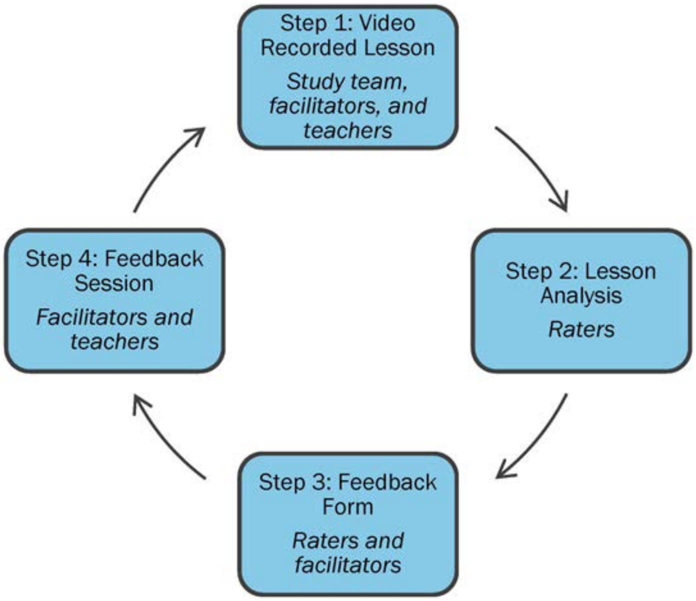
To support teachers in making connections to their instruction, the Mathematics Learning Community meetings are typically led by two district-based facilitators who are familiar with the district context. For the study, the two district-based facilitators additionally led the third component of the PD program (the Video Feedback Cycle) and helped teachers make connections among the three components of the PD program. To select the facilitators, the study team worked with each district to identify individuals who were familiar with the district context; had experience teaching mathematics; and had experience working with teachers in professional learning environments, including instructional coaching.

All 12 of the facilitators hired for the study had prior experience teaching math. All but one had experience working with teachers in professional learning environments, and six had prior experience as instructional coaches in mathematics. To prepare for their roles, facilitators attended Intel Math with the teachers in their district; completed training activities led by the Mathematics Learning Community program developers and by study team members from American Institutes for Research and Harvard University's Center for Education Policy Research; and completed a crosswalk that outlined the alignment among the content of Intel Math, the Mathematics Learning Community, and the district curriculum. This crosswalk was intended to support the facilitators in helping teachers make connections between the PD and their curriculum.

*Design of the Video Feedback Cycles.* The Video Feedback Cycle component was designed for the study to further support teachers in enacting the knowledge gained in Intel Math and the Mathematics Learning Community in their classrooms. It entailed three rounds of video-based, instructional feedback for the grade 4 treatment teachers and was structured to provide explicit feedback on math instruction in a standard way across the study districts. The feedback focused on the math-specific aspects of instruction rather than classroom management or relationships with students. The basis for the feedback was analysis of teachers' lessons with the Mathematical Quality of Instruction (MQI) by trained raters associated with the Center for Education Policy Research at Harvard University. As noted, the district-based facilitators who led the Mathematics Learning Community facilitated the Video Feedback Cycles. Each teacher worked with the same facilitator for all three cycles.

Each cycle consisted of four steps, illustrated in Exhibit 3.4.

**Exhibit 3.4. Steps in a Video Feedback Cycle**



In step one, the study team and one of the district-based facilitators worked with each treatment teacher to identify and video-record a lesson on which the teacher would receive feedback. Lessons were selected on the basis of two criteria: (a) they addressed content previously covered in Intel Math and a prior Mathematics Learning Community session; and (b) they introduced students to new material (as opposed to a review). The target topics were the key grade 4 topics covered in Intel Math: multiplication, division, and fractions.

In step two, a certified rater analyzed the lesson using the MQI instrument, focusing on two of the three dimensions of mathematical quality of instruction measured by the MQI: *Richness of Mathematics* and *Errors and Imprecision*. To be certified, MQI raters complete a standardized training and assessment program that involves coding videos and demonstrating high levels of agreement with master-coders for selected videos. The *Richness of Mathematics* elements emphasize the conceptual aspects of mathematics, including sense-making, multiple procedures, linking between representations, and remediation of student errors. The *Errors and Imprecision* elements focus on content errors, imprecision, and lack of clarity in presentation of mathematical content. The conceptual aspects of mathematics and correct, clear, and precise mathematical communication are emphasized in Intel Math and the Mathematics Learning Community. Teachers could, therefore, refer to the materials from each program to consider ways of designing instruction that emphasized the conceptual aspects of mathematics, with clarity and few errors. The *Student Participation in Mathematics* dimension was not included in the Video Feedback Cycles because the elements in this dimension focus on ways students engage in mathematics (e.g., providing explanations, raising mathematical questions, making conjectures), and Intel Math and the Mathematics Learning Community do not provide direct and explicit guidance on pedagogical moves to engage students in these ways.

Exhibit 3.5 shows the elements of each dimension that were the focus of the analysis. Brief definitions of each element and dimension can be found in Exhibit 2.5 in chapter II.)

**Exhibit 3.5. Dimensions and Elements Used for Lesson Analysis in the Video Feedback Cycles**

| Dimension | Elements of Interest for Video Feedback Cycle |
|---|---|
| *Richness of Mathematics* | Linking between representations |
| | Explanations |
| | Mathematical sense-making |
| | Multiple procedures or solution methods |
| | Mathematical language |
| | Remediation of student errors and difficulties |
| *Errors and Imprecision* | Mathematical content errors |
| | Imprecision in language or notation |
| | Lack of clarity in presentation of mathematical content |

After viewing the videos for instances of strengths or areas for improvement with respect to the elements listed in Exhibit 3.5, the rater identified three or four short video clips, totaling no more than 10 minutes in length, that illustrated a minimum of one area of strength and two areas for improvement.

In step three, the rater and district-based facilitator assigned to that teacher completed a feedback form. First, the rater described the strengths and weaknesses associated with each video clip and for the lesson overall. Then, for each of the weaknesses associated with each video clip, the rater provided actionable suggestions for improvement. The suggestions for improvement were informed by the rater's analysis of what the teacher could have done to improve their MQI scores. Although the rater paid attention to the MQI scores in constructing the feedback, the focus of Video Feedback Cycles was not the scores. Instead, the focus was on qualitative feedback, and teachers did not see their MQI scores. They only saw descriptions of strengths, descriptions of weaknesses, and suggestions for improvement. After the rater completed this part of the form, the facilitator read the information, watched the video, and entered information in the form identifying places in the Intel Math and Math Learning Community materials that the teacher could reference to improve instruction in the two focal areas of the Video Feedback Cycles: emphasizing the conceptual aspects of mathematics and reducing instances of errors, lack of clarity, and imprecision.

In the fourth step, the facilitator met individually with the teacher for approximately 1 hour to review the feedback and watch the video clips, discuss the connections to Intel Math and the Mathematics Learning Community, and identify next steps the teacher could take to improve instruction in the current and next unit they were teaching. These next steps were entered into the feedback form during the meeting so that the teacher had a record of the feedback and ways of addressing issues raised in that feedback.

These sections of the chapter have focused on describing the *design* of the three PD components and how they were intended to build upon each other. The following sections below present findings from analyses of data collected during *actual implementation* of the study PD. These analyses focused on the fidelity of implementation and the features of each component of the PD program as implemented.

## The PD program was well implemented with mathematical instructional quality evident most of the time, and it provided opportunities for teachers to solve problems, analyze student work, and receive feedback

More specifically, key findings include the following:

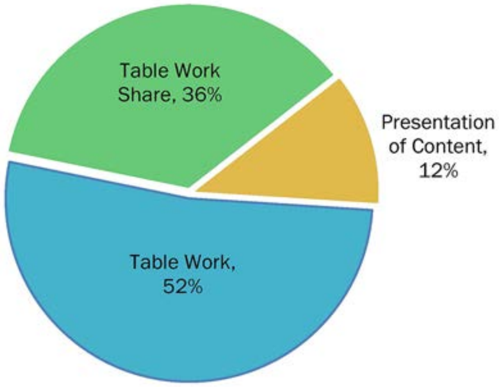
- All three components of the PD program were implemented with high fidelity.
- The PD program provided teachers extended time to solve mathematics problems, analyze student work, explain their solutions to mathematics problems, and share their analyses of student work.
- Scores on the MQI indicated that mathematical instructional quality was evident at a low, mid, or high level in most of the whole-group discussions of math content and solution strategies.
- The video-based feedback provided to teachers on their mathematics instruction emphasized the richness of mathematical presentations and discussions.

These results are elaborated in the following three sections, which examine the implementation of each component.

***Implementation of Intel Math.*** We found that Intel Math was implemented with high fidelity. Analyses of the fidelity forms completed during in-person observations indicated that, on average, 96 percent of the expected 80 hours were delivered in the six districts (mean: 77 hours, range: 72–80 hours), with four of the six districts covering each of the 42 sessions across the eight Intel Math units and the remaining two districts covering 40 of the 42 sessions. As expected, on average, approximately 90 percent of the time spent in the Intel Math sessions was spent in math-focused sessions, and approximately 10 percent was spent in those sessions that focused on student approaches to mathematics. Also as expected, 30 percent of the time, on average, was spent in units focused on the key grade 4 topics of multiplication, division, and fractions.[12]

Our analysis of video recordings of the Intel Math sessions indicated that teachers spent the majority of time solving problems and sharing their thinking with others. As shown in Exhibit 3.6, 12 percent of the time was spent in presentation of content, where teachers and instructors discussed new material. More than half of the time (52 percent) was spent in table work, in which teachers worked individually or in small groups; approximately one third of the time (36 percent) was spent in table work share, where teachers participated in whole-group discussion of the table work. Additional analyses of video recordings of table work indicated that 89 percent of the time was dedicated to solving mathematics problems and 11 percent to analyzing student work.[13]

**Exhibit 3.6. Mean Percentage of Time Spent in Presentation of Content, Table Work, and Table Work Share in Intel Math Sessions**



Note: Sample size = 6 districts. Percentage of time spent in these three activities is out of total time spent in Intel Math sessions; that is, nonsession time (daily evaluations, discussion of homework, and other nonsession activities) is excluded from the denominator.

Source: Intel Math videos.

To further describe the features of Intel Math as delivered, we used the MQI to analyze the mathematical quality of the whole-group discussions (i.e., presentation of content and table work share) and found frequent evidence of mathematical quality in Intel Math.[14] That is, rich mathematics and teacher participation in mathematical reasoning and communication were common, and uncorrected content

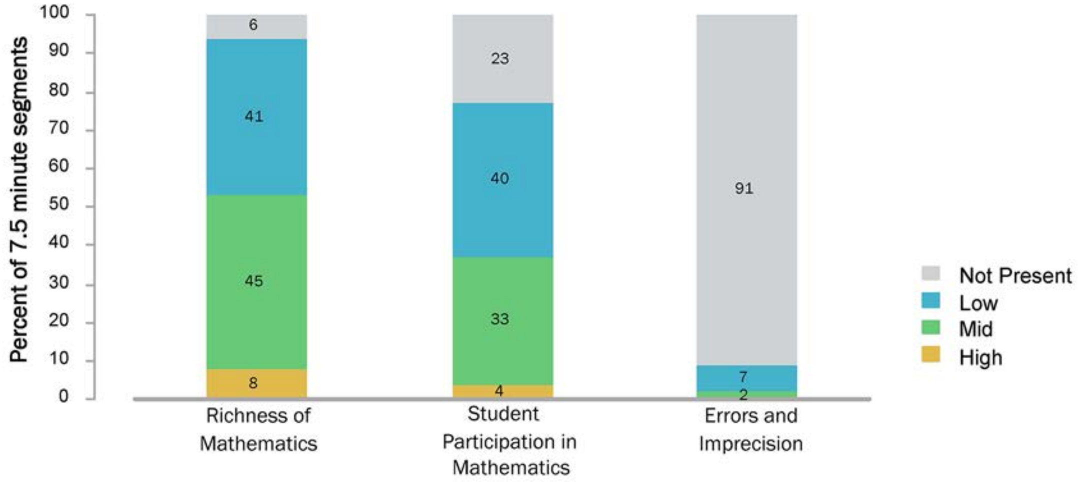
[12] Analyses of time spent in Intel Math content sessions and on grade 4 units excluded time spent on evaluations, homework, and the Intel Math-developed pre- and post-test.

[13] Analyses of time spent in table work, table work share, and presentation of content in Intel Math excluded time spent on evaluations, homework, and the Intel Math-developed pre- and post-test.

[14] We did not use the MQI to analyze table work because the audio was insufficient to assign valid scores to those portions of Intel Math.

errors, lack of clarity, and imprecision were rare. To conduct this analysis, certified raters used the MQI to code each 7.5-minute segment of whole-group discussion during a sample of units across the Intel Math program, including some that focused on grade 4 topics.[15] Exhibit 3.7 shows the percentage of 7.5-minute segments in which each of the three dimensions of the MQI was rated not present, low, mid, and high across the six districts.[16] Note that, for Intel Math, the "students" are the participating teachers.

**Exhibit 3.7. Percentage of Segments in Intel Math Rated Not Present, Low, Mid, and High on Each MQI Dimension**



Note: Sample size = 654 7.5-minute segments across 6 districts. The coded segments occurred during Presentation of Content and Table Work Share in Intel Math mathematics content-focused sessions. Percentages shown are the percent of segments rated not present, low, mid, or high for each dimension (*Richness of Mathematics*, *Errors and Imprecision*, and *Student Participation in Mathematics*).

Source: Intel Math videos.

The results shown in Exhibit 3.7 indicate that mathematical instructional quality was evident during much of the whole-group discussion of math content and solution strategies. For example, rich mathematics was evident at a mid or high level in more than half (53 percent) of the coded segments. *Student Participation in Mathematics*, including participants' engagement in mathematical reasoning and communication, was evident at a mid or high level in 37 percent of the coded segments. Errors, instances of imprecision, and lack of clarity were rare (not present in more than 90 percent of the coded segments).

To draw firm conclusions about the extent to which Intel Math, as implemented in the study, exhibited high mathematical quality, it would be necessary to compare these results with MQI data from other PD programs or another external standard, but these are not available. To create a proxy comparison, we examined the distributions of MQI scores of the lessons of control teachers in the study. Specifically, we examined the scores assigned by raters to the 7.5-minute segments for whole-group, mathematical conversations in the three lessons per control teacher (258 coded lessons total from the 86 control

[15] Analyses focused on the mathematics content-focused sessions of Units 1, 3, 5, and 7.

[16] As noted in chapter II, raters assign holistic scores for each overall dimension (*Richness, Student Participation,* and *Errors and Imprecision*) to each segment, in addition to scores for the individual elements within dimensions. For simplicity of presentation, the MQI analyses of the PD videos use these overall dimension codes.

teachers).[17] We found rich mathematics was evident at a mid or high level in 32 percent of the coded segments in teachers' lessons compared with 53 percent for Intel Math. Student engagement in mathematical reasoning and communication was evident at a mid or high level in 23 percent of the coded segments in teachers' lessons compared with 37 percent in Intel Math. Finally, uncorrected errors, instances of imprecision, and lack of clarity were not present in 73 percent of the coded segments in teachers' lessons, compared with 91 percent in Intel Math. These results illustrate that mathematical quality was more frequently evident in Intel Math, relative to typical classroom teachers' lessons.[18]

Further analyses of the features of Intel Math as implemented indicated that teachers had additional opportunities to solve problems and discuss mathematics through homework assignments. Analyses of the fidelity forms completed during in-person observations showed that instructors assigned homework on 82 percent of the Intel Math days across the six districts, on average. Analysis of the homework completion sheets indicated that teachers, in turn, submitted, on average, 95 percent of the assigned homework. Although teachers did not get individual feedback on their homework, the beginning of each day was devoted to discussing homework answers and solution methods. Analyses of the video recordings indicated that, on average, Intel Math instructors and participants spent three hours discussing homework in total, providing additional opportunities for teachers to receive feedback on their solutions and explanations.[19]

Taken together, analyses of the features of Intel Math as implemented suggest that teachers had extensive opportunities to solve mathematics problems individually and in groups. Teachers also had opportunities to participate in mathematical reasoning and communication in rich, whole-group conversations. These are all activities that were hypothesized to deepen teachers' knowledge and support enactment (see Exhibit 1.1 in chapter I).

*Implementation of the Mathematics Learning Community.* Like Intel Math, the Mathematics Learning Community meetings were implemented with high fidelity. Analyses of the facilitator logs and the video recordings collected for each session indicated that on average, 100 percent of the planned Mathematics Learning Community hours were delivered by the study districts (mean: 10.0 hours, range: 9.9–10.1 hours). Facilitators in five districts delivered all four parts of each of the five meetings, and in the sixth district, the

[17] We examined only the control teachers' MQI scores for this purpose because treatment teachers' scores could have been affected by the study PD. Control teachers' scores should reflect typical grade 4 teachers' MQI scores. We excluded segments of teachers' classroom lessons that were identified as comprising small group work or "other" in order to isolate those segments in which public discussion of math occurred, as these were the portions of Intel Math that were coded with the MQI. The teachers' MQI scores used for this analysis were the holistic scores for each overall dimension (*Richness, Student Participation,* and *Errors*) assigned by raters to each 7.5-minute segment of the lesson, paralleling the method used to code PD segments. These holistic dimension-level scores are different from the scores for the individual elements within dimensions used for the impact analysis of classroom practice (see chapter II and Appendix A).

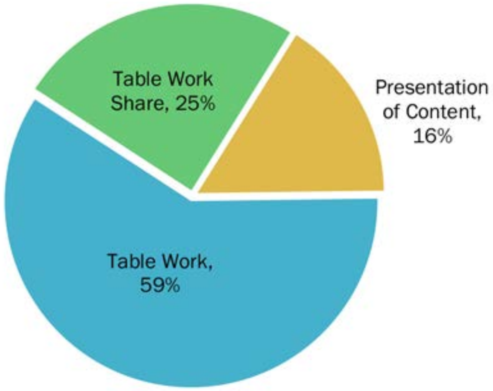
[18] This comparison is limited by a number of factors, including the different content in Intel Math and the grade 4 control teachers' classroom lessons (although the math topics did overlap) and the different audience (adult teachers for Intel Math and grade 4 students for the classroom lessons). Given these differences in content and audience, the comparison may be interpreted as a lower bound test of the mathematical quality of instruction in Intel Math.

[19] In one district, video recordings were not completed for homework discussions. We removed this district to calculate the average.

facilitators delivered all but one part of one meeting.[20] Analyses of the video recordings indicated that, as planned, on average, close to 40 percent (38 percent) of the meeting time was spent on parts one and two, which focused on math content, and a little more than 60 percent (62 percent) on parts three and four, which focused on student approaches to mathematics and classroom connections.[21]

Additional analyses of the video recordings indicated that, similar to Intel Math, the most prominent activity structure was table work, followed by time spent discussing the work completed during table work. As illustrated in Exhibit 3.8, 59 percent of the time spent in the Mathematics Learning Community was dedicated to table work, 25 percent to table work share, and 16 percent to presentation and discussion of new content. Further analyses indicated that more than three fourths (76 percent) of the time spent in table work was dedicated to analysis of student work and classroom connections.

**Exhibit 3.8. Mean Percentage of Time Spent in Presentation of Content, Table Work, and Table Work Share in the Mathematics Learning Community**



Note: Sample size = 6 districts. Percentage of time spent in these three activities is out of total time spent in Mathematics Learning Community parts; that is, daily evaluations and other administrative activities were excluded from the denominator.

Source: Mathematics Learning Community videos.

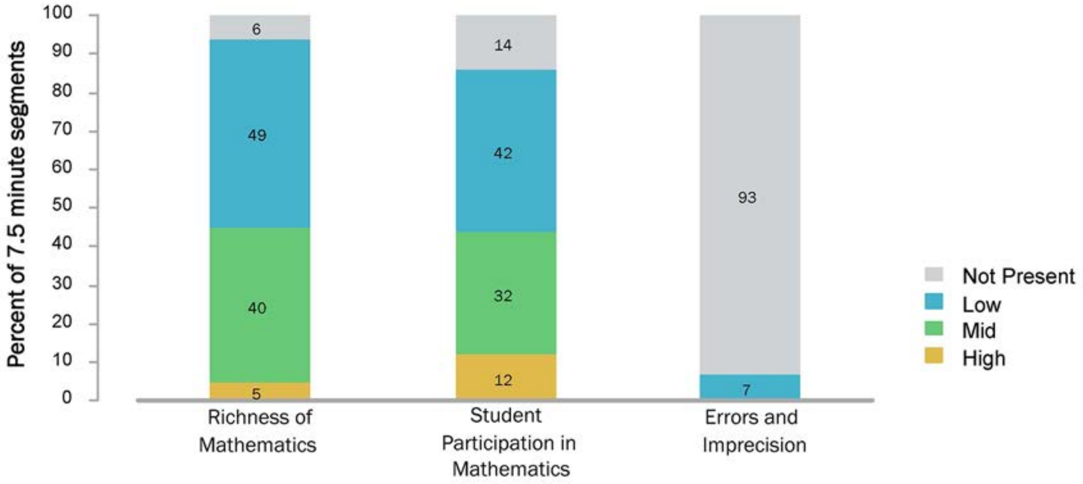
MQI analysis of the video recordings of whole-group mathematical conversations indicated that these conversations demonstrated rich mathematics, teacher participation in mathematical reasoning and communication, and few instances of uncorrected errors, lack of clarity, and imprecision.[22] The percentage of coded 7.5-minute segments that were rated not present, low, mid, and high for each MQI dimension is shown in Exhibit 3.9. Note that, as in the analysis of whole-group mathematical discussion in Intel Math, the students here are the participating teachers.

[20] Video for one Mathematics Learning Community meeting in one district was missing. All analyses involving video recordings, therefore, did not include that meeting.

[21] Analyses of the video recordings to determine fidelity and features of the Mathematics Learning Community excluded time spent completing and discussing meeting evaluations as well as time spent discussing study-related issues.

[22] MQI analyses of whole mathematical conversations focused on presentation of content and table work share within the math content-focused parts of each Mathematics Learning Community meeting (i.e., parts one and two).

**Exhibit 3.9. Percentage of Segments in the Mathematics Learning Community Rated Not Present, Low, Mid, and High on Each MQI Dimension**



Note: Sample size = 97 7.5-minute segments across 6 districts. The coded segments occurred during Presentation of Content and Table Work Share in the Mathematics Learning Community mathematics content-focused parts. Percentages shown are the percent of segments rated not present, low, mid, or high for each dimension (*Richness of Mathematics, Errors and Imprecision*, and *Student Participation*).

Source: Mathematics Learning Community videos.

As in Intel Math, the results shown in Exhibit 3.6 indicate that mathematical quality was evident much of the time during the Mathematics Learning Community meetings. Rich mathematics was evident at a mid or high level in 45 percent of the coded segments. Participants' engagement in mathematical reasoning and communication was evident at a mid or high level in 44 percent of the coded segments. Unresolved errors, instances of imprecision, and lack of clarity were not present in 93 percent of the coded segments.

Comparison with the MQI data from whole-group discussion in control teachers' classroom lessons suggests that the Mathematics Learning Community, as delivered in the study, reflected a greater degree of *Richness of Mathematics* and *Student Participation of Mathematics* than may be typical in grade 4 teachers' classroom lessons. *Errors and Imprecision* were rarer in the Mathematics Learning Community meetings than in typical grade 4 teachers' classroom lessons.[23]

Taken together, the analyses of the features of the Mathematics Learning Community meetings suggest that, similar to Intel Math, this component of the PD program provided teachers extended opportunities to work in groups and to share their thinking in whole-group discussions that had evidence of richness. However, unlike Intel Math, in which small-group work focused mostly on mathematics content, more of the time in small-group work in the Mathematics Learning Community was dedicated to analysis of student thinking and classroom connections. In so doing, the Mathematics Learning Community provided teachers with

[23] Again, these results may be expected given that the coded segments of the Mathematics Learning Community sessions reflected discussions among adults, and teachers' grade 4 lessons reflect instruction delivered to children. This comparison may thus be interpreted as a lower bound test of the mathematical quality of instruction in the Mathematics Learning Community.

opportunities to participate in the activities that were hypothesized to support enactment, as well as enhance their knowledge, consistent with the conceptual framework presented in Exhibit 1.1 in chapter I.

*Implementation of the Video Feedback Cycles.* We found that the Video Feedback Cycles were also implemented with high fidelity. Analyses of facilitator logs and feedback forms indicated that all three feedback cycles were implemented as expected with treatment teachers, with each cycle having close to 100 percent of the required characteristics. Nearly all (97 percent) of the feedback sessions provided feedback on lessons that covered content addressed in Intel Math and a prior Math Learning Community session. All of the feedback forms were completed as intended. All feedback forms included descriptions of strengths and weaknesses for the lesson overall and for three to four video clips. Each clip illustrated no more than three MQI codes, with suggestions for improvement as well as connections to Intel Math and the Mathematics Learning Community. All feedback forms also identified next steps for the teacher. Facilitators reported reviewing each of the clips in 96 percent of the feedback sessions and discussing next steps in all of them.

Analyses of the features of the Video Feedback Cycles indicated that, while the Video Feedback Cycles provided feedback to teachers on both the *Richness of Mathematics* and *Errors and Imprecision* elements, most of the feedback was on elements associated with *Richness of Mathematics*, with particular emphasis on Linking and Connections, Mathematical Meaning and Sense-Making, and Remediation of Student Errors and Difficulties. Exhibit 3.10 shows the frequency with which the feedback addressed each element and overall.

**Exhibit 3.10. Prevalence of MQI Elements on Teachers' Feedback Forms, Across Three Video Feedback Cycles**

| MQI Dimensions and Elements | Percent of All Elements Identified on Feedback Forms | |
|---|---|---|
| Richness of Mathematics | 82% | |
| Linking and Connections | | 18% |
| Explanations | | 13% |
| Multiple Procedures or Solution Methods | | 9% |
| Mathematical Meaning and Sense-Making | | 17% |
| Mathematical Language | | 7% |
| Remediation of Student Errors and Difficulties | | 18% |
| Errors and Imprecision | 18% | |
| Major Mathematical Errors | | 3% |
| Imprecision in Language and Notation | | 9% |
| Lack of Clarity in Presentation of Mathematical Content | | 6% |

Note: Across the three cycles, raters and facilitators completed 237 feedback forms for treatment teachers. These forms identified a total of 1,390 elements. The exhibit shows how the 1,390 elements identified on the forms were distributed among the *Richness of Mathematics* and *Errors and Imprecision* dimensions. The percentages within each cycle were similar to those shown across cycles.

Source: Feedback forms.

Together, these analyses indicate that teachers received feedback on their mathematics instruction, of the kind hypothesized to support the enactment of knowledge in the classroom (see conceptual framework in Exhibit 1.1 in chapter I).

### Treatment teacher attendance at the PD was high, and the contrast between treatment and control teachers' math-related PD participation was considerable

Analyses of log data and teacher attendance forms indicate that treatment teachers' rates of participation in all three components of the PD program were high. On average, treatment teachers participated in more than 90 percent of the implemented hours for each component (Exhibit 3.11).

**Exhibit 3.11. Number of Hours of PD Intended, Implemented, and Attended, and Percentage of Implemented Hours Received by the Average Treatment Teacher**

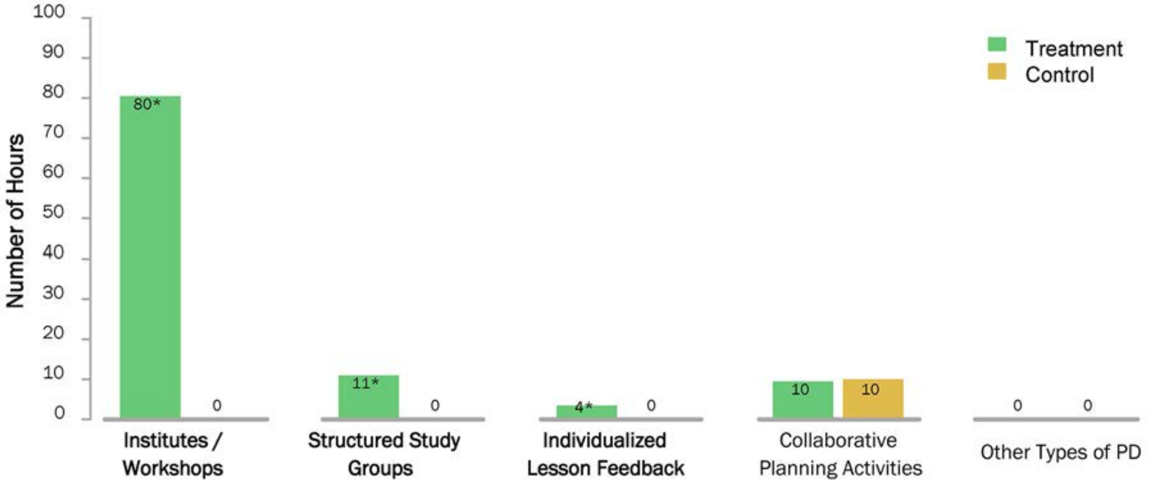
| PD Component | Hours Intended | Mean Hours Implemented | Mean Hours Attended | Percentage of Implemented Hours Received |
|---|---|---|---|---|
| Intel Math | 80.0 | 76.9 | 75.2 | 97.8 |
| Mathematics Learning Community | 10.0 | 10.1 | 9.0 | 90.0 |
| Video Feedback Cycles | 3.0 | 3.0 | 2.9 | 96.7 |
| Total | 93.0 | 89.9 | 87.1 | 96.9 |

Source: Intel Math fidelity forms; Mathematics Learning Community and Video Feedback Cycle facilitator logs; Intel Math and Mathematics Learning Community attendance sheets.

Analyses of spring 2014 survey data indicate statistically significant differences between treatment and control teachers in their reported participation in the types of math PD that were offered as part of the study intervention (institutes/workshops, structured study groups, and individualized lesson feedback) but not in other types of PD. Based on data from the study-administered teacher survey, close to 100 percent of the treatment teachers reported having participated in all three of these types of math PD over the summer of 2013 and during the 2013–14 school year, versus 14 percent to 41 percent of control teachers, for a participation rate difference of 56 to 81 percentage points.

Teachers who reported participating in these types of math-related PD also reported the number of hours they spent in them. Treatment teachers participated in 80 more hours in institutes and workshops than control teachers, 11 more hours in structured study groups, and 4 more hours in individualized lesson feedback, as shown in Exhibit 3.12. Overall, treatment teachers participated in 95 more hours of math-related PD than did control teachers, which is close to the approximately 93 hours of math PD provided by the study. This finding indicates that as expected, the PD provided in this study was substantially more PD than these teachers would have typically received and that as intended, the study PD was layered on top of the business-as-usual PD.

**Exhibit 3.12. Median Number of Hours of Mathematics-Related PD in Which Teachers Participated During Summer 2013 and the 2013–14 School Year**



Note: Sample size = 73 schools; 79 treatment teachers and 86 control teachers.

* Difference between the median treatment teacher hours and the median control teacher hours is statistically significant at the 0.05 level, two-tailed test.

Source: Spring 2014 Teacher Survey.

Apart from these differences in amount of PD, the survey data also indicate that the PD in which treatment teachers participated differed in features from the PD attended by control teachers. As expected, treatment teachers reported a greater focus on K–8 math content and student thinking activities during their workshop, study group, and feedback-related PD than did control teachers who reported participating in these types of math-related PD. Readers who are interested in seeing more details on the results presented in this section should consult Appendix C.

Overall, treatment teachers had high rates of participation in the PD program, and there appeared to be a strong contrast in the amount and type of math-related PD experienced by treatment and control teachers. The next chapter presents results of analyses testing the PD's impact on teacher and student outcomes.

# IV. Impact of the Professional Development Program

This chapter presents findings about the impact of the study professional development (PD) program on teacher knowledge, instructional practice, and student achievement. As described in chapter II, the teacher outcomes (knowledge and instructional practice) were measured in both the fall (after Intel Math) and spring (after the full PD program); results reported in this chapter examine impacts of the PD on these outcomes in the fall and spring separately. Both of the student achievement outcomes were measured in spring 2014, after the full PD program was completed.

## The PD had a positive impact on teacher knowledge

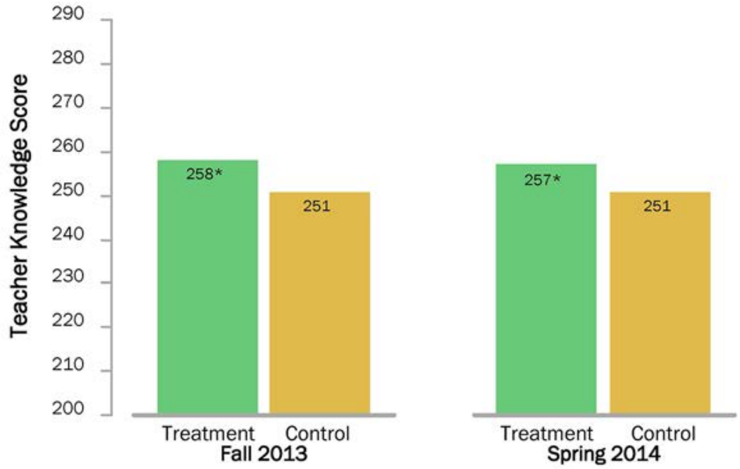
As hypothesized, the PD program had a substantial positive impact on teacher knowledge in the fall, which was largely sustained through spring.[24] The average knowledge score in the fall was 258 points on the NWEA RIT (**R**asch Un**it**) scale for treatment teachers, compared to 251 for control teachers (see Exhibit 4.1). The 7-point difference in the fall corresponds to an effect size of 0.63 and an improvement index of 24 percentile points, implying that the percentile rank of the average control teacher would increase from the 50th percentile to the 74th percentile had the teacher received the treatment.[25] The 6-point difference in the spring corresponds to an effect size of 0.55 and an improvement index of 21 percentile points (from the 50th to 71st percentile).[26]

[24] These results were not sensitive to sample definition or the inclusion of covariates. (See Exhibit D.1 in Appendix D for the main impact analyses and sensitivity analyses results.)

[25] The "improvement index" refers to the expected change in the percentile rank of an average control teacher had the teacher received the treatment, based on the outcome distribution within the control group (What Works Clearinghouse, 2014).

[26] As another way to understand the size of these impacts, the fall effect size corresponds to a 10-percentile-point difference based on the NWEA spring norms for 11th graders (the oldest students in the norming sample), with treatment teachers scoring at the 84th percentile and control teachers at the 74th percentile, on average. The spring effect size corresponds to a 8-percentile-point difference based on the NWEA spring norms for 11th graders (82nd percentile for treatment teachers versus 74th percentile for control teachers).

**Exhibit 4.1. Average Teacher Knowledge Scores in Fall 2013 and Spring 2014**



Note: Sample size = 73 schools; 79 treatment teachers and 86 control teachers.

The teacher knowledge score is reported in the scale used by NWEA, which takes into account the difficulty of individual test items in measuring teacher knowledge. The assessment is not typically given to adults; 11th graders are the oldest students for whom norming data are available. The scale shown ranges from 200, the score that corresponds approximately to the 1st percentile for 11th graders, to 290, the score that corresponds approximately to the 99th percentile for 11th graders. In the fall, the average scores correspond to the 84th percentile for treatment teachers and 74th percentile for control teachers. In the spring, the average scores correspond to the 82nd percentile for treatment teachers and 74th percentile for control teachers.

The analyses are based on a teacher-level regression controlling for school fixed effects and teacher background characteristics. The means for the treatment group are unadjusted. The means for the control group were computed based on the unadjusted treatment group means and estimated impact.

* Difference between the average treatment teacher score and the average control teacher score is statistically significant at the 0.05 level, two-tailed test.

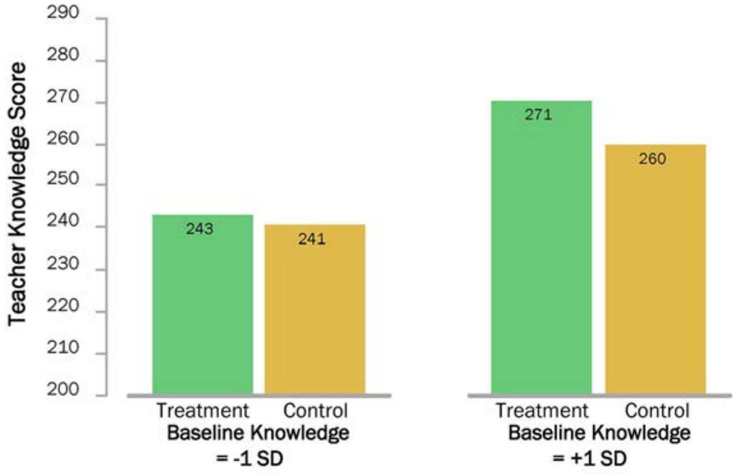
Source: Fall 2013 and Spring 2014 Teacher Knowledge Tests.

Although the PD had a positive impact on teachers' knowledge overall, it is possible that the PD was mathematically too easy or too difficult for some teachers—in which case, it might have had a smaller impact on content knowledge for some teachers than others. To examine whether this was the case, we assessed whether the impact of the PD differed for teachers with different levels of prior math knowledge, as measured by the study's baseline teacher knowledge test. We found that the impact of the PD on teachers' knowledge in the fall did not differ for teachers with different levels of prior knowledge.[27] However, the impact of the PD on teachers' knowledge in the spring was statistically significantly larger for teachers with higher baseline knowledge scores than for teachers with lower scores (see Exhibit 4.2, which presents the predicted impact for teachers with a baseline knowledge score of +/− 1 standard deviation).[28]

[27] The estimated impact of the PD on fall teacher knowledge was positive and statistically significant for teachers with a baseline knowledge score 1 standard deviation above average (improvement index of 25 percentile points), as well as for teachers 1 standard deviation below average (improvement index of 23 percentile points).

[28] The estimated impact of the PD on spring teacher knowledge was positive and statistically significant for teachers with a baseline knowledge score 1 standard deviation above average (improvement index of 34 percentile points), but was not statistically significant for teachers 1 standard deviation below average (improvement index of 8 percentile points). Teachers' baseline knowledge scores were standardized using the control group mean and standard deviation within the teacher analysis sample. See Exhibit D.4 in Appendix D for results of analyses testing for differential impact by teachers' baseline knowledge.

**Exhibit 4.2. Average Teacher Knowledge Score in Spring 2014 for Teachers With a Baseline Knowledge Score of +/− 1 Standard Deviation**



Note: Sample size = 73 schools, 79 treatment teachers, and 86 control teachers.

The teacher knowledge score is reported in the scale used by NWEA, which takes into account the difficulty of individual test items in measuring teacher knowledge. The assessment is not typically given to adults; 11th graders are the oldest students for whom norming data are available. The scale shown ranges from 200, the score that corresponds approximately to the 1st percentile for 11th graders, to 290, the score that corresponds approximately to the 99th percentile for 11th graders.

The analyses are based on a teacher-level regression controlling for school fixed effects and teacher background characteristics. The means for the treatment group are unadjusted. The means for the control group were computed based on the unadjusted treatment group means and estimated impact.

Source: Baseline and Spring 2014 Teacher Knowledge Tests.

### The PD had a positive impact on some dimensions of instructional practice, particularly *Richness of Mathematics*

This section reports the results of analyses comparing treatment and control teachers' scores on the three Mathematical Quality of Instruction (MQI) dimensions of instructional practice. As noted previously, for two of the MQI dimensions—*Richness of Mathematics* and *Student Participation in Mathematics*—higher scores indicate better practice than lower scores. Therefore, on these dimensions, the hypothesized impact of the study PD is positive (i.e., treatment teachers were expected to score higher than control teachers). On the third MQI dimension—*Errors and Imprecision*—lower scores indicate fewer content errors and less imprecision than higher scores. Therefore, on this dimension, the hypothesized impact of the study PD is negative (i.e., treatment teachers were expected to score lower than control teachers).

The impact of the PD on the *Richness of Mathematics* dimension was positive but not statistically significant in the fall, and positive and statistically significant in the spring. An average treatment teacher in the spring demonstrated *Richness of Mathematics* at a mid or high level during 63 percent of a typical lesson, compared to 46 percent for an average control teacher (see Exhibit 4.3).[29] This difference of 17 percentage points corresponds to an effect size of 0.61 (an improvement index of 23 percentile points).

Exhibit 4.3 also shows the impact of the PD on the *Student Participation in Mathematics* dimension, which was positive in both fall and spring but statistically significant only in the fall.[30] In the fall, an average treatment teacher's instruction demonstrated *Student Participation in Mathematics* at a mid or high level during 33 percent of a typical lesson, compared to 23 percent for an average control teacher. This difference of 10 percentage points corresponds to an effect size of 0.29 and an improvement index of 11 percentile points.

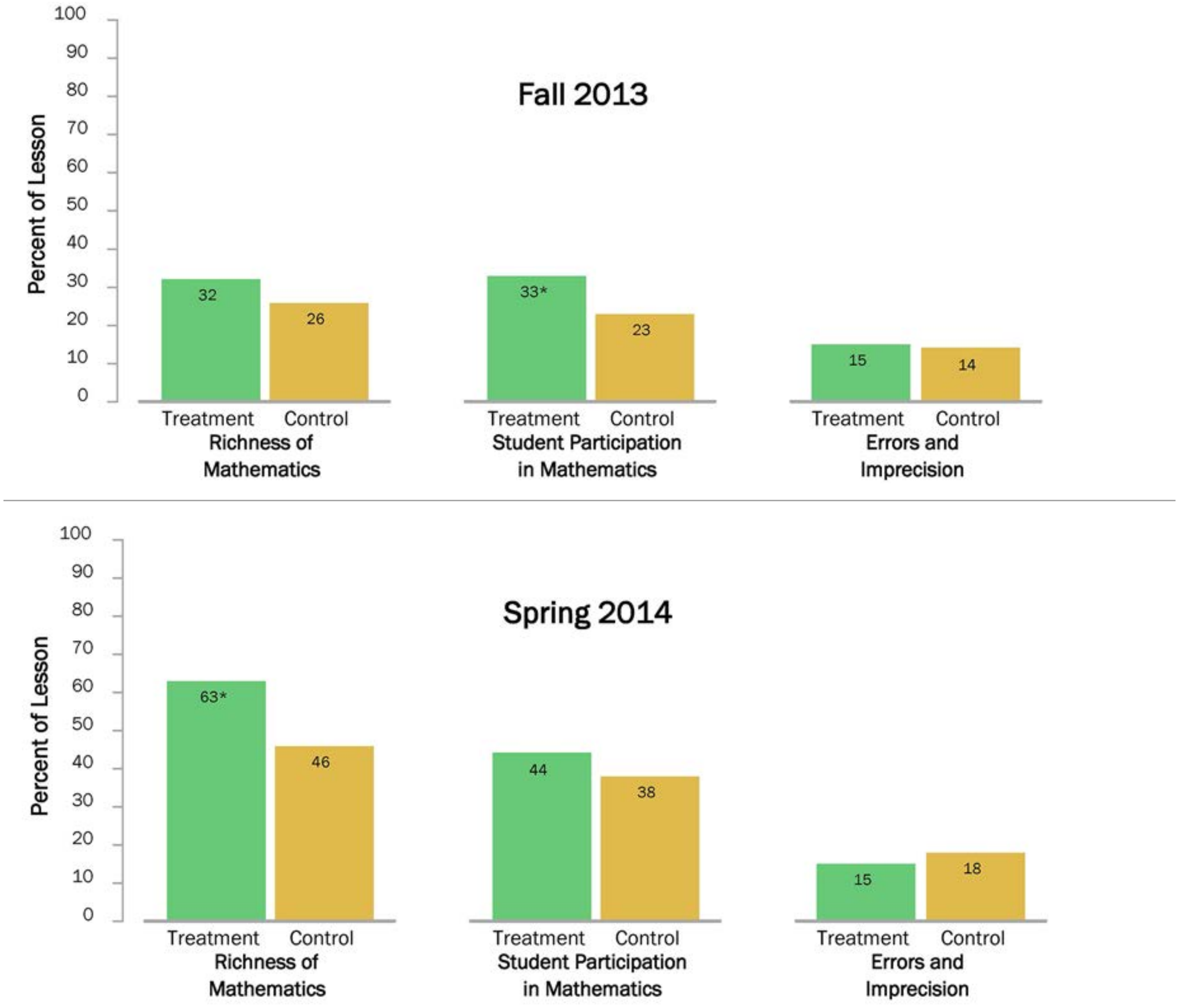
The PD did not have a statistically significant impact on the third MQI dimension—*Errors and Imprecision*—in either the fall or the spring (see Exhibit 4.3).[31]

[29] As described in chapter II, each teacher's instruction was coded based on seven elements of *Richness of Mathematics*. We used Rasch scaling to generate a *Richness* dimension score for each segment, based on the scores for the seven elements. We then used the Rasch scores to determine the probability that a teacher demonstrated Richness during a typical segment. We defined demonstrating *Richness* as scoring mid or high on one or more elements of *Richness* during the segment. A typical lesson lasted about 60 minutes, or eight segments. If, for example, an average teacher had a 25 percent chance of demonstrating *Richness* during a typical segment, the teacher would have been expected to demonstrate *Richness* in two segments over the course of a typical lesson, or 25 percent of the lesson.

[30] The fall impact for the *Student Participation in Mathematics* dimension was sensitive to sample definition and the inclusion of covariates. The other findings about the impacts of the PD on classroom practice dimensions were robust to alternative sample definition and covariate adjustment. (Exhibits D.2 and D.3 in Appendix D show results for the main impact and sensitivity analyses from fall 2013 and spring 2014.)

[31] In the spring, an average treatment teacher demonstrated *Errors and Imprecision* at a low, mid, or high level during 15 percent of a typical lesson, compared to 18 percent for an average control teacher. This difference of 3 percentage points (corresponding to an effect size of −0.22 and an improvement index of −9 percentile points) was not statistically significant.

**Exhibit 4.3. Percentage of an Average Teacher's Lesson Demonstrating Three Dimensions of Mathematical Quality of Instruction in Fall 2013 and Spring 2014**



Note: Sample size for fall 2013 = 73 schools; 79 teachers, 79 lessons, and 708 7.5-minute segments for the treatment group; 86 teachers, 86 lessons, and 739 7.5-minute segments for the control group. Sample size for spring 2014 = 73 schools; 79 teachers, 158 lessons, and 1,277 7.5-minute segments for the treatment group; 86 teachers, 172 lessons, and 1,352 7.5-minute segments for the control group.

The graph shows the percent of a typical lesson in which an average treatment or control teacher demonstrated each of the three MQI dimensions of instructional quality. Demonstrating *Richness* or *Student Participation* is defined as scoring mid or high on one or more of the elements that comprise the dimension. Demonstrating *Errors* is defined as scoring present (low, mid, or high) on one or more of the elements that comprise the dimension. Scores of low, mid, and high reflect the intensity and quality of the element of practice observed. *Richness of Mathematics* emphasizes the conceptual aspects of math, such as the use and quality of mathematical explanations; *Student Participation in Mathematics* focuses on student mathematical contributions, explanations, and reasoning; and *Errors and Imprecision* focuses on incorrect, unclear, and imprecise use of math. Lower error and imprecision scores are desirable and indicate fewer content errors and less imprecision than higher scores.

The fall *Richness of Mathematics* and *Student Participation in Mathematics* analyses are based on two-level models (segments within teachers), and the spring *Richness of Mathematics* and *Student Participation in Mathematics* analyses are based on three-level models (segments within lessons within teachers), controlling for school fixed effects and covariates at the segment, lesson, and teacher levels. The fall *Errors and Imprecision* analysis is based on a teacher-level regression, and the spring *Errors and Imprecision* analysis is based on a two-level model (lessons within teachers), controlling for school fixed effects and covariates at the lesson and teacher levels.

* Difference between the average treatment teacher percentage and the average control teacher percentage is statistically significant at the 0.05 level, two-tailed test.

Source: MQI scores of video-recorded lessons from the 2013–14 school year.

As we did for teacher knowledge outcomes, we tested whether the impact of the PD on classroom practice differed for teachers with different levels of prior knowledge, on the theory that teachers with higher levels of knowledge might be better able to translate their knowledge into practice. We found that the impact of the PD on MQI scores did not vary by teachers' prior math knowledge (see Exhibit D.4 in Appendix D).

We also tested whether the impact of the PD on instructional practice differed for teachers teaching classes with different levels of average prior achievement, on the theory that teachers might more easily or more effectively find opportunities to demonstrate teaching behaviors reflected in the MQI dimensions of practice in classes with relatively lower achieving or relatively higher achieving students. In the fall, the impact of the PD on teachers' scores on the *Richness of Mathematics* and *Student Participation in Mathematics* dimensions of classroom practice was larger in classrooms with lower achieving students than in classrooms with higher achieving students. The impact of the PD on *Richness of Mathematics* corresponds to an improvement index of 43 (effect size = 1.44) for teachers whose classroom average prior achievement scores were 1 standard deviation below the state average, compared with an improvement index of −31 (effect size = −0.89) for teachers whose classroom average prior achievement scores were 1 standard deviation above the state average. For the *Student Participation in Mathematics* dimension, the improvement index was 35 for teachers with low classroom average prior achievement (effect size = 1.04), and it was −18 for teachers with high classroom average prior achievement (effect size = −0.46). Findings for these two dimensions in the spring show a similar pattern, although the differential impacts were not statistically significant (see Exhibit D.5 in Appendix D). The differential impact on *Errors and Imprecision* by classroom average prior achievement was not statistically significant in either the fall or the spring.

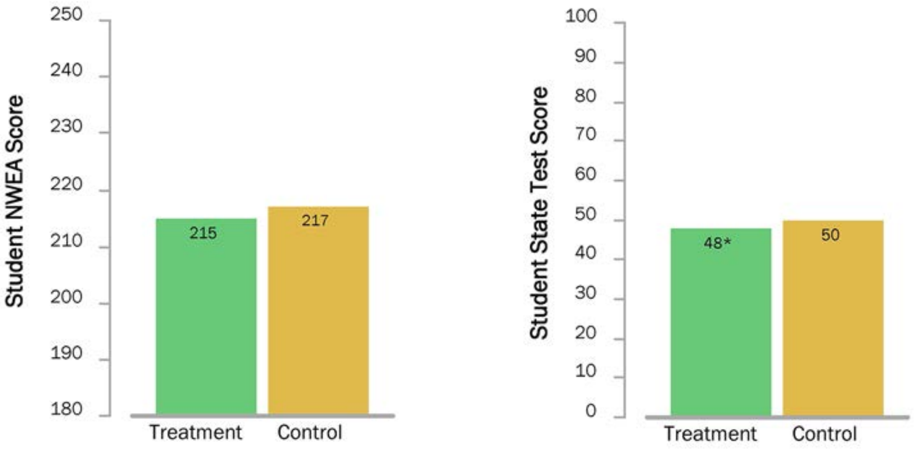
### The PD had no positive impacts on student achievement

The PD program did not have a positive impact on the two measures of student achievement in the spring. On the NWEA test, students in the treatment group scored 215 points on the RIT (**R**asch Un**it**) scale, compared with 217 for control students (see Exhibit 4.4). The difference, corresponding to an effect size of −0.05 and an improvement index of −2 percentile points, was not statistically significant.[32]

On the state assessment, treatment students scored 48 points on a normal curve equivalent scale, which can be interpreted as an approximate state percentile, compared to 50 for control students (see Exhibit 4.4). This difference, corresponding to an effect size of −0.06 and an improvement index of −2 percentile points, was statistically significant. However, this impact was sensitive to sample definition and the inclusion of covariates—it was not statistically significant based on any of the three sensitivity analyses we conducted. (See Exhibit D.6 in Appendix D for sensitivity analyses results, and Exhibits D.7 and D.8 for impact estimates for the two achievement measures for individual study districts.)

[32] This finding was not sensitive to sample definition or the inclusion of covariates. (See Exhibit D.6 in Appendix D for sensitivity analyses results.) As another way to understand the size of this impact, the effect size corresponds to a 5-percentile-point difference based on the NWEA spring norms for 4th graders, with treatment students scoring at the 58th percentile and control students at the 63rd percentile, on average.

**Exhibit 4.4. Average Student Mathematics Achievement in Spring 2014**



Note: Sample size for analysis of NWEA scores = 73 schools; 79 teachers and 806 students in the treatment group; 86 teachers and 891 students in the control group. Sample size for analysis of state test scores = 73 schools; 79 teachers and 1,760 students in the treatment group; 86 teachers and 1,917 students in the control group.

The student NWEA score is reported on the scale used by NWEA, which takes into account the difficulty of individual test items in measuring student achievement. The scale shown ranges from 180, the score that corresponds approximately to the 1st percentile for fourth graders, to 250, the score that corresponds approximately to the 99th percentile for fourth graders.

The state score is reported using Normal Curve Equivalent (NCE) scores. NCE scores measure a student's position on the normal curve, relative to other students in their state. NCE values run from 0 to 100. They are similar to percentile ranks, but on an equal-interval scale.

The analyses are based on a two-level model controlling for school fixed effects and student characteristics. The means for the treatment group are unadjusted. The means for the control group were computed based on the unadjusted treatment group means and estimated group differences.

* Difference between the average treatment student score and the average control student score is statistically significant at the 0.05 level, two-tailed test.

Source: Spring 2014 NWEA Test; District administrative records.

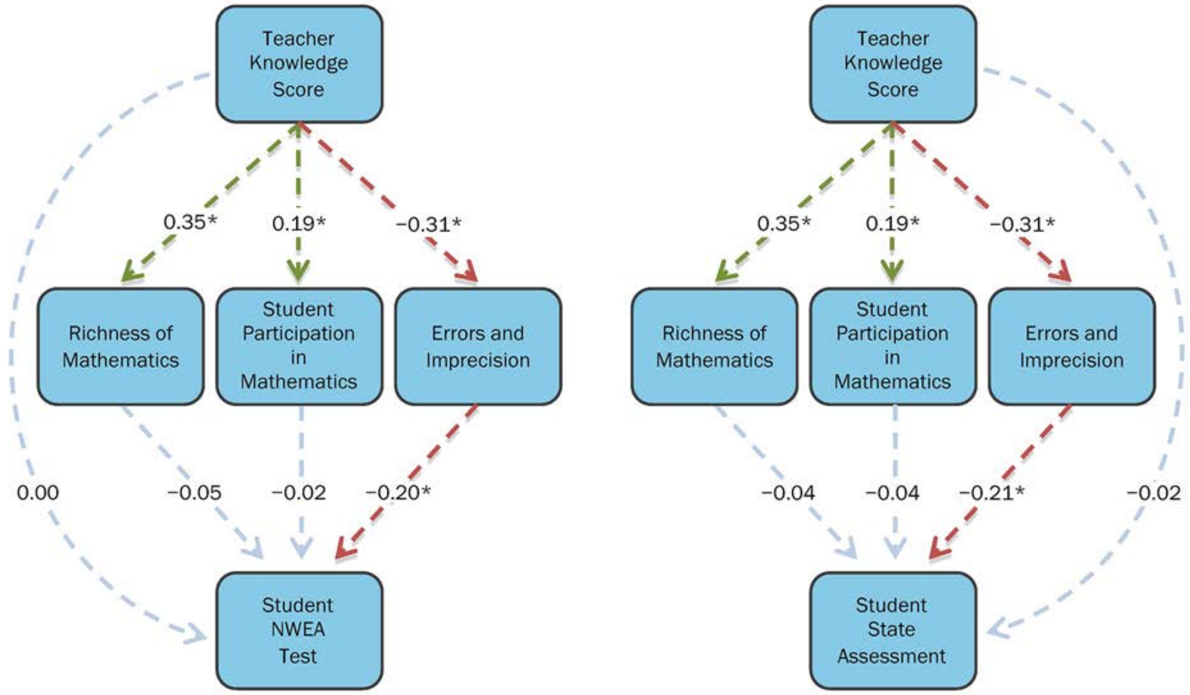
We also examined whether the average impact of the PD on student achievement masked differential impacts for teachers or classrooms with different characteristics. For example, we examined whether the effect of the PD on achievement differed for students whose teachers were higher or lower in baseline math knowledge, on the theory that teachers with higher levels of baseline knowledge might be better able to translate what they learned in the PD into improvements in student learning. We found no differential impact. We conducted similar analyses for teachers with more or less teaching experience, on the theory that less experienced teachers might have more need for the PD, or be more willing to profit from it, but again there was no differential effect. We also examined whether the impact of the PD on student achievement might differ in impact in classrooms with higher versus lower average prior student achievement or for students with higher or lower prior achievement, on the theory that the students with lower prior achievement might be more sensitive to improvements in instruction. Here too we found no differential impact of the PD. (See Exhibit D.9 in Appendix D for full results of these analyses.)

# V. Understanding the Impacts

Although the PD we tested in this study had an impact on teacher knowledge and some aspects of classroom practice, it did not translate into a positive impact on student achievement. These results are generally consistent with the few other recent larger-scale rigorous studies of content-focused math PD, including PD delivered by established providers on a wide scale, such as Pearson Achievement Solutions and America's Choice (Garet et al., 2011), Developing Mathematical Ideas (Hammerman, Demers, & Higgins, 2015), and Math Solutions (Jacob, Hill, & Corey, 2015).[33]

Why did the PD tested in this study not have an impact on student achievement—especially given its relatively large impact on teacher knowledge and classroom practice? To investigate this question, we assessed whether knowledge, classroom practice, and achievement were related to one another as hypothesized in the conceptual framework discussed in chapter I. First, we predicted classroom practice based on teacher knowledge. Second, we predicted students' achievement based on teacher knowledge and the three classroom practice dimensions measured in the study, controlling for students' prior achievement. The estimates, shown in Exhibit 5.1, were based on hierarchical linear models conducted separately for each association (see Appendix A for more details on the estimation procedures). Note that these analyses are correlational and do not provide causal impacts of teacher knowledge and practice on achievement.

**Exhibit 5.1. Associations Among Teacher Knowledge, Three Dimensions of Instructional Practice, and Student Achievement**



[33] To identify content-focused math PD studies, we examined experimental studies in Yoon et al. (2007), Gersten et al. (2014), and those that were subsequently published. The few studies that found a positive impact on student achievement tended to be smaller-scale efficacy trials, including studies of Cognitively Guided Instruction (Carpenter et al., 1989; Jacobs et al., 2007) and lesson study (Perry & Lewis, 2011).

Note: Sample size = 73 schools; 79 treatment teachers and 86 control teachers.

The model was estimated based on data for treatment and control teachers. All models included the covariates that were included in the main impact models, including school fixed effects. Teacher and student outcome variables of interest were standardized based on the control-group mean and standard deviation. For more detail on the models estimated, see Appendix A. In addition to the models shown, we conducted sensitivity analyses for the model predicting achievement, including classroom practice but not teacher knowledge, and including the three dimensions of knowledge one at a time. The results, shown in Appendix E, are similar.

* Association is statistically significant at the 0.05 level.

Source: Fall 2013 Teacher Knowledge Tests; MQI scores of video-recorded lessons from Spring 2013; district administrative records; and Spring 2014 NWEA Test.

## Teachers' content knowledge was related to their instructional practice, as expected

The estimates in Exhibit 5.1 show that teacher knowledge as measured in the study was associated with all three Mathematical Quality of Instruction (MQI) dimensions. For example, a teacher 1 standard deviation above average in teacher knowledge was predicted to be 0.35 standard deviations above average in *Richness of Mathematics*, 0.19 standard deviations above average in *Student Participation in Mathematics*, and 0.31 standard deviations below average in *Errors and Imprecision*. These estimates of association are consistent with the link between teachers' content knowledge and their instructional practice hypothesized by the study's conceptual framework. In particular, the conceptual framework assumes that the primary route through which teachers' content knowledge affects student achievement is through their math-related instructional practices, as reflected in the MQI dimensions.

## Teachers' content knowledge and instructional practice were generally unrelated to student achievement

Teacher knowledge was not associated with student achievement, and of the three dimensions of practice, only *Errors and Imprecision* was associated with student achievement. For example, a teacher who was a standard deviation above average in *Errors and Imprecision* was predicted to have students 0.21 standard deviations below average in achievement on the state test, and 0.20 standard deviations below average on the study-administered NWEA assessment.

Other studies also have estimated the relationship between various measures of teachers' math content knowledge, math instructional practice, and student achievement, with mixed results. Some studies have obtained results similar to ours, with generally no statistically significant relationships between knowledge and achievement or between practice and achievement. Others have found statistically significant positive relationships.

For example, we found that the relationship between teacher knowledge and student achievement was not statistically significant (0.00 for the NWEA assessment and −0.02 for the state assessment). Estimates from other studies range from a not statistically significant 0.02 (Rockoff, Jacob, Kane, & Staiger, 2011) to a statistically significant 0.05 (Hill, Rowan, & Ball, 2005). We also found that the relationship between two of the three MQI dimensions and student achievement was not statistically significant (−0.05 for *Richness of Mathematics* and −0.02 for *Student Participation in Mathematics* for the NWEA assessment, and −0.04 for both dimensions for the state assessment). The relationship between the third MQI dimension, *Errors and Imprecision*, and achievement was statistically significant (−0.20 for the NWEA and −0.21 for the state assessment). In comparison, estimates from the Measures of Effective Teaching study (Kane & Staiger, 2012) obtained an estimate of the relationship between the overall score on an abbreviated version of the

MQI and student achievement of 0.02 (statistical significance was not reported in the study). More recently, Blazar (2015), using a version of the MQI similar to the one used in the present study, found that the relationship between student achievement and a measure combining *Richness of Mathematics* and *Student Participation in Mathematics* ranged from a not statistically significant 0.06 to a statistically significant 0.11 depending on the analysis model. The relationship between student achievement and *Errors and Imprecision* was not statistically significant, ranging from −0.03 to −0.05.[34]

While these various estimates are mixed, they provide evidence that *some* aspects of teachers' mathematical knowledge and instructional practice may be related to student achievement. However, these estimates imply that any correlations with achievement are at best modest. PD would thus likely need to have a relatively large impact on those aspects of knowledge and practice to produce a measureable impact on achievement. For example, the Hill and colleagues estimate implies that a PD intervention targeting knowledge alone would need to obtain an impact of roughly 2 standard deviations on knowledge to obtain an impact of 0.1 standard deviations on student achievement. Similarly, the most optimistic Blazar estimate implies that the impact on teacher practice would need to be roughly 1 standard deviation to improve student achievement by 0.1 standard deviations. To summarize, future research might focus on identifying PD that will impact these knowledge and practice outcomes to a larger degree. Future research might also seek to identify other aspects of knowledge and practice to target with PD that are more strongly related to improved student achievement.

[34] Results reported in different studies differ in part depending on how the association was computed and what other variables were controlled. In addition, some studies report the association between classroom practice measures and teacher value-added scores; others report the association with student achievement. The results in the text were all scaled in terms of the association with achievement and are thus comparable to each other. The association with value-added scores can be converted to the association with achievement by multiplying by roughly 0.2, the approximate standard deviation of teacher value-added scores in student standard deviation units. All estimates of association can be interpreted as the expected change in standard deviations of student achievement if teacher knowledge or practice improved by 1 standard deviation.

This page has been left blank for double-sided copying.

# Appendix A. Samples, Measures, and Analyses

This appendix provides additional technical information about the study to supplement the information in chapter II. The first section provides details on the study samples, and the second discusses the power analyses. The third section addresses the key measures, and the fourth section presents the statistical models for the main impact and correlational analyses.

## Study samples

This section presents information about the teacher and student samples.

*Teacher sample.* At the time of random assignment, study schools had two, three, or four volunteer teachers participating in the study. Exhibit A.1 presents the distribution of schools and teachers according to the number of participating teachers within a school when random assignment took place.

**Exhibit A.1. Number of Schools and Teachers**

| Number of Teachers in School | Schools in Study Sample | Teachers in Treatment Condition | Teachers in Control Condition |
|---|---|---|---|
| Schools with 2 teachers | 68 | 68 | 68 |
| Schools with 3 teachers | 19 | 22 | 35 |
| Schools with 4 teachers | 7 | 14 | 14 |
| All schools | 94 | 104 | 117 |

Note: When there were three volunteers in a school, the study typically assigned two teachers to control and one to treatment, to minimize the likelihood that treatment teachers would share what they learned from the PD with control teachers.

Source: Study records.

Exhibit A.2 compares the characteristics of the teachers in the study sample with grade 4 teachers in the national population of schools. The study teachers in Exhibit A.2 include those who were randomly assigned to condition, for whom we had baseline knowledge data or spring 2014 survey data, from which most of the background characteristics were collected. As shown in Exhibit A.2, teachers in the study sample generally had characteristics similar to those of teachers in the national population.

**Exhibit A.2. Background Characteristics of Grade 4 Teachers in the Study Sample and Grade 4 Teachers in the National Population**

| Characteristics | Grade 4 Study Teachers | National Population of Grade 4 Teachers | Difference | P value |
|---|---|---|---|---|
| Standard certification (percent) | 94.4 | 98.9 | −4.5* | 0.007 |
| Years of teaching experience (percent) | | | | |
| 3 years or fewer | 15.2 | 9.9 | 5.3 | 0.051 |
| 4–10 years | 35.9 | 32.5 | 3.4 | 0.386 |
| 11–20 years | 31.3 | 33.1 | −1.8 | 0.637 |
| More than 20 years | 17.7 | 24.5 | −6.8* | 0.025 |
| Master's degree or higher (percent) | 62.6 | 58.1 | 4.5 | 0.229 |

Note: Sample size = 184–198 grade 4 teachers in study sample and 3,300 grade 4 teachers in schools serving grade 4 students in the national population. There is a range for the study sample size due to item-level missing data on the study's spring 2014 teacher survey.

The SASS-based estimates for the national population of grade 4 teachers are approximate, as the teacher weights in SASS are not designed specifically to generate estimates representative of this population.

* Difference between grade 4 study teachers and the teachers in the national population is statistically significant at the 0.05 level, two-tailed test.

Source: Spring 2014 Teacher Survey; U.S. Department of Education, National Center for Education Statistics, *Schools and Staffing Survey* (SASS), 2011–12 school year.

We examined the baseline equivalence of teachers across treatment conditions based on the sample at random assignment using information collected during the administration of the baseline teacher knowledge test and the spring 2014 teacher survey. As shown in Exhibit A.3, no statistically significant differences by treatment status were detected using a two-tailed t-test.

**Exhibit A.3. Teacher Background Characteristics for All Teachers with Baseline Data**

| Characteristics | Treatment Group | Control Group | Estimated Difference | P value |
|---|---|---|---|---|
| Teacher knowledge at baseline (standardized)[a] | 252.3 | 254.6 | −2.3 | 0.172 |
| Standard certification (percent) | 96.7 | 90.6 | 6.1 | 0.085 |
| Years of teaching experience (percent) | | | | |
| 3 years or fewer | 15.6 | 14.8 | 0.7 | 0.889 |
| 4–10 years | 36.7 | 37.0 | −0.4 | 0.961 |
| 11–20 years | 33.3 | 31.5 | 1.8 | 0.800 |
| More than 20 years | 14.4 | 16.6 | −2.1 | 0.694 |
| Master's degree or higher (percent) | 65.6 | 62.7 | 2.9 | 0.686 |
| Calculus course (percent) | 27.9 | 22.2 | 5.7 | 0.389 |
| Number of mathematics courses | 4.2 | 3.7 | 0.5 | 0.249 |

Note: Sample size for teacher survey sample = 83 schools; 90 treatment teachers and 102 control teachers. Sample size for teacher knowledge sample = 89 schools; 97 treatment teachers and 112 control teachers.

[a] Teacher knowledge scores at baseline are based on a computer-adaptive math assessment provided by the Northwest Evaluation Association (NWEA), administered in summer 2014. The assessment is typically given to students, not adults. Grade 11 students are the oldest students for whom norming data are available; scores for 11th graders range from 197 (1st percentile) to 286 (99th percentile).

The analyses are based on a teacher-level regression controlling for school fixed effects. Treatment group means are unadjusted means. For continuous measures of teacher characteristics, the control group mean was computed by subtracting the estimated difference from the treatment group mean. Seven teachers were missing data on whether or not they took a calculus course (4 percent).

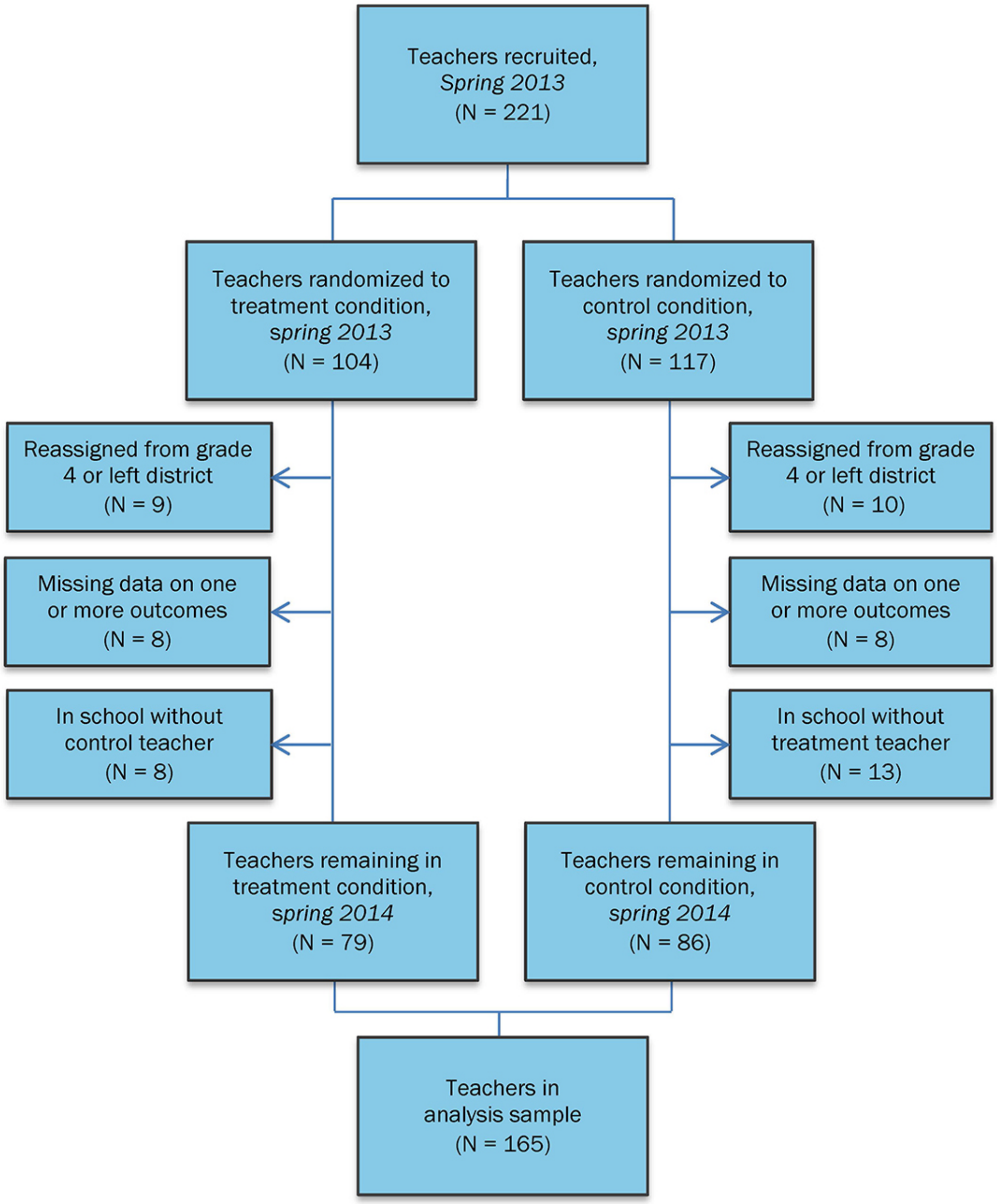
* Difference between treatment teachers and control teachers is statistically significant at the 0.05 level, two-tailed test.

A likelihood ratio test of overall baseline equivalence confirmed that there was not a statistically significant difference between the treatment and control groups across the full set of teacher baseline characteristics ($p = 0.17$).

Source: Baseline Teacher Knowledge Test; Spring 2014 Teacher Survey.

After random assignment, some teachers left the study for various reasons. Exhibit A.4 summarizes the number of teachers who left the sample, by treatment status. From the original sample of 221 teachers who participated in the random assignment in spring 2013 (104 treatment, 117 control), 165 teachers remained in the study's analytic sample (79 treatment, 86 control).

**Exhibit A.4. Teacher Sample Over the Course of the Evaluation**


Teachers recruited, *Spring 2013* (N = 221)
Teachers randomized to treatment condition, *spring 2013* (N = 104)
Teachers randomized to control condition, *spring 2013* (N = 117)
Reassigned from grade 4 or left district (N = 9)
Missing data on one or more outcomes (N = 8)
In school without control teacher (N = 8)
Reassigned from grade 4 or left district (N = 10)
Missing data on one or more outcomes (N = 8)
In school without treatment teacher (N = 13)
Teachers remaining in treatment condition, *spring 2014* (N = 79)
Teachers remaining in control condition, *spring 2014* (N = 86)
Teachers in analysis sample (N = 165)

Source: Study records.

*Student sample for the study-administered assessment.* As explained in chapter II (see “Student samples” section), we randomly selected about 10 students from each classroom of teachers in the analysis sample to participate in a study-administered assessment in spring 2014. Among the 2,140 students (1,020 treatment and 1,120 control) who were sampled, some students did not end up participating in the assessment. A total of 152 students were omitted from the sample because their teachers determined that accommodations related to their individualized education program or English learner status could not be met. Of the remaining students, 85 percent (806 treatment and 891 control) were tested; 6 percent had withdrawn from study schools, 8 percent were opted out by parents or guardians,[35] and 1 percent were not tested for other reasons (e.g., absence on the day of testing).

Among the students tested, treatment and control students were comparable in pretreatment characteristics, with two exceptions (see Exhibit A.5): The treatment group had a higher proportion of female students than the control group (53 percent versus 46 percent) and a lower proportion of Asian/Pacific Islanders (5 percent versus 7 percent).

**Exhibit A.5. Student Background Characteristics (Study-Administered Assessment Sample)**

| Characteristics | Treatment Group | Control Group | Estimated Difference | P value |
|---|---|---|---|---|
| Third-Grade Math Standardized Score on Spring 2013 State Assessment | 0.28 | 0.26 | 0.02 | 0.676 |
| Age (years) | 10.5 | 10.6 | −0.1 | 0.109 |
| Female (percent) | 53.2 | 45.8 | 7.4* | 0.003 |
| Race (percent) | | | | |
| White, non-Hispanic | 44.0 | 43.9 | 0.1 | 0.985 |
| Black, non-Hispanic | 14.4 | 14.9 | −0.5 | 0.724 |
| Asian, Pacific Islander, non-Hispanic | 4.9 | 7.4 | −2.5* | 0.049 |
| Hispanic | 33.5 | 30.9 | 2.6 | 0.155 |
| Other | 3.3 | 2.9 | 0.4 | 0.644 |
| Eligibility for free or reduced-price lunch (percent) [a] | 59.2 | 61.2 | −2.0 | 0.416 |
| English language learner (percent) | 12.2 | 10.0 | 2.2 | 0.232 |
| Special education status (percent) | 7.6 | 8.8 | −1.2 | 0.396 |

Note: Sample size = 73 schools; 79 teachers and 806 students in the treatment group; 86 teachers and 891 students in the control group.

[a] Estimates for free or reduced-price lunch status were unavailable for two of the six study districts. Sample size = 57 teachers and 578 students in the treatment group; 63 teachers and 645 students in the control group.

The analyses are based on a two-level model controlling for school fixed effects. Treatment group means are unadjusted means; control group means were computed by subtracting the estimated differences from the treatment group means. Sixteen students were missing race/ethnicity data (1 percent), and 95 students were missing the prior-year math achievement score (6 percent).

* Difference between treatment students and control students is statistically significant at the 0.05 level, two-tailed test.

A likelihood ratio test of overall baseline equivalence showed a statistically significant difference across the full set of student baseline characteristics (excluding free or reduced-price lunch, which was missing in two districts) (p = 0.02). A likelihood ratio test excluding gender showed no statistically significant difference across the remaining set of student baseline characteristics (p = 0.14).

Source: District administrative records.

[35] Parental or guardian consent for student participation in the study-administered assessment was obtained through opt-out procedures that were approved by the study's Institutional Review Board and research departments in each of the participating districts. Parents or guardians of all students in study classrooms received a letter explaining the study purpose, their teacher's participation in the study, and that their child could be selected by lottery to participate in a math test that spring. A denial of permission was available for families to sign and return if they did not want their child to participate. Based on guidance from study teachers about student home languages, these forms were translated and distributed as needed in Spanish, simplified Mandarin, Russian, and Vietnamese.

## Statistical power for impacts on teachers and students

We had initially estimated that the study would be powered to detect a minimum effect of 0.30–0.40 on teacher outcomes and 0.12 on student outcomes, based on a sample of 200 teachers. We recalculated the minimum detectable effect size (MDES) based on the actual analysis sample. Exhibit A.6 reports the MDES estimates for the teacher and student outcomes, using the available data for the realized analysis samples. The MDES was 0.22–0.29 for teacher knowledge outcomes, 0.27–0.46 for classroom practice outcomes, and 0.08–0.10 for student achievement outcomes.

**Exhibit A.6. Minimum Detectable Effect Sizes for Study Outcomes with the Teacher and Student Impact Analysis Samples**

| | Minimum Detectable Effect Size | |
|---|---|---|
| Outcome | *Fall* | *Spring* |
| Teacher Knowledge | 0.22 | 0.29 |
| Classroom Practice | | |
| Richness of Mathematics | 0.46 | 0.39 |
| Student Participation in Mathematics | 0.38 | 0.39 |
| Errors and Imprecision | 0.27 | 0.42 |
| Study-administered student assessment | | 0.10 |
| State assessment | | 0.08 |

Note: MDESs are based on the standard errors of impact estimates and standard deviations of the control group for the analysis sample.

Source: Fall 2013 and Spring 2014 Teacher Knowledge tests; MQI scores of video-recorded lessons from the 2013–14 school year; district administrative records; and Spring 2014 NWEA Test.

## Measures

This section describes the key measures used in the study. The study collected data at multiple time points, summarized in Exhibit A.7.

**Exhibit A.7. Data Collection Schedule**

| | 2013 | | Implementation Year: 2013–14 | | | | |
|---|---|---|---|---|---|---|---|
| Data Collection Activities | *May – June* | *July – Sept* | *Oct – Nov* | *Dec – Jan* | *Feb – Apr* | *May – June* | *July – Sept* |
| Teacher Knowledge Test | X | X | | | | X | |
| Teacher Survey | | | | | | X | |
| Video observations of classroom practice | | X | | | X[a] | | |
| Study-administered student test | | | | | | X | |
| District administrative data (records for students in participating fourth-grade teachers' classes) | | | | | | | X |
| Fidelity form and log data on the PD (Intel Math, Mathematics Learning Community, and Video Feedback Cycles) | X | X | X | X | X | | |

[a] The study collected two video observations of classroom practice for each teacher during the February to April data collection window.

The number and percentage of teachers and students who participated in data collection activities are presented in Exhibit A.8.

**Exhibit A.8. Number and Percentages of Teachers and Students Who Provided Outcome Data**

| | Treatment | | Control | |
|---|---|---|---|---|
| Data Collection Activities | *N* | *Percent* | *N* | *Percent* |
| Teacher Knowledge Test | | | | |
| Baseline | 98 | 94.2 | 114 | 97.4 |
| Fall 2013 | 95 | 91.3 | 110 | 94.0 |
| Spring 2014 | 92 | 88.5 | 110 | 94.0 |
| Teacher Survey | 93 | 89.4 | 110 | 94.0 |
| Video observations of classroom practice | | | | |
| Fall 2013 | 95 | 91.3 | 104 | 88.9 |
| Spring 2014, Observation 1 | 89 | 85.6 | 102 | 87.2 |
| Spring 2014, Observation 2 | 88 | 84.6 | 101 | 86.3 |
| Study-administered student assessment | 806 | 84.9 | 891 | 85.8 |
| State assessment | 1760 | 94.4 | 1917 | 92.7 |

Note: The percentages for teacher outcome data are based on 104 treatment teachers and 117 control teachers in the sample at random assignment. The percentages for study-administered assessment data are based on 949 treatment students and 1,039 control students in grade 4 classrooms taught by 79 treatment teachers and 86 control teachers from the teacher analysis sample. The percentages for state assessment data are based on 1,865 treatment students and 2,067 control students in grade 4 classrooms taught by 79 treatment teachers and 86 control teachers from the teacher analysis sample.

Source: Study records.

We next describe the measures that were derived on the basis of the data collected. First, we describe the measures of implementation and service contrast. We then discuss the outcome measures, including teacher knowledge, classroom practice, and student achievement.

*Measures of Intel Math implementation*. The measures used to assess the fidelity, features, and participation rates for the implementation of Intel Math are listed in Exhibit A.9.

**Fidelity of implementation of Intel Math.** The fidelity of implementation of Intel Math was measured using daily fidelity forms completed by trained AIR staff. The fidelity forms recorded the length and type of each session and activity for each Intel Math unit.

**Features of Intel Math.** We documented the features of Intel Math based on in-person observations and video recordings of each Intel Math session in each district. Video recordings captured each day of Intel Math in each district using thereNow HD Insight 2 video cameras. Each camera contains two smaller cameras, one of which focused on the board at the front of the room while the other focused on the teachers, who were seated at tables in groups of three to six.[36] This configuration provided opportunities to capture the main presentation during whole-group time and also allowed a view of individual tables during the individual or small-group activities. The videos were coded to document the time spent in each activity

[36] Due to auto-focusing problems encountered with some of the thereNow cameras, we focused both cameras on the board at the front of the room in some districts. In the event that one of the cameras failed to work properly, we wanted to ensure that we captured lead instructor presentations to the whole group.

structure (e.g., table work) and were also coded using the MQI instrument. For more information on how the videos were coded to capture the features of the PD, see chapter III.

We tracked the assignment of homework through the fidelity forms. The Intel Math instructors kept daily records of who completed the Intel Math homework (marked as complete, partially complete, or not complete).

**Teacher participation in Intel Math.** We tracked participation in Intel Math through attendance records. Each day, teachers were asked to sign in and out with their name and the time. We used these data to calculate the number of hours of Intel Math each teacher attended.

**Exhibit A.9. Measures and Data Sources for the Analysis of Implementation of Intel Math**

| Measure | Data Source |
|---|---|
| *Fidelity of Implementation of Intel Math* | |
| Number and percentage of intended hours of Intel Math delivered | Intel Math fidelity forms |
| Number and percentage of Intel Math sessions covered across Intel Math units *(Note:* Each Intel unit is divided into multiple sessions addressing different topics.) | Intel Math fidelity forms |
| Percentage of time spent in mathematics content and student thinking sessions | Intel Math fidelity forms |
| Percentage of time spent on each unit of Intel Math | Intel Math fidelity forms |
| *Features of Intel Math* | |
| Percentage of time in Intel Math dedicated to whole-group discussion (presenting new content and sharing work) and to individual and small-group work | Intel Math videos |
| Distribution of MQI scores, based on analysis of video of whole-group discussion of mathematics | Intel Math videos |
| Percentage of days on which homework was assigned and percentage of teachers who submitted their homework | Intel Math fidelity forms, Intel Math homework completion forms |
| *Teacher Participation in Intel Math* | |
| Number and percentage of implemented hours of Intel Math attended by treatment teachers | Intel Math attendance sheets |

*Measures of Mathematics Learning Community implementation.* The fidelity of implementation of the Mathematics Learning Community meetings, the features of the meetings, and teacher participation in the meetings were tracked through facilitator logs, attendance sheets, and video recordings, as summarized in Exhibit A.10.

**Fidelity of implementation of the Mathematics Learning Community.** Logs completed by facilitators tracked the duration and content coverage of the sessions, including the time spent on each part within each session. Video recordings also were used to confirm the facilitator estimates of time entered in the logs.

**Features of the Mathematics Learning Community.** To document the features of the Mathematics Learning Community sessions, each session was video-recorded and subsequently coded similarly to the video recording and coding of Intel Math. Analyses focused on the distribution of time across activity structures and the coded MQI scores (for more details, see chapter III).

**Teacher participation.** To determine participation rates in Mathematics Learning Community sessions, teachers were asked to sign attendance sheets at each session, which provided data on the number of hours teachers spent at Mathematics Learning Community sessions.

**Exhibit A.10. Measures and Data Sources for the Analysis of the Implementation of the Mathematics Learning Community Meetings**

| Measure | Data Source |
|---|---|
| *Fidelity of Implementation of the Mathematics Learning Community Meetings* | |
| Number and percentage of hours of Mathematics Learning Community meetings delivered | Mathematics Learning Community facilitator logs and videos |
| Number and percentage of Mathematics Learning Community parts covered across sessions *(Note:* Each Mathematics Learning Community meeting is divided into multiple parts.) | Mathematics Learning Community facilitator logs |
| Percentage of time spent in mathematics content and student thinking parts of the Mathematics Learning Community meetings | Mathematics Learning Community videos |
| *Features of the Mathematics Learning Community Meetings* | |
| Percentage of time in the Mathematics Learning Community dedicated to whole-group discussion (presenting new content and sharing work) and to individual and small-group work | Mathematics Learning Community videos |
| Distribution of MQI scores, based on analysis of video of whole-group discussion of mathematics | Mathematics Learning Community videos |
| *Teacher Participation in Mathematics Learning Community Meetings* | |
| Number and percentage of implemented Mathematics Learning Community hours attended by treatment teachers | Mathematics Learning Community attendance sheets |

*Measures of implementation of the Video Feedback Cycles.* The fidelity of implementation, features, and teacher participation in the Video Feedback Cycles were tracked through facilitator logs, and the video feedback forms were completed by MQI raters and the facilitators, as summarized in Exhibit A.11.

**Fidelity of the implementation of Video Feedback Cycles.** Logs completed by the facilitators (one log per Video Feedback Cycle per teacher) and feedback forms provided to the teachers (one form per Video Feedback Cycle per teacher) were used to document fidelity of the Video Feedback Cycles. The logs recorded the total time the facilitators spent in each feedback session with each teacher, and the types of activities that occurred during the sessions (e.g., watching video clips, examining Intel Math and/or Mathematics Learning Community materials). In addition, we used the feedback forms constructed by MQI raters and facilitators to document fidelity, noting whether intended elements were included, such as providing specific references to Intel Math and/or Mathematics Learning Community materials, and the next steps identified by the facilitator and the teacher at the conclusion of each feedback session.

**Features of the Video Feedback Cycles.** We also used the feedback forms to describe the features of the Video Feedback Cycles. For these analyses, we tallied the prevalence of each of the MQI elements in the feedback forms.

**Teacher participation.** Using the facilitator logs, we calculated the time each teacher spent in the Video Feedback Cycle sessions.

**Exhibit A.11. Measures, Data Sources, and Units of Analysis for the Analysis of the Implementation of the Video Feedback Cycles**

| Measure | Data Source |
|---|---|
| *Fidelity of the Implementation of Video Feedback Cycles* | |
| Number of Video Feedback Cycles implemented | Video Feedback Cycles study records |
| Percentage of lessons recorded addressing content covered in Intel Math and a prior Mathematics Learning Community session | Video Feedback Cycles feedback forms and study records |
| Percentage of feedback forms meeting all requirements | Video Feedback Cycles feedback forms |
| Percentage of feedback sessions in which all of the highlighted video clips were reviewed | Video Feedback Cycles facilitator logs |
| *Features of the Video Feedback Cycles* | |
| Prevalence of MQI elements on teachers' feedback forms | Video Feedback Cycles feedback forms |
| *Teacher Participation in Video Feedback Cycles* | |
| Percentage of intended hours of feedback received by treatment teachers | Video Feedback Cycles facilitator logs |

*Measures of service contrast.* We used a teacher survey to capture the amount of math PD in which treatment and control group teachers participated during summer 2013 and the 2013–14 school year, as well as the features of these PD activities. We administered the survey in May/June 2014 to all teachers who had been randomly assigned: 203 of the 221 teachers completed the survey, resulting in a response rate of 92 percent.

Items on the survey were used to create indices describing features of math PD that treatment and control teachers experienced during the year of the evaluation. The items used for each index are listed in Exhibit A.12. Teachers responded to items in the mathematical and student thinking activities index, math topic foci index, and lesson feedback features index using a 4-point scale representing the frequency with which the activity occurred or the topic was a focus during the PD (1 = never/rarely, 2 = sometimes, 3 = often, and 4 = most or all of the time). Items in the coherence with goals, materials, and expectations index were based on a 4-point scale from "strongly disagree" to "strongly agree." The score for each index was computed by averaging the items in the index.

**Exhibit A.12. Items Included in Indices Measuring Features of Math PD Experiences**

| Mathematical and Student Thinking Activities (10 items) |
|---|
| Solve mathematics problems |
| Share your solutions to mathematics problems with other teachers in small- or large-group settings |
| Discuss mathematics topics that appear in mathematics curricula above or below the current grade you teach |
| Explore the conceptual underpinnings of K–8 mathematics concepts |
| Practice using a variety of representations to illustrate a given mathematical concept |
| Practice analyzing student work on mathematics problems |
| Practice writing story problems for students you teach |
| Explore the connections between mathematics topics and solution methods |
| Explore the ways students commonly approach mathematics problems |
| Explore common student misconceptions and errors in mathematics |
| **Math Topic Foci (7 items)** |
| Addition, subtraction, multiplication, and division |
| Connections between addition, subtraction, multiplication, and division |
| Fractions and operations with fractions |
| Algebra |
| Connections between algebra and arithmetic |
| Decimals |
| Exponents |
| **Coherence With Goals, Materials, and Expectations (6 items)** |
| Your experience was consistent with your own goals for your professional development. |
| Your experience was complementary to your use of district-adopted curricular materials. |
| Your experience was related to the mathematics content you taught this year. |
| Your experience was logically connected from one day or session to the next. |
| Your experience was clear about how you could use what you learned from the professional development experience in your classroom. |
| Your experience was focused on practices that district or school leaders expect you to demonstrate in your classroom. |
| **Lesson Feedback Features (6 items)** |
| How you made connections between mathematics topics and/or representations |
| How you conveyed the meaning of mathematical procedures |
| How you remediated student errors |
| How you responded to student thinking |
| The clarity, precision, and correctness of your mathematics presentations |
| Provide you with useful feedback about your teaching of mathematics |

Source: Spring 2014 Teacher Survey.

*Measure of teacher knowledge.* We measured teachers' content knowledge three times: at baseline (summer 2013), following completion of the Intel Math course (fall 2013), and at the end of the school year (May or June 2014). Teacher content knowledge was measured with a computer-adaptive mathematics assessment composed of items selected from the Measures of Academic Progress (MAP) Mathematics assessment item bank, developed by the Northwest Evaluation Association.

MAP is a computer-adaptive assessment that is widely used for students. To adapt the MAP assessment for use with teachers, we first identified indices (NWEA's pre-designated collections of items addressing a math concept) for potential inclusion that aligned with the mathematical content covered in Intel Math. We then culled any indices from this set that did not meet a sufficient threshold of difficulty, based on the item RIT (**R**asch Un**it**) score scale established for students. We restricted the indices to those that had a minimum difficulty above the fourth-grade level. We chose a RIT cutoff of 220, which is the RIT score of an average fifth-grade student. We then grouped the remaining indices of items into the following five goal areas: (1) whole numbers; (2) fractions; (3) rational numbers; (4) ratio, proportion, and rate; and (5) linear equations and functions. Structuring the indices within goals ensured that the MAP adaptive software would test teachers on a consistent number of items for each of the goal areas, which all together roughly represented the content covered in Intel Math. Teachers received seven items per goal area, for a total of 35 items on the test. Completed tests were assigned a RIT score based on Rasch scaling. Each wave of the teacher knowledge test was administered in person by study team members who were trained on procedures for proctoring the computer-adaptive test. At the completion of each test, teachers' item responses and calculated RIT scores were automatically stored in secure NWEA servers.

In the summer of 2013, study staff administered the baseline teacher knowledge assessment to all treatment teachers on the first morning of Intel Math. We administered the baseline test to control teachers at a specified time on either the first or second day of Intel Math, at the location where Intel Math was being offered, but in a different room.

The first follow-up assessment was administered by study staff to all treatment and control teachers in the fall of 2013, during the first month of school. All teachers were invited to come to a central location at a specified time on one of two dates to take the first follow-up assessment. The third teacher knowledge assessment was administered in May or June 2014. We again invited all teachers to come to a central location at a specified time on one of two dates to take the test. (Teachers were asked to complete the end-of-year survey prior to taking the third test.)

After completing each wave of the test, teachers received a gift card to thank them for their time. Each teacher received a total stipend of $325 for taking three teacher knowledge tests and the survey.

***Measures of classroom instructional practice.*** We measured teachers' classroom practice based on video recordings of their instruction, coded using the MQI to assess three dimensions of math instructional practice: *Richness of Mathematics*, *Student Participation in Mathematics*, and *Errors and Imprecision*. To capture classroom lessons for MQI scoring, we conducted video observations of each study teacher once in the fall of 2013 and twice in the spring of 2014. The study team scheduled video observations individually with each of the participating treatment and control grade 4 teachers. The team worked with teachers to identify lessons that met the following criteria:

- Lesson was expected to be at least 40 minutes in length.
- Lesson was expected to have a clear instructional focus.
- Lesson was not expected to include large blocks of time devoted to test-taking, silent work, or test review.

- Lesson was expected to introduce new material in content areas covered in Intel Math about:
  - place value, multi-digit addition or subtraction, or multiplication (fall observation), or
  - fraction concepts, fraction operations, or decimals (spring observation).[37]

The study team communicated these guidelines to teachers when scheduling the observations. In addition, the study team worked with teachers to understand when they anticipated teaching the content areas covered in Intel Math. The video observations for participating teachers within a school were scheduled so that they would capture the same or similar content areas for all participating teachers in that school.

At the time of the scheduled observation, district-based videographers, hired for the study and trained by the study team, were responsible for setting up the equipment to video-record the scheduled lessons. They were responsible for setting up the thereNow HD Insight 2 video cameras in classrooms, monitoring consent procedures (e.g., ensuring that students whose parents opted them out were not viewed in the video recording), recording the lessons, and securely uploading the data to thereNow's online system.

After the video observations were recorded and uploaded, the study team conducted a quality control screening process for each video to assess the suitability of the video and audio for coding with the MQI. The process focused primarily on assessing whether writing on the classroom chalkboard or white board could be viewed, and if the teacher could be sufficiently heard. In cases where a teacher's video could not be used, the study team rescheduled observations with all of the participating teachers within the school to ensure the similarity of lesson content for teachers within schools. Ten percent of the recorded lessons for teachers in the analysis sample were retakes.

Once the videos passed the quality control screening, the lessons were transcribed and the videos and associated transcripts were shared with Harvard's Center for Education Policy Research, which oversees the certification and ongoing training of MQI raters. To become certified in the MQI, raters must complete an online training, pass a certification test, and score at least three lessons with a lead MQI rater during a follow-up apprentice period. The MQI raters have strong backgrounds in mathematics or math education and are often retired math teachers, mathematicians, or math education graduate students. For the study, 40 raters were identified and trained by the MQI developers to code the observed lessons. Among these raters, 38 had previous experience teaching K–12 math, 26 had at least a B.A. degree in mathematics or a graduate degree in math education, and 23 were newly trained and certified for this study.

Video-recorded lessons were systematically assigned to raters (two raters per lesson) to ensure a balanced mix of treatment and control teachers' videos for each rater. Raters were blind to condition. The process for scoring the observed lessons followed standard MQI procedures. First, lessons were divided into 7.5-minute segments. Then, using a 4-point scale (1 = not present, 2 = low, 3 = mid, 4 = high), the two raters per lesson

[37] In some cases, due to timing, teacher preference, camera malfunction, or other factors, it was not possible to record a lesson that introduced new material. In these cases (6 percent of recorded videos), we worked with the teacher to record during a review lesson that focused on a study-relevant topic. In addition, in some cases (due to pacing issues within districts), it was not possible to record both spring observations on fractions or decimals. In these cases (5 percent of recorded videos), we recorded lessons on other related topics, including measurement concepts such as rate and conversion.

independently assigned scores to the MQI elements and dimensions for each 7.5-minute segment within each lesson. For each 7.5-minute segment, raters also assigned a holistic score for each of the three MQI dimensions overall. At the end of the lesson, raters assigned lesson-level holistic scores for each dimension and for "mathematical quality of instruction" as a whole, on a 5-point scale (1 = low, 2 = low/mid, 3 = mid, 4 = mid/high, 5 = high). Only element scores assigned at the 7.5-minute segment level were used in the study's impact analyses, as these scores provide more detailed information than the segment-level overall dimension scores or lesson-level scores.[38] Following Agodini, Harris, Thomas, Murphy, and Gallagher (2010) and Garet et al. (2010), Exhibit A.13 provides the average percent agreement for each MQI element, based on raters' scores for all double-coded segments from treatment and control teachers' lessons used for the fall and spring analyses of the PD's impact on instructional practice. On average across elements, the raters agreed on the exact score for 65 percent of the coded segments and were within one scale point for 94 percent of the coded segments.[39]

**Exhibit A.13. Percentage of Coded Segments in Which Raters Agreed on the Exact Score (1–4) or Were Off by One Scale Point for Each MQI Element (Across Time Points and Condition)**

| Dimension | Item | Number of Segments | Percent Exact Agreement | Percent Exact Agreement or Off by 1 |
|---|---|---|---|---|
| *Richness of Mathematics* | Linking Between Representations | 4,011 | 69 | 90 |
| | Explanations | 4,011 | 56 | 89 |
| | Mathematical Sense-Making | 4,010 | 38 | 85 |
| | Multiple Procedures or Solution Methods | 4,007 | 77 | 92 |
| | Patterns and Generalizations | 4,009 | 92 | 98 |
| | Mathematical Language | 4,012 | 53 | 96 |
| | Remediation of Student Errors and Difficulties | 4,008 | 55 | 92 |
| *Errors and Imprecision* | Mathematical Content Errors | 4,013 | 93 | 98 |
| | Imprecision in Language or Notation | 4,011 | 76 | 95 |
| | Lack of Clarity in Presentation of Mathematical Content | 4,013 | 78 | 95 |
| *Student Participation in Mathematics* | Teacher Uses Student Mathematical Contributions | 4,014 | 53 | 95 |
| | Students Provide Explanations | 4,014 | 67 | 95 |
| | Student Mathematical Questioning and Reasoning | 4,011 | 69 | 95 |
| | Students Communicate about the Mathematics of the Segment | 4,013 | 53 | 95 |
| | Task Cognitive Demand | 4,012 | 43 | 91 |
| | Students Work with Contextualized Problems | 4,010 | 77 | 96 |

[38] However for ease of analysis and presentation, descriptive summaries of the PD implementation videos did use the segment-level holistic dimension scores, rather than the individual element scores.

[39] These rates were similar to those observed in two other studies that used the same version of the MQI as the current study (i.e., elements for segments rated on a 4-point scale ranging from not present to high). Exact and within-one-scale-point agreement rates are 72 percent and 94 percent for a total of 2,844 lesson segments from the Middle School Mathematics Teachers and Teaching Survey study, and 62 percent and 91 percent for a total of 1,388 lesson segments from the Mathematical ACES: Algebraic Concepts for Elementary Students study (H. Hill, personal communication, June 12, 2016).

Note: Sample size = 73 schools; 165 teachers (79 treatment teachers and 86 control teachers), and 495 lessons from fall and spring. The number of double-coded segments differs from the full number of coded segments (N = 4,076) because a small percentage of segments (less than 2 percent) had only one coder.

Source: MQI scores of video-recorded lessons from the 2013–14 school year.

**Conversion of segment scores to Rasch scale scores**. We created scale scores for the three MQI dimensions (*Richness*, *Student Participation*, and *Errors*), based on the element-level results. We used Item Response Theory (IRT) scaling to derive the scales because the 4-point metric used to rate each element (1 = not present, 2 = low, 3 = mid, 4 = high) is ordinal rather than interval. Also, the distributions of the segment-level data from the MQI are quite skewed (see Exhibit A.14). In particular, for five of the seven *Richness of Mathematics* elements, all three of the *Errors and Imprecision* elements, and three of the six *Student Participation in Mathematics* elements, more than 50 percent of segments received a score of 1 (indicating not present) on the 1–4 metric. For six of the 16 total elements, more than 75 percent of segments received a score of 1.

**Exhibit A.14. Percentage of Ratings in Each MQI Score Category for Each MQI Element (Across Time Points, Condition, and Raters)**

| Dimension 1: *Richness of Mathematics* | | | | | | | |
|---|---|---|---|---|---|---|---|
| Score | Linking Between /Within Representations | Explanations | Mathematical Sense-Making | Multiple Procedures | Patterns and Generalizations | Mathematical Language | Remediation of Student Errors and Difficulties |
| 1 | 58.8 | 65.5 | 27.9 | 82.2 | 94.7 | 9.6 | 52.6 |
| 2 | 24.2 | 23.6 | 34.2 | 9.4 | 3.6 | 58.9 | 32.3 |
| 3 | 12.7 | 9.0 | 29.4 | 7.6 | 1.4 | 28.2 | 13.5 |
| 4 | 4.3 | 1.9 | 8.5 | 0.9 | 0.3 | 3.3 | 1.6 |
| Dimension 2: *Student Participation in Mathematics* | | | | | | | |
| Score | Teacher Uses Student Contributions | Students Provide Explanations | Student Mathematical Questioning | Students Communicate About Math | Task Cognitive Demand | Students Work With Contextualized Problems | |
| 1 | 14.6 | 70.4 | 75.2 | 11.4 | 34.3 | 73.7 | |
| 2 | 55.7 | 23.1 | 19.7 | 46.1 | 43.4 | 18.1 | |
| 3 | 26.3 | 5.4 | 4.3 | 38.8 | 20.1 | 6.5 | |
| 4 | 3.5 | 1.0 | 0.7 | 3.8 | 2.2 | 1.7 | |
| Dimension 3: *Errors and Imprecision* | | | | | | | |
| Score | Mathematical Content Errors | Imprecision in Language or Notation | Lack of Clarity in Presentation | | | | |
| 1 | 95.6 | 83.0 | 86.5 | | | | |
| 2 | 2.7 | 13.1 | 9.6 | | | | |
| 3 | 1.4 | 3.5 | 3.3 | | | | |
| 4 | 0.4 | 0.5 | 0.6 | | | | |

Note: Sample size = 73 schools; 165 teachers (79 treatment teachers and 86 control teachers), 495 lessons, and 4,076 segments.

A score of 1 is “Not Present”; 2 is “Low” (low incidence and basic level); 3 is “Mid” (moderate incidence and level); and 4 is “High” (extensively present or high level).

Source: MQI scores of video-recorded lessons from the 2013–14 school year.

We followed a two-step process to construct the measures for the *Richness of Mathematics* and *Student Participation in Mathematics* dimensions. In the first step, we used Facets (Linacre, 2014) to generate segment-level scale scores for each dimension based on Maximum Likelihood Estimation (MLE), using a 1-parameter partial credit model.[40] Raters (two per lesson) were treated as a facet in the model to take potential additive effects of raters into account. The segment-level MLE scores were then used as the outcome measure for the analyses conducted in the second step, where we estimated the treatment effect using a two-level model for the analysis of fall observations, based on a single lesson observation for each teacher (with segments nested within teachers), and a three-level model for the analysis of spring observations, based on two recorded lessons for each teacher (with segments nested within lessons within teachers).

For the *Errors and Imprecision* dimension, we also used Facets to generate MLE scores. The scores were created at the lesson level instead of the segment level in the first step, thus reducing the percentage of cases with extreme values from 60 percent (of segments) to 13 percent (of lessons).[41] To generate lesson-level MLE scores, we treated segments as facets (analogous to raters).[42] We then used these lesson-level scores in a teacher-level regression for fall observations in the second step and a two-level impact model (lessons within teachers) for spring observations. Exploratory analyses revealed that the two-step approach yielded similar results to a one-step, multi-level IRT approach (e.g., Cheong & Raudenbush, 2000), which was considered but ultimately not used because of the computational power demanded.

**Conversion of Rasch scale scores to probabilities for each MQI dimension.** For each MQI dimension, Facets provided item difficulty estimates,[43] and the item difficulties were averaged across elements to produce the average element difficulty. We then calculated the probability for an average treatment teacher to score “low,” “mid,” “high,” or “not present” during a typical 7.5-minute segment based on a 1-parameter partial credit model, using the unadjusted treatment group mean as the ability parameter and the average element difficulty as the difficulty parameter. Similarly, we calculated the probabilities for the control group based on the estimated control group mean and the average element difficulty. We then used the Rasch

[40] Facets is a Rasch measurement software package that allows for applications of the Rasch model in the areas of performance assessment where there is a need to adjust for aspects that might affect the scores (for example, different tasks faced by examinees, different teachers scoring the exam, or different locations where exams are given). We checked the Rasch assumption of a common discrimination parameter (item slope), by using Parscale (Muraki & Bock, 2003) to estimate a 2-parameter partial credit model. We found little evidence of variation in item discrimination, affirming the appropriateness of using Rasch model scaling.

[41] For a segment, if all items within a dimension were rated at the lowest or highest level, the segment had an extreme value. Because MLE estimates for cases with extreme values do not exist, Facets assigns arbitrary scores to cases with extreme values, using an approach standard in IRT scaling.

[42] The number of segments within a lesson ranged from 4 to 17. In exploratory analyses for the *Richness of Mathematics* and *Student Participation in Mathematics* dimensions, we included covariates at the segment level to mark the first and last segment of each lesson, and found that the first and last segments in a lesson on average tend to have lower MQI scores than the remaining segments. To take into account the potential effect of segment order in the Facets model for the *Errors and Imprecision* dimension, we recoded the first segment as segment 1, the last segment as segment 2, and all other segments as their original number plus one.

[43] For an ordinal item with 4 levels, there are 3 item difficulty parameters (also called threshold parameters): one at level 2 versus level 1, one at level 3 versus level 2, and one at level 4 versus level 3.

scores to determine the probability that a teacher demonstrated *Richness*, *Student Participation,* or *Errors* during a typical segment. We defined demonstrating *Richness* or demonstrating *Student Participation* in a segment as scoring mid or high on one or more elements of the dimension during the segment. We defined demonstrating *Errors* in a segment as scoring low, mid, or high on one or more elements during the segment.

*Measures of student achievement.* We assessed the impact of the study PD on student achievement with two measures. First, we administered a computer-adaptive assessment—a modified version of the NWEA Measures of Academic Progress—to a random sample of students in participating grade 4 teachers' classes. Second, we requested scores on the state mathematics assessment for all students in participating grade 4 teachers' classes.

- **Study-administered assessment.** We administered a student achievement test as part of the study to provide a common achievement measure across sites and to obtain data on student performance on items aligned in content with the study PD. To carry out the alignment, we used the teacher content knowledge assessment (described previously) as a starting point, and selected indices (collections of items pre-designated by NWEA addressing a math concept) for students that aligned with the whole numbers, fractions, and decimal portions because these were the topics that would be covered in fourth grade. We did not remove indices based on RIT score difficulty as we had with the teacher assessment. We again grouped the selected indices into goal areas to ensure the adaptive software would test students on an equal number of items within the targeted content areas: (1) whole numbers and decimals, and (2) fractions. RIT scores for students on this customized assessment were based on the Rasch scale typical for the NWEA MAP.

  After the student assessment sample was drawn, we worked with each participating teacher to devise a plan for the study team or trained local data collectors to administer the assessment to his or her students in a way that minimized disruption and missed instructional time. We administered the assessment in school computer labs, working with the district liaison and school representatives to schedule the sessions. Students took approximately 30 minutes on the computer to complete the assessment.

- **State mathematics assessment.** Scores on the spring 2014 state mathematics assessments (used for accountability purposes) served as a policy-relevant outcome for this study. The 6 participating districts were located in 5 states, each of which used a different math assessment for accountability purposes in the 2013–14 school year. We requested these scores for all students assigned to participating grade 4 teachers' classes, and we standardized them separately for each state, based on the state mean and standard deviation.

## Analyses

This section presents the models used to assess the average impact of the PD on teacher knowledge, classroom practice, and student achievement, the models used to examine the differential impact of the PD for teachers and students who differed in measured background characteristics, and the models used to estimate the relationships among the three types of outcomes.

*Analyses of the PD impact on teacher knowledge.* Analyses of the impact of the PD on teacher knowledge as measured by the NWEA test were estimated separately for teacher knowledge scores in fall 2013 and spring 2014 based on the regression model shown in Exhibit A.15.

**Exhibit A.15. Impact Model for Teacher Knowledge**

$$Y_k = \sum_{s=1}^{73} \beta_{0s} S_{sk} + \sum_{d=1}^{6} \beta_{1d} (T * D_d)_k + \sum_{c=1}^{C} \beta_{2c} W_{ck} + r_k \qquad (1)$$

Where:

- $Y_k$ is the teacher knowledge RIT score in fall 2013 or spring 2014 for teacher *k*;
- $S_{sk}$ , s = 1, 2, … 73, are 73 dummy variables indicating whether teacher *k* taught in school s;
- $(T * D_d)_k$ , *d* = 1–6, are six treatment-by-district interactions, indicating whether teacher *k* was in the treatment group in district *d*;
- $W_{ck}$ , c = 1, 2, … C, is a vector of C background characteristics for teacher *k*;
- $\beta_{0s}$ represents the average knowledge test score among control teachers in school *s*, adjusted for teacher background characteristics;
- $\beta_{1d}$ captures the treatment effect on teacher knowledge in district *d*;
- $\beta_{2c}$ represents the relationship between teacher background characteristic *c* and teacher knowledge test scores; and
- $r_k$ is a random error associated with teacher *k*.

To improve the precision of the estimates, we included a set of covariates for teacher background characteristics ($W_{ck}$) that were expected to be correlated with teacher knowledge. In addition to baseline teacher knowledge score, the covariates included teaching experience (coded as dummy variables representing 0–3 years, 4–10 years, 11–20 years, and 20+ years of experience), an indicator for whether the teacher reported taking calculus in college, and whether the teacher had a graduate degree.

The model generated an estimate for the impact of the PD on teacher knowledge in each district ($\beta_{1d}$), and the average impact across all six districts was computed as a precision-weighted average.

*Analyses of the PD impact on classroom practice.* The main impact analyses for classroom practice were conducted separately using data from the fall and the spring of the 2013–14 school year. The classroom observation conducted in the fall took place prior to the start of the Mathematics Learning Community and Video Feedback Cycles components of the intervention; thus, the fall analysis provides an estimate of the impact of Intel Math only. The analysis based on spring data provides an estimate of the combined effect of the three program components.

The analysis was conducted separately for each of the three MQI dimensions: *Richness of Mathematics*, *Student Participation in Mathematics*, and *Errors and Imprecision*. As previously explained, the *Richness of Mathematics* and *Student Participation in Mathematics* scales were created using Rasch scaling at the segment level. Because the analyses of the impact of the PD in fall 2013 was based on only one lesson per teacher, we estimated the impact on the *Richness of Mathematics* and *Student Participation in Mathematics* outcomes using a two-level model that nested segment-level Rasch scores within teachers, as shown in Exhibit A.16.

**Exhibit A.16. Impact Model for Classroom Practice (*Richness of Mathematics* and *Student Participation in Mathematics* Dimensions), Fall Outcomes**

Level 1 (Segments):

$$Y_{ik} = \pi_{0k} + \pi_{1k}(FIRST)_{ik} + \pi_{2k}(LAST)_{ik} + \varepsilon_{ik} \qquad (2)$$

Where:

- $Y_{ik}$ is the Rasch score for either the *Richness of Mathematics* or *Student Participation in Mathematics* dimension for segment *i* of the fall lesson taught by teacher *k*;
- $(FIRST)_{ik}$ and $(LAST)_{ik}$ are dummy variables indicating whether segment *i* was the first or last segment of the lesson taught by teacher *k*, grand-mean centered, with the rest of the segments of the lesson (i.e., "middle segments") as the omitted reference;[44]
- $\pi_{0k}$ is the adjusted average score for the dimension across all segments in the lesson taught by teacher *k*;
- $\pi_{1k}$ and $\pi_{2k}$ represent the difference in the average scores between the first segment and the middle segments and between the last segment and the middle segments, respectively, of the lesson taught by teacher *k*; and
- $\varepsilon_{ik}$ is a random error associated with segment *i* of the lesson taught by teacher *k*.

Level 2 (Teachers):

$$\pi_{0k} = \sum_{s=1}^{73}\gamma_{00s}S_{sk} + \sum_{d=1}^{6}\gamma_{01d}(T*D_d)_k + \gamma_{02}AVGACH_k + \sum_{c=1}^{C}\gamma_{03c}W_{ck} +$$

$$\gamma_{04}(TUE)_k + \gamma_{05}(WED)_k + \gamma_{06}(THU)_k + \gamma_{07}(FRI)_k + u_{0k} \qquad (3)$$

$$\pi_{1k} = \gamma_{10} \qquad (4)$$

[44] Dummy indicators were included to control for segment order because instruction might differ at the beginning, middle, and end of a lesson. Further, the length of the last segment of each lesson was often shorter than 7.5 minutes. Thus, controlling for segment order also may adjust for segment length.

$$\pi_{2k} = \gamma_{20} \tag{5}$$

Where:

- $S_{sk}$, $(T*D_d)_k$, and $W_{ck}$ are defined as in Exhibit A.15;
- $AVGACH_k$ is the classroom average prior year (grade 3) achievement of the students taught by teacher *k*, grand-mean centered;
- $(TUE)_k$, $(WED)_k$, $(THU)_k$, and $(FRI)_k$ are a set of dummy indicators for different days of the week that the lesson taught by teacher *k* was observed, grand-mean centered, with Monday as the omitted reference;[45]
- $\gamma_{00s}$ is the average score for the given dimension across all control teachers in school *s*;
- $\gamma_{01d}$ represents the treatment effect on the dimension score in district *d*;
- $\gamma_{02}$ represents the relationship between classroom average prior achievement and teacher dimension scores;
- $\gamma_{03c}$ represents the relationship between teacher background characteristic *c* and teacher dimension scores;
- $\gamma_{04}$~$\gamma_{07}$ represent the average difference in the dimension score between a lesson observed on each weekday (Tuesday through Friday) relative to a lesson observed on Monday across all teachers;
- $\gamma_{10}$ and $\gamma_{20}$ represent the average differences in the average dimension scores between the first segment and the middle segments, and between the last segment and the middle segments of a lesson, respectively, taught by all teachers; and
- $u_{0k}$ is a random error associated with teacher *k*.

The coefficient $\gamma_{01d}$ from the level-2 model represents the treatment effect on the Rasch score for the given dimension in district *d*, and the overall treatment effect was computed as a precision-weighted average effect across the six school districts. The teacher-level covariates used in this set of analyses included teacher experience, whether the teacher had a graduate degree, baseline teacher knowledge, and class size.

Because the study had data from two lessons in the spring, the impact on these two MQI dimensions in the spring were estimated with a three-level model that nested segments within lessons, which were in turn nested within teachers. The model was specified as displayed in Exhibit A.17.

[45] Dummy indicators were included for days of the week because the quality of instruction might have differed on different days of the week.

**Exhibit A.17. Impact Model for Classroom Practice (*Richness of Mathematics* and *Student Participation in Mathematics* Dimensions), Spring Outcomes**

Level 1 (Segments):

$$Y_{ijk} = \pi_{0jk} + \pi_{1jk}(FIRST)_{ijk} + \pi_{2jk}(LAST)_{ijk} + \varepsilon_{ijk} \tag{6}$$

Where:

- $Y_{ijk}$ is the Rasch score for either the *Richness of Mathematics* or *Student Participation in Mathematics* dimension for segment *i* of lesson *j* taught by teacher *k*;
- $(FIRST)_{ijk}$ and $(LAST)_{ijk}$ are defined as in Exhibit A.16;
- $\pi_{0jk}$ is the adjusted average score for the dimension across all segments in lesson *j* taught by teacher *k*;
- $\pi_{1jk}$ and $\pi_{2jk}$ represent the difference in the average scores between the first segment and the middle segments and between the last segment and the middle segments, respectively, of lesson *j* taught by teacher *k*; and
- $\varepsilon_{ijk}$ is a random error associated with segment *i* of lesson *j* taught by teacher *k*.

Level 2 (Lessons):

$$\pi_{0jk} = \beta_{00k} + \beta_{01k}(TUE)_{jk} + \beta_{02k}(WED)_{jk} + \beta_{03k}(THU)_{jk} + \beta_{04k}(FRI)_{jk} + r_{0jk} \tag{7}$$

$$\pi_{1jk} = \beta_{10k} \tag{8}$$

$$\pi_{2jk} = \beta_{20k} \tag{9}$$

Where:

- $(TUE)_{jk}$, $(WED)_{jk}$, $(THU)_{jk}$, and $(FRI)_{jk}$ are defined as in Exhibit A.16 for lesson *j* taught by teacher *k*;
- $\beta_{00k}$ is the adjusted average score for the given dimension across all lessons taught by teacher *k*;
- $\beta_{01k} \sim \beta_{04k}$ represent the differences in the average score across segments between lessons observed on each weekday (Tuesday through Friday) relative to lessons observed on Monday for teacher *k*;
- $\beta_{10k}$ and $\beta_{20k}$ represent the average differences in the average scores between the first segment and the middle segments, and between the last segment and the middle segments, respectively, across all lessons taught by teacher *k*; and
- $r_{0jk}$ is a random error associated with lesson *j* taught by teacher *k*.

Level 3 (Teachers):

$$\beta_{00k} = \sum_{s=1}^{73} \gamma_{000s} S_{sk} + \sum_{d=1}^{6} \gamma_{001d} (T * D_d)_k + \gamma_{002} AVGACH_k + \sum_{c=1}^{C} \gamma_{003c} W_{ck} + u_{00k} \quad (10)$$

$$\beta_{0nk} = \gamma_{0n0} \text{, n = 1, 2, 3, and 4} \quad (11)$$

$$\beta_{10k} = \gamma_{100} \quad (12)$$

$$\beta_{20k} = \gamma_{200} \quad (13)$$

Where:

- $S_{sk}$ , $(T * D_d)_k$ , and $W_{ck}$ are defined as in Exhibit A.15;
- $AVGACH_k$ is the classroom average prior year (grade 3) achievement of the students taught by teacher *k*, grand-mean centered;
- $\gamma_{000s}$ is the average score for the given dimension across all control teachers in school *s*;
- $\gamma_{001d}$ represents the treatment effect on the dimension score in district *d*;
- $\gamma_{002}$ represents the relationship between classroom average prior achievement and teacher dimension scores;
- $\gamma_{003c}$ represents the relationship between teacher background characteristic *c* and teacher dimension scores;
- $\gamma_{0n0}$ represents the average difference in the dimension score between lessons observed on each weekday (Tuesday through Friday) relative to lessons observed on Monday across all teachers;
- $\gamma_{100}$ and $\gamma_{200}$ represent the average differences in the average dimension scores between the first segment and the middle segments, and between the last segment and the middle segments, respectively, across all lessons taught by all teachers; and
- $u_{00k}$ is a random error associated with teacher *k*.

The same set of teacher-level covariates was used as in the fall analyses, and we again computed the average treatment effect as a precision-weighted average effect across the six school districts.

*Errors and Imprecision* was measured using lesson-level scores created through Rasch scaling; therefore, the models used to assess the impact of the PD on this dimension do not have a segment level, but are otherwise similar to the models described above.

***Analyses of the PD impact on student achievement.*** To test the impact of the PD on student achievement at the end of the 2013–14 school year, we estimated the two-level model in Exhibit A.18.

**Exhibit A.18. Impact Model for Student Achievement**

Level 1 (Students):

$$Y_{mk} = \pi_{0k} + \sum_{g=1}^{G} \pi_{1kg} X_{gmk} + \varepsilon_{mk} \tag{14}$$

Where:

- $Y_{mk}$ is the test score of student *m* taught by teacher *k*;
- $X_{gmk}$, g = 1, 2, … G, is a vector of G background characteristics of student *m* taught by teacher *k*, grand-mean centered;
- $\pi_{0k}$ is the average test score for students taught by teacher *k*, adjusted for student characteristics;
- $\pi_{1kg}$ is the relationship between student characteristic *g* and student test scores among students taught by teacher *k*; and
- $\varepsilon_{mk}$ is a random error associated with a given student.

Level 2 (Teachers):

$$\pi_{0k} = \sum_{s=1}^{73} \beta_{00s} S_{sk} + \sum_{d=1}^{6} \beta_{01d} (T * D_d)_k + \sum_{c=1}^{C} \beta_{02c} W_{ck} + r_{0k} \tag{15}$$

$$\pi_{1kg} = \beta_{10g} \tag{16}$$

Where:

- $S_{sk}$, $(T * D_d)_k$, and $W_{ck}$ are defined as in Exhibit A.15;
- $\beta_{00s}$ represents the average student test score among control teachers in school *s*;
- $\beta_{01d}$ represents the treatment effect on student achievement in district *d*;
- $\beta_{02c}$ represents the relationship between teacher characteristic *c* and student average test scores;
- $\beta_{10g}$ represents the average relationship between student characteristic *g* and student test scores across all teachers; and
- $r_{0k}$ is a random error associated with teacher *k*.

To improve the precision of the impact estimates, the model in Exhibit A.18 incorporates covariates at both student and teacher levels. The student-level covariates include gender, age, race/ethnicity, English learner status, special education status, and prior (grade 3) achievement scores.[46] The teacher-level covariates include baseline teacher knowledge score and average class prior achievement. The average impact across the six districts was computed as a precision-weighted average impact.

*Treatment of missing data on covariates.* The analysis samples for all impact analyses included teachers and students with complete outcome data, but there are missing data on some covariates. (Specifically, as noted in chapter II, 4 percent of teachers were missing data on whether or not they had taken a calculus course, 1 percent of students were missing race/ethnicity data, and 6 percent of students were missing the prior year math achievement score.) Missing covariate data were handled using the dummy variable adjustment approach (Puma, Bell, Olsen, & Price, 2009). For each covariate with missing data, we set the missing value to zero and included a missingness indicator in the impact model.

*Sensitivity analyses.* For each main impact analysis, we conducted three types of sensitivity analyses. The first type of sensitivity analysis included only the school fixed effects and treatment-by-district interactions, and excluded all other covariates. The impact analyses without covariates make fewer assumptions and are expected to produce similar point estimates but larger standard errors for the treatment effects relative to impact analyses with covariates for randomized controlled trials.

The second type of sensitivity analysis was based on a fully specified model including all teachers or students with the relevant outcome data, not just those in the impact analysis sample. The expanded sample analysis was not conducted for the classroom practice outcomes, however, because the time and effort to convert the raw MQI scores to dimension-specific Rasch scales for teachers not in the analysis sample was substantial.

The third type of sensitivity analysis was based on the fully specified model for the main impact analysis but excluded a study district in which there were particularly large differences at baseline between the treatment and control groups, for both the teacher and student samples. In this district, we observed higher scores on the baseline teacher knowledge test among control teachers than treatment teachers. Control teachers' students also had higher scores on their prior-year state math assessment than did treatment teachers' students.

*Differential impact analyses.* For each of the teacher and student outcomes, we explored whether the impact of the PD varied based on certain teacher or student characteristics, as described below.

**Analyses of the differential impact of the PD on teacher knowledge.** We explored whether the impact of the PD varied depending on teachers' prior mathematics knowledge. Teachers with lower levels of prior knowledge, for example, may have benefited less from the PD if the PD offered was too challenging for them. Conversely, teachers with higher levels of prior knowledge may have benefited less if the PD offered was too easy. To test for the differential impact of the PD on teacher outcomes, we modified the main teacher impact models by adding a treatment-by-district-by-baseline knowledge interaction term to the teacher-level equation, allowing the potential differential impact to vary by district. The model for testing the differential PD impact on teacher knowledge was adapted from Equation (1), as shown in Exhibit A.19.

[46] Two districts did not provide free or reduced-price lunch eligibility status for students, so it was not included as a covariate in the model.

**Exhibit A.19. Model to Estimate the Differential Impact of the PD on Teacher Knowledge by Teachers' Baseline Knowledge**

$$Y_k = \sum_{s=1}^{73} \beta_{0s} S_{sk} + \sum_{d=1}^{6} \beta_{1d} (T * D_d)_k + \sum_{d=1}^{6} \beta_{2d} (T * D_d * BK)_k + \beta_3 (BK)_k + \sum_{c=1}^{C} \beta_{4c} W_{ck} + r_k \qquad (17)$$

Where:

- $Y_k$ is the teacher knowledge RIT score in fall 2013 or spring 2014 for teacher *k*;
- $S_{sk}$, $(T * D_d)_k$, and $W_{ck}$ are defined as in Exhibit A.15;
- $(BK)_k$ is the baseline knowledge test score for teacher *k*, standardized based on the control group mean and standard deviation;
- $(T * D_d * BK)_k$ is a three-way interaction equaling the baseline knowledge score for teacher *k* if the teacher is in the treatment group in district *d*, and 0 otherwise;
- $\beta_{0s}$ represents the average knowledge test score among control teachers with a baseline knowledge score of 0 (i.e., control group mean) in school *s*, adjusted for teacher background characteristics;
- $\beta_{1d}$ captures the treatment effect on teacher knowledge score for teachers with an average baseline knowledge score of 0 in district *d*;
- $\beta_{2d}$ represents the change in treatment effect per 1 standard deviation increase in teachers' baseline knowledge score in district *d*;
- $\beta_3$ represents the relationship between teachers' baseline knowledge score and their fall or spring knowledge test scores; and
- $\beta_{4c}$ represents the relationship between teacher background characteristic *c* and teacher knowledge test scores.

A statistically significant positive value of $\beta_{2d}$ would suggest that teachers with higher baseline knowledge scores benefited more from the PD than teachers with lower baseline knowledge scores in district *d*. The overall differential PD impact was computed as a precision-weighted average differential impact across the six study districts.

**Analyses of the differential impact of the PD on classroom practice.** To examine whether baseline teacher knowledge and average class prior achievement moderated the impact of the PD on the classroom practice measures, we added interaction terms to the teacher-level equations in the main impact model for classroom practice, and computed the overall differential effect as a precision-weighted average differential impact across the six study districts. We estimated the differential impact model separately for each of the three MQI dimensions and for lessons observed in fall 2013 and lessons observed in spring 2014.

**Analyses of the differential impact of the PD on student achievement.** We tested for differential impacts of the PD on student achievement by including appropriate interactions at the teacher and student levels in the main impact model for student achievement. The potential moderators that we examined for the impact of the PD on student achievement included baseline teacher knowledge, teacher experience, class average prior achievement, and student's prior year (grade 3) state math assessment score.[47]

*Correlational analyses.* In addition to impact analyses, we also conducted a set of correlational analyses to examine the relationships among teacher knowledge, classroom practice, and student achievement in grade 4 mathematics. Although these analyses are not causal, they provide suggestive evidence on the validity of the theory of action underlying the study PD (see Exhibit 1.1 in chapter I for the study's conceptual framework).

To examine the relationships between teacher knowledge and the measures of classroom practice, we estimated a three-level model that used the teacher knowledge score on the fall 2013 test to predict classroom practice scores measured on the basis of the two lessons observed in spring 2014. The model is identical to the main impact model for classroom practice as presented in Exhibit A.17, except that the treatment indicator was replaced by the fall 2013 teacher knowledge score and the covariate for baseline teacher knowledge was removed.

To examine the relationships between teacher outcomes and student achievement, we modified the student achievement impact model (Equations 14 to 16) by replacing the treatment indicator with the teacher knowledge measure that averaged a teacher's two scores over the fall and spring teacher knowledge assessments or a teacher-level measure of classroom practice (rather than the segment- or lesson-level MQI dimension scores previously described).

We took additional steps to generate the teacher-level classroom practice measure. As described earlier (see "Measures of classroom instructional practice"), the approach to scaling used for *Richness of Mathematics* and for *Student Participation in Mathematics* generated scores for each 7.5-minute segment for each of the three lessons taught by each teacher. To estimate the models relating classroom practice to student achievement, it was necessary to generate a teacher-level measure of classroom practice. To obtain teacher-level classroom practice measures for the *Richness of Mathematics* and *Student Participation in Mathematics* dimensions, we estimated a three-level model adapted from the main impact model for classroom practice, based on the lessons observed in both the fall and the spring, as specified in Exhibit A.20. For the *Errors and Imprecision* dimension, a two-level model (lessons nested within teachers) was used to obtain the teacher-level measure for the dimension.

[47] To facilitate the interpretation of the analysis with teacher experience as a moderator, we used a dichotomous measure of teacher experience (5 or fewer years of experience versus 6 or more years) rather than the four teacher experience dummy variables used as covariates in the main impact analyses for teacher outcomes.

**Exhibit A.20. Model to Create Teacher-Level Classroom Practice Measures, *Richness of Mathematics* and *Student Participation in Mathematics* Dimensions**

Level 1 (Segments):

$$Y_{ijk} = \pi_{0jk} + \pi_{1jk}(FIRST)_{ijk} + \pi_{2jk}(LAST)_{ijk} + \varepsilon_{ijk} \tag{18}$$

Where all terms are defined as in Exhibit A.17.

Level 2 (Lessons):

$$\pi_{0jk} = \beta_{00k} + \beta_{01k}(TUE)_{jk} + \beta_{02k}(WED)_{jk} + \beta_{03k}(THU)_{jk} + \beta_{04k}(FRI)_{jk} + r_{0jk} \tag{19}$$

$$\pi_{1jk} = \beta_{10k} \tag{20}$$

$$\pi_{2jk} = \beta_{20k} \tag{21}$$

Where all terms are defined as in Exhibit

A.17. Level 3 (Teachers):

$$\beta_{00k} = \gamma_{000} + \gamma_{001}T_k + u_{00k} \qquad (22)^{48}$$

$$\beta_{0nk} = \gamma_{0n0}\text{, } n = 1, 2, 3, \text{ and } 4 \tag{23}$$

$$\beta_{10k} = \gamma_{100} \tag{24}$$

$$\beta_{20k} = \gamma_{200} \tag{25}$$

Where:

- $T_k$ is a treatment indicator, coded 1 if teacher *k* was in the treatment group and 0 otherwise;
- $\gamma_{000}$ is the average score for the given dimension across all control teachers;
- $\gamma_{001}$ is the difference between treatment teachers and control teachers in the average score for the given dimension; and
- all other terms are defined as in Exhibit A.17.

Based on the estimates from the model in Exhibit A.20, we calculated the teacher-level score for the given dimension as the Empirical Bayes residual for the teacher ( $u_{00k}$ ) plus the grand mean ( $\gamma_{000}$ ) for each control teacher, and as the Empirical Bayes residual plus the grand mean plus the treatment effect estimate ( $\gamma_{001}$ ) for each treatment teacher.[49] These dimension-specific, teacher-level measures of classroom practice were used as the main predictors for the analysis of the relationship between classroom practice and student achievement.

[48] This equation does not include school fixed effects and teacher background characteristics because those are controlled for in the model assessing the relationship between teacher practice and student achievement.

[49] The treatment indicator was included in the measurement model because otherwise the Empirical Bayes residuals for the treatment and control teachers would have been shrunk to a common mean, biasing any treatment-control difference toward zero.

This page has been left blank for double-sided copying.

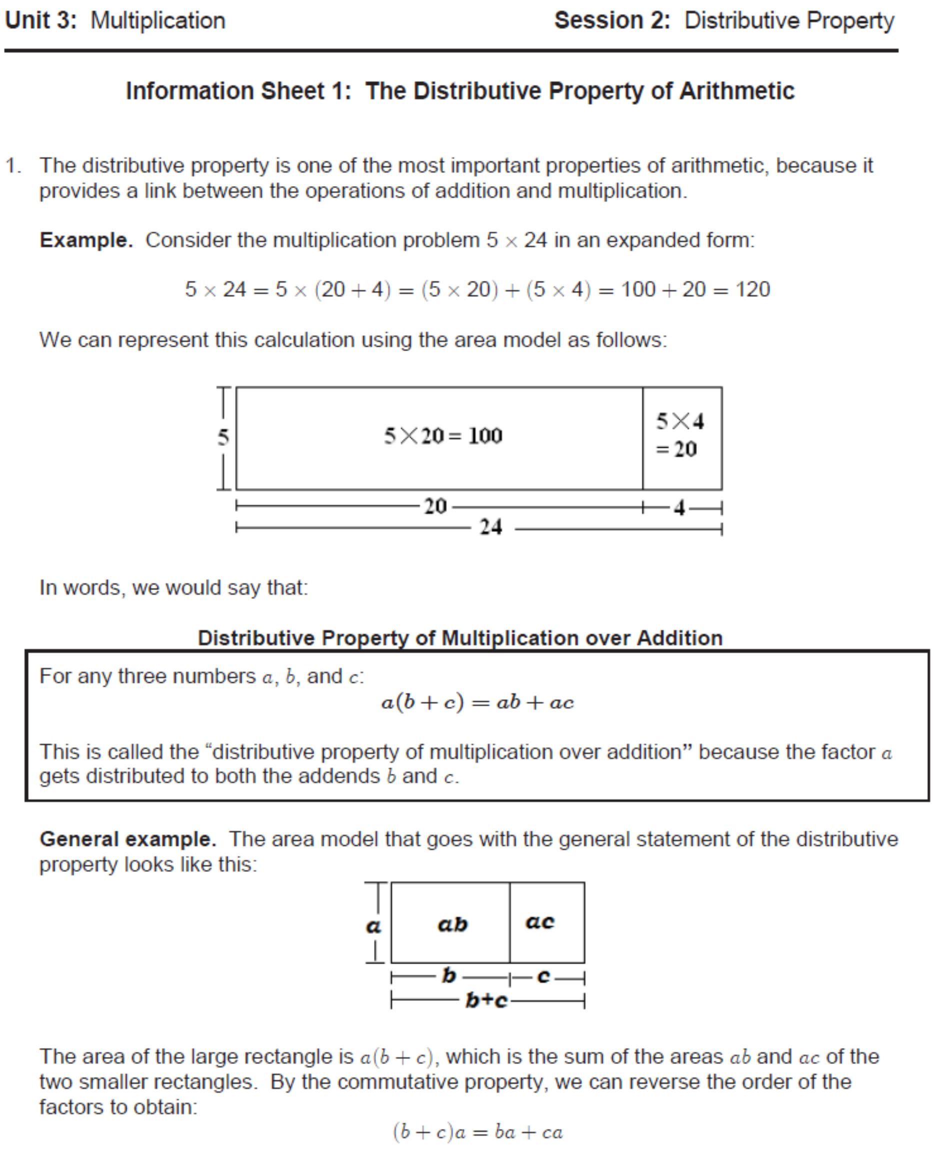
# Appendix B. Supplemental Information about the Study PD

This appendix includes information about each of the three study PD components to supplement the descriptions provided in chapter III.

## Intel Math

Exhibits B.1–B.3 show examples of the three common types of Intel Math materials: information sheets, problem sets, and exercise sets. Exhibit B.1 includes a sample information sheet from Unit 3 of the program. Information sheets are one way that topics are introduced in Intel Math.

**Exhibit B.1. Example Intel Math Information Sheet**



**Unit 3:** Multiplication **Session 2:** Distributive Property

**Information Sheet 1: The Distributive Property of Arithmetic**

1. The distributive property is one of the most important properties of arithmetic, because it provides a link between the operations of addition and multiplication.

**Example.** Consider the multiplication problem 5 × 24 in an expanded form:

$$5 \times 24 = 5 \times (20 + 4) = (5 \times 20) + (5 \times 4) = 100 + 20 = 120$$

We can represent this calculation using the area model as follows:

In words, we would say that:

**Distributive Property of Multiplication over Addition**

For any three numbers $a$, $b$, and $c$:

$$a(b + c) = ab + ac$$

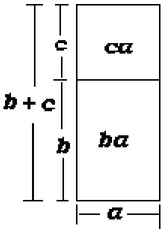
This is called the "distributive property of multiplication over addition" because the factor $a$ gets distributed to both the addends $b$ and $c$.

**General example.** The area model that goes with the general statement of the distributive property looks like this:

The area of the large rectangle is $a(b + c)$, which is the sum of the areas $ab$ and $ac$ of the two smaller rectangles. By the commutative property, we can reverse the order of the factors to obtain:

$$(b + c)a = ba + ca$$

This equation corresponds geometrically to rotating the preceding rectangle by 90 degrees.



2. When we read the formula for the distributive property from left to right:

   $$a(b+c) = ab + ac$$

   we **distribute the factor $a$ through the parentheses** to each of the addends $b$ and $c$.

   When we read the formula from right to left (allowable because of symmetry):

   $$ab + ac = a(b+c)$$

   we say that $a$ is a **common factor** of the terms $ab$ and $ac$, and we have **factored** the expression $ab + ac$ into two factors, $a$ and $(b+c)$.

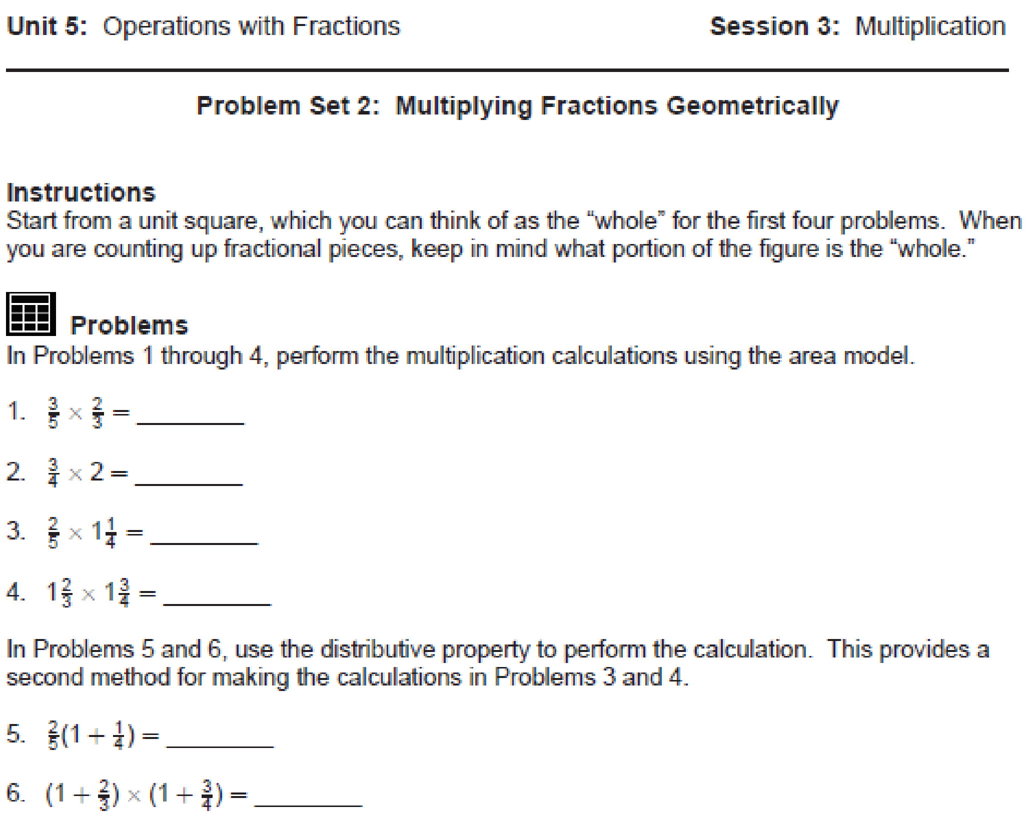
### ≠ Common Error

Note that $a(b+c) \neq ab + c$ because the distributive property requires that we distribute factor $a$ through the parentheses.

Source: Intel Math materials (Intel Foundation, 2009).

Exhibit B.2 contains a sample problem set from Unit 5 of Intel Math. This problem set focused on multiplying fractions and asked teachers to use the area model and apply the distributive property when solving the problems.

**Exhibit B.2. Example Intel Math Problem Set**



**Unit 5:** Operations with Fractions **Session 3:** Multiplication

**Problem Set 2: Multiplying Fractions Geometrically**

**Instructions**
Start from a unit square, which you can think of as the "whole" for the first four problems. When you are counting up fractional pieces, keep in mind what portion of the figure is the "whole."

**Problems**
In Problems 1 through 4, perform the multiplication calculations using the area model.

1. $\frac{3}{5} \times \frac{2}{3} =$ ________

2. $\frac{3}{4} \times 2 =$ ________

3. $\frac{2}{5} \times 1\frac{1}{4} =$ ________

4. $1\frac{2}{3} \times 1\frac{3}{4} =$ ________

In Problems 5 and 6, use the distributive property to perform the calculation. This provides a second method for making the calculations in Problems 3 and 4.

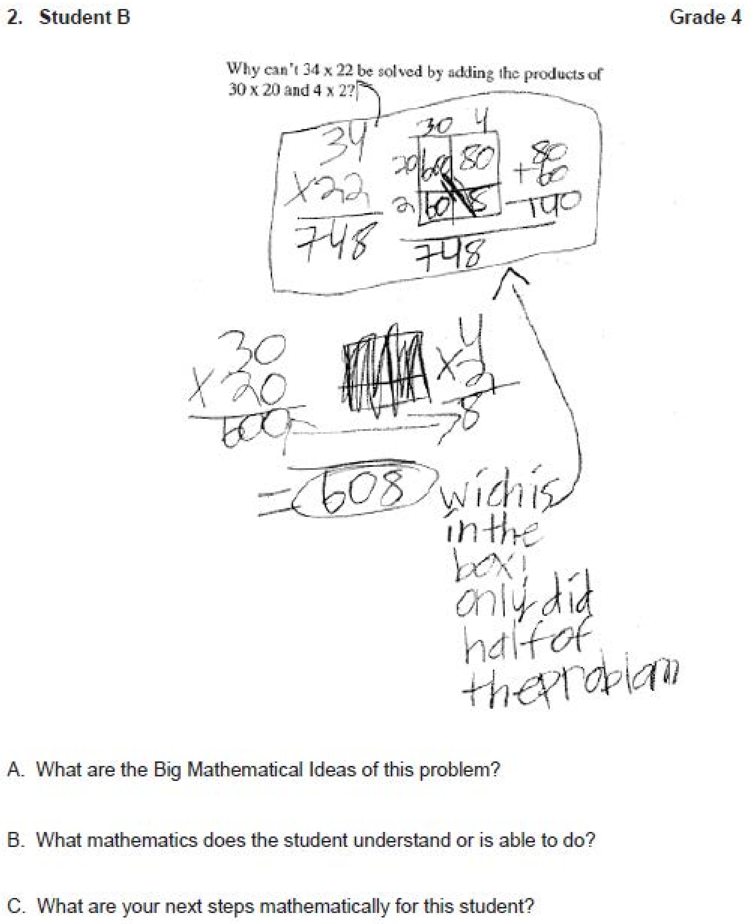
5. $\frac{2}{5}(1 + \frac{1}{4}) =$ ________

6. $(1 + \frac{2}{3}) \times (1 + \frac{3}{4}) =$ ________

Source: Intel Math materials (Intel Foundation, 2009).

Exhibit B.3 includes an exercise set from Unit 3 of Intel Math. Exercise sets appear once per unit and provide opportunities for teachers to analyze student work samples.

**Exhibit B.3. Example Exercise Set**



**2. Student B** **Grade 4**

Why can't 34 x 22 be solved by adding the products of 30 x 20 and 4 x 2?

$$\begin{array}{r} 34 \\ \times 22 \\ \hline 748 \end{array}$$

| | 30 | 4 |
|---|---|---|
| 20 | 600 | 80 |
| 2 | 60 | 8 |

$$\begin{array}{r} 80 \\ +60 \\ \hline 140 \end{array}$$

748

$$\begin{array}{r} 30 \\ \times 20 \\ \hline 600 \end{array} \qquad \begin{array}{r} 4 \\ \times 2 \\ \hline 8 \end{array}$$

= 608 wich is in the box! only did half of theproblam

A. What are the Big Mathematical Ideas of this problem?

B. What mathematics does the student understand or is able to do?

C. What are your next steps mathematically for this student?

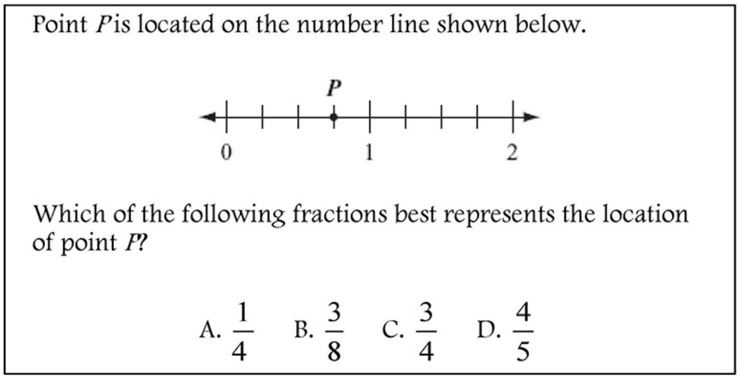
Source: Intel Math materials (Intel Foundation, 2009).

## Mathematics Learning Community

Exhibits B.4 and B.5 include examples of two common types of materials used in the Mathematics Learning Community: math metacognition problems and student work examples. Exhibit B.4 illustrates a math metacognition problem from the Representing and Interpreting Fractions meeting.

**Exhibit B.4. Example Math Metacognition Problem**

**Solve the problem**



Point *P* is located on the number line shown below.

Which of the following fractions best represents the location of point *P*?

A. $\frac{1}{4}$ B. $\frac{3}{8}$ C. $\frac{3}{4}$ D. $\frac{4}{5}$

**Discuss** each of the four possible answer choices:

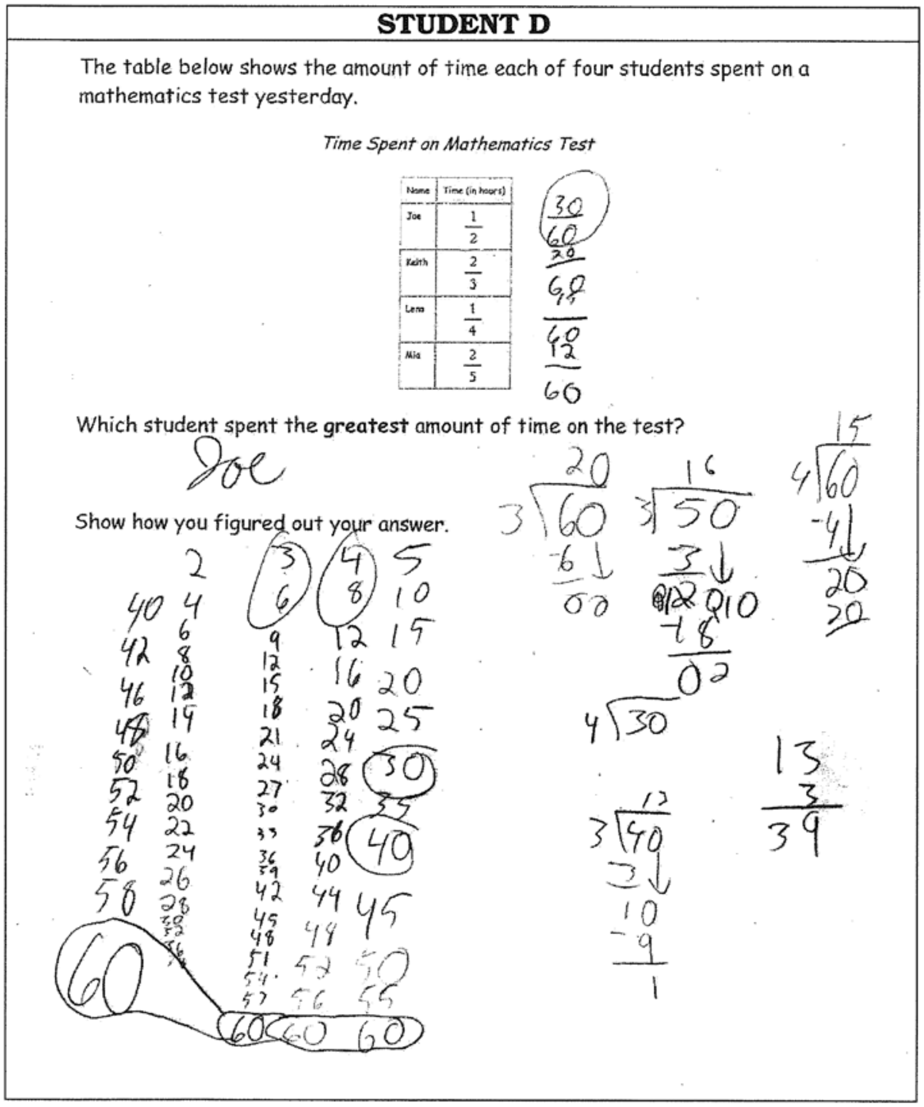
a. What would be the reasoning behind choices A, B and D?

| | |
|---|---|
| Choice **A:** $\frac{1}{4}$<br>Reasoning behind this choice:<br>▪ P is ¼ unit away from 1.<br>▪ Intervals are marked at ¼ units. | Choice **B:** $\frac{3}{8}$<br>Reasoning behind this choice:<br>▪ P is 3/8 of the way to 2.<br>▪ Consider 2 to be the whole, rather than 1. |
| Choice **D:** $\frac{4}{5}$<br>Reasoning behind this choice:<br>▪ 5 tic marks from 0 to 1.<br>▪ P is on the 4th tic mark.<br>▪ Counting tic marks rather than considering the length as the quantity to be measured. | |

Source: Mathematics Learning Community materials (Regional Science Resource Center at the University of Massachusetts Medical School, 2011).

Exhibit B.5 presents one of four student work samples from Representing and Interpreting Fractions, which teachers analyze and discuss during the meeting.

**Exhibit B.5. Example Student Work Analysis**



STUDENT D

The table below shows the amount of time each of four students spent on a mathematics test yesterday.

*Time Spent on Mathematics Test*

| Name | Time (in hours) |
|---|---|
| Joe | $\frac{1}{2}$ |
| Keith | $\frac{2}{3}$ |
| Lena | $\frac{1}{4}$ |
| Mia | $\frac{2}{5}$ |

Which student spent the **greatest** amount of time on the test?

Joe

Show how you figured out your answer.

30/60
20/60
40/60
12/60
2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 60
3 6 9 12 15 18 21 24 27 30 33 36 39 42 45 48 51 54 57 60
4 8 12 16 20 24 28 32 36 40 44 48 52 56 60
5 10 15 20 25 30 35 40 45 50 55 60
40 42 46 48 50 52 54 56 58 60
20
3)60
-6
00
16
3)50
3
012 010
18
02
15
4)60
-4
20
20
4)30
13
3)40
3
10
-9
1
13
3
39

Source: Mathematics Learning Community materials (Regional Science Resource Center at the University of Massachusetts Medical School, 2011).

## Video Feedback Cycles

This section includes excerpts from a feedback form from the Video Feedback Cycles.

*Lesson-level feedback.* The first page of the feedback form included a statement of the lesson goal, a description of the lesson activities, and feedback on the overall mathematical quality of the lesson. Exhibit B.6 illustrates an example of feedback on the latter.

**Exhibit B.6. Example Feedback on Overall Mathematical Quality of the Lesson**

| Overall Mathematical Quality of the Lesson |
|---|
| The teacher clearly explains a rule for adding like fractions, including very good reasoning for why we do not add the denominators. She incorporates the meaning of the numerator and the denominator (in terms of numbers of pieces) in many of her explanations. She also uses visual representations to illustrate examples of adding fractions. She uses very good mathematical vocabulary throughout most of the lesson, including making the point that "decompose" means to break down and not to make smaller. There was one instance where she used a less precise term ("timsing"). The lesson was mostly error-free, except for one error in labeling the units to a solution to a word problem. The teacher showed two methods for adding like fractions, but she did not make many comparisons between the two methods. She also provided good remediation, but it was often more procedural than conceptual. |

Source: Feedback forms.

*Clip-level feedback.* After providing information on the lesson overall, the form continues with clip-level feedback. The rater provides identifying information regarding the clip, provides the codes the clip illustrates, describes the strengths/limitations that the clip illustrates, and gives actionable suggestions for improving this part of the lesson. The facilitator reviews the feedback and adds connections to Intel Math and Math Learning Community Materials. Exhibit B.7 illustrates an example of clip-level feedback.

**Exhibit B.7. Example Feedback for a Video Clip**

| | | | |
|---|---|---|---|
| **Clip 1:** Ducks in Rows | | **Time:** 00:22:20–00:25:20 | |
| MQI Code | Strength/Limitation | To Improve This Part of the Lesson | Link to Intel Math/MLC |
| Linking Between Representations | The video shown by the teacher draws strong links between the array of ducks and the repeated addition and the array of ducks and the multiplication, but the teacher only links the repeated addition and multiplication to each other in a very brief way by saying they are the same. | To strengthen the links between the repeated addition and multiplication, the teacher could explicitly point out specific areas of correspondence in the two expressions. For example, count the number of threes in the sum and point out that it corresponds to the factor of 4. | Intel math, Unit 3 Session 1 Information sheet 1 Linking Repeated addition to the Meaning of Multiplication pages 187–189. MLC 1. |
| Multiple Procedures or Solution Methods | The teacher did a very nice job of presenting, through the video, two procedures for finding the total number of ducks. She discussed both procedures in detail and momentarily compared them for efficiency when she said that multiplication is faster and easier. | To make this even more substantial, the two methods could be linked to each other for similarities and differences. See above. | |

Source: Feedback forms.

*Next steps.* At the end of the feedback session, the facilitator and teacher identify next steps. These address ways in which the teacher could respond to the feedback in the current unit as well as in future units of study. Exhibit B.8 provides an example of next steps.

**Exhibit B.8. Example Next Steps**

Next Steps:

Immediate next steps: Remediate students in a more conceptual way focusing on the WHY. When appropriate, ask them to explain their thinking out loud. When the breakdown or misconception is identified, use models and strategies to help correct their thinking.

Future next steps (Division): Relate division to sharing of equal groups. Connect the relationship to multiplication, identifying number of groups and size of groups.

Source: Feedback forms.

# Appendix C. Supplemental Information Regarding the Comparison of Treatment and Control Teachers' Math PD During the Year of the Study (Service Contrast)

This appendix provides additional information about the "service contrast" or the degree to which the PD experiences of treatment and control teachers in the analytic sample varied during the year of the evaluation.

**Exhibit C.1. Percentage of Teachers Participating in Different Types of Mathematics-Related PD During Summer 2013 and the 2013–14 School Year**

| Mathematics-Related PD Activities | Treatment Group (Percent) | Control Group (Percent) | Estimated Difference (Percent) | P value |
|---|---|---|---|---|
| Attending traditional course/workshop/seminars | 100.0 | 25.1 | 74.9* | <0.001 |
| Participating in mathematics-related structured study groups | 94.9 | 13.8 | 81.1* | <0.001 |
| Being observed in class and given feedback on teaching | 96.2 | 40.6 | 55.6* | <0.001 |
| Participating in collaborative planning activities | 69.6 | 67.5 | 2.1 | 0.770 |
| Participating in other types of PD | 26.6 | 19.6 | 7.0 | 0.216 |

Note: Sample size = 73 schools; 79 treatment teachers and 86 control teachers.

The analyses are based on teacher-level linear probability models controlling for school fixed effects. The percentages for the treatment group are unadjusted. The percentages for the control group were computed based on the unadjusted treatment group percentages and estimated group differences.

* Difference between the treatment teacher percentage and the control teacher percentage is statistically significant at the 0.05 level, two-tailed test.

Source: Spring 2014 Teacher Survey.

**Exhibit C.2. Number of Hours of Mathematics-Related PD in Which Teachers Participated During Summer 2013 and the 2013–14 School Year**

| Mathematics-Related PD Activities | Treatment Group (Median) | Control Group (Median) | Estimated Difference (H L Estimate) | P value |
|---|---|---|---|---|
| Attending traditional course/workshop/seminars | 80.5 | 0.0 | 80.5* | <0.001 |
| Participating in mathematics-related structured study groups | 11.1 | 0.0 | 11.1* | <0.001 |
| Being observed in class and given feedback on teaching | 3.5 | 0.0 | 3.5* | <0.001 |
| Participating in collaborative planning activities | 9.8 | 10.1 | -0.3 | 0.808 |
| Participating in other types of PD | 0.0 | 0.0 | 0.0 | 0.592 |

Note: Sample size = 73 schools; 79 treatment teachers and 86 control teachers.

The analyses are based on the aligned rank sum test developed by Hodges and Lehmann (1963). The medians for the control group are unadjusted. The medians for the treatment group were computed based on the unadjusted control group medians and estimated group differences.[50]

* Difference between the median treatment teacher hours and the median control teacher hours is statistically significant at the 0.05 level, two-tailed test.

Source: Spring 2014 Teacher Survey.

[50] The Hodges and Lehmann method computes the treatment-control difference for each possible pair that consists of one treatment unit and one control unit, and identifies the median of all pairwise differences as the overall treatment effect estimate. Thus, strictly speaking, the Hodges and Lehmann estimates represent "median differences," although they are commonly interpreted as "differences in median."

### Exhibit C.3. Teachers' Ratings of Their Experiences in Different Types of Mathematics-Related PD Activities, for Teachers Who Participated in PD, by Treatment Condition

| Features of PD | Treatment Group (Mean) | Control Group (Mean) | Estimated Difference (Mean) | P value |
|---|---|---|---|---|
| **Traditional course/workshop/seminar** | | | | |
| Frequency of mathematical and student thinking activities in *traditional PD* | 3.4 | 2.3 | 1.1* | <0.001 |
| Frequency K–8 math topics were focus of *traditional PD* | 3.0 | 2.1 | 0.9* | <0.001 |
| Coherence of *traditional PD* with goals, materials, and expectations | 3.4 | 3.2 | 0.2 | 0.186 |
| **Mathematics-related structured study groups** | | | | |
| Frequency of mathematical and student thinking activities in *structured study group* | 3.4 | 2.1 | 1.3* | <0.001 |
| Frequency K–8 math topics were focus of *structured study groups* | 2.9 | 2.1 | 0.8* | <0.001 |
| Coherence of *structured study groups* with goals, materials, and expectations | 3.4 | 3.1 | 0.3 | 0.161 |
| **Being observed and given feedback on teaching** | | | | |
| Frequency of *lesson feedback* focused on mathematical topics and student thinking | 3.5 | 2.8 | 0.7* | <0.001 |
| **Other PD** | | | | |
| Frequency of mathematical and student thinking activities in *other PD* | 2.4 | 2.4 | 0.0 | 0.906 |
| Frequency K–8 math topics were focus of *other PD* | 2.2 | 2.0 | 0.2 | 0.404 |
| Coherence of *other PD* with goals, materials, and expectations | 3.1 | 3.4 | -0.3 | 0.108 |

Note: Traditional course/workshop/seminar sample size = 73 schools, 79 treatment teachers, and 21 control teachers. Mathematics-related structured study groups sample size = 70 schools, 75 treatment teachers, and 13 control teachers. Being observed and given feedback on teaching sample size = 71 schools, 76 treatment teachers, and 35 control teachers. Other PD sample size = 28 schools, 21 treatment teachers, and 17 control teachers.

The analyses are based on teacher-level regression models without controlling for school fixed effects. The means for the treatment group are unadjusted. The means for the control group were computed based on the unadjusted treatment group means and estimated group differences.

Items comprising the indexes reported here are shown in Exhibit A.12 in Appendix A. Frequency items were based on a scale where 1 = never/rarely, 2 = sometimes, 3 = often, and 4 = most or all of the time. Coherence items were based on a scale where 1 = strongly disagree, 2 = disagree, 3 = agree, 4 = strongly agree.

* Difference between the average treatment teacher rating and the average control teacher rating is statistically significant at the 0.05 level, two-tailed test.

Source: Spring 2014 Teacher Survey.

This page has been left blank for double-sided copying.

# Appendix D. Supporting Exhibits for Impact Analyses

This appendix includes further detail on the impact analyses to supplement the results presented in chapter IV.

### Exhibit D.1. Impact of the PD on Teacher Knowledge Based on the Main Impact Analyses and Sensitivity Analyses

| Sample | | Covariates | Outcome | Treatment Group Mean (SD) | Control Group Mean (SD) | Estimated Difference | P value |
|---|---|---|---|---|---|---|---|
| Main Impact Analyses | | | | | | | |
| (1) | Impact Sample | All | Fall | 0.43<br>(1.13) | −0.20<br>(1.00) | 0.63* | <0.001 |
| | | | Spring | 0.34<br>(1.28) | −0.21<br>(1.00) | 0.55* | <0.001 |
| Sensitivity Analyses [a] | | | | | | | |
| (2) | Impact Sample | School fixed effects | Fall | 0.43<br>(1.13) | −0.02<br>(1.00) | 0.45* | 0.006 |
| | | | Spring | 0.34<br>(1.28) | −0.03<br>(1.00) | 0.37* | 0.036 |
| (3) | All teachers with data | All | Fall | 0.46<br>(1.14) | −0.14<br>(1.00) | 0.60* | <0.001 |
| | | | Spring | 0.34<br>(1.32) | −0.20<br>(1.02) | 0.54* | <0.001 |
| (4) | Impact Sample, excluding District 2 | All | Fall | 0.50<br>(1.10) | −0.14<br>(1.00) | 0.64* | <0.001 |
| | | | Spring | 0.38<br>(1.25) | −0.15<br>(1.00) | 0.53* | <0.001 |

Note: Sample size for (1) and (2) = 73 schools, 79 treatment teachers, and 86 control teachers. Sample size for (3) = 92 schools, 96 treatment teachers, and 113 control teachers. Sample size for (4) = 65 schools, 69 treatment teachers, and 75 control teachers.

Analyses (1), (3), and (4) are based on a teacher-level regression controlling for school fixed effects and teacher background characteristics; analysis (2) is based on a teacher-level regression controlling only for school fixed effects. (See Exhibit A.15 in Appendix A for details about the main impact analyses of teacher knowledge.)

[a] The first sensitivity analysis is based on an impact model that included school fixed effects but no other covariates. The second sensitivity analysis is based on an expanded sample that included all teachers with knowledge scores, not just teachers in the impact sample. The third sensitivity analysis excluded District 2 where there were substantial differences in teacher knowledge and student achievement between the treatment and control groups at baseline.

The means for the treatment group are unadjusted. The means for the control group were computed based on the unadjusted treatment group means and estimated group differences.

* Difference between the average treatment teacher score and the average control teacher score is statistically significant at the 0.05 level, two-tailed test.

Source: Fall 2013 and Spring 2014 Teacher Knowledge Tests.

## Exhibit D.2. Impact of the PD on Fall 2013 Classroom Practice Based on the Main Impact Analyses and Sensitivity Analyses

| | Sample | Covariates | Outcome | Treatment Group Mean (SD) | Control Group Mean (SD) | Estimated Difference | P value |
|---|---|---|---|---|---|---|---|
| Main Impact Analyses | | | | | | | |
| (1) | Impact Sample | All | Richness | −0.33<br>(0.66) | −0.49<br>(0.61) | 0.16 | 0.113 |
| | | | Student Participation | −0.20<br>(0.77) | −0.49<br>(0.97) | 0.29* | 0.029 |
| | | | Errors | −0.36<br>(1.07) | −0.43<br>(1.17) | 0.07 | 0.521 |
| Sensitivity Analyses | | | | | | | |
| (2) | Impact Sample | School fixed effects | Richness | −0.33<br>(0.66) | −0.44<br>(0.61) | 0.11 | 0.261 |
| | | | Student Participation | −0.20<br>(0.77) | −0.42<br>(0.97) | 0.22 | 0.076 |
| | | | Errors | −0.36<br>(1.07) | −0.51<br>(1.17) | 0.15 | 0.220 |
| (3) | Impact Sample, excluding District 2 | All | Richness | −0.38<br>(0.64) | −0.51<br>(0.63) | 0.13 | 0.239 |
| | | | Student Participation | −0.21<br>(0.71) | −0.48<br>(1.00) | 0.27 | 0.058 |
| | | | Errors | −0.41<br>(1.11) | −0.48<br>(1.14) | 0.07 | 0.553 |

Note: Sample size for (1) and (2) = 73 schools; 79 teachers, 79 lessons, and 708 7.5-minute segments for the treatment group; 86 teachers, 86 lessons, and 739 7.5-minute segments for the control group. Sample size for (3) = 65 schools; 69 teachers, 69 lessons, and 634 7.5-minute segments for the treatment group; 75 teachers, 75 lessons, and 654 7.5-minute segments for the control group.

The analyses for the *Richness of Mathematics* and *Student Participation in Mathematics* dimensions are based on a two-level model (segments within teachers), and the analyses for the *Errors and Imprecision* dimension are based on a teacher-level regression. Analyses (1) and (3) controlled for school fixed effects and covariates at the segment and teacher levels; analysis (2) controlled only for school fixed effects. (See Exhibit A.16 in Appendix A for details about the main impact analyses of fall 2013 classroom practice.)

Analysis (3) excluded District 2 where there were substantial differences in teacher knowledge and student achievement between the treatment and control groups at baseline.

The means for the treatment group are unadjusted. The means for the control group were computed based on the unadjusted treatment group means and estimated group differences.

* Difference between the average treatment teacher score and the average control teacher score is statistically significant at the 0.05 level, two-tailed test.

Source: MQI scores of video-recorded lessons from fall 2013 (one per teacher).

### Exhibit D.3. Impact of the PD on Spring 2014 Classroom Practice Based on the Main Impact Analyses and Sensitivity Analyses

| | Sample | Covariates | Outcome | Treatment Group Mean (SD) | Control Group Mean (SD) | Estimated Difference | P value |
|---|---|---|---|---|---|---|---|
| Main Impact Analyses | | | | | | | |
| (1) | Impact Sample | All | Richness | 0.24<br>(0.47) | −0.07<br>(0.50) | 0.31* | <0.001 |
| | | | Student Participation | 0.04<br>(0.62) | −0.10<br>(0.71) | 0.14 | 0.160 |
| | | | Errors | −0.34<br>(0.92) | −0.15<br>(0.86) | −0.19 | 0.160 |
| Sensitivity Analyses | | | | | | | |
| (2) | Impact Sample | School fixed effects | Richness | 0.24<br>(0.47) | −0.06<br>(0.50) | 0.30* | <0.001 |
| | | | Student Participation | 0.04<br>(0.62) | −0.08<br>(0.71) | 0.12 | 0.185 |
| | | | Errors | −0.34<br>(0.92) | −0.21<br>(0.86) | −0.13 | 0.334 |
| (3) | Impact Sample, excluding District 2 | All | Richness | 0.23<br>(0.46) | −0.08<br>(0.50) | 0.31* | <0.001 |
| | | | Student Participation | 0.07<br>(0.61) | −0.08<br>(0.71) | 0.15 | 0.150 |
| | | | Errors | −0.36<br>(0.95) | −0.18<br>(0.87) | −0.18 | 0.204 |

Note: Sample size for (1) and (2) = 73 schools; 79 teachers, 158 lessons, and 1,277 7.5-minute segments for the treatment group; 86 teachers, 172 lessons, and 1,352 7.5-minute segments for the control group. Sample size for (3) = 65 schools; 69 teachers, 138 lessons, and 1,137 7.5-minute segments for the treatment group; 75 teachers, 150 lessons, and 1,217 7.5-minute segments for the control group.

The analyses for the *Richness of Mathematics* and *Student Participation in Mathematics* dimensions are based on a three-level model (segments within lessons within teachers), and the analyses for the *Errors and Imprecision* dimension are based on a two-level model (lessons within teachers). Analyses (1) and (3) controlled for school fixed effects and covariates at the segment, lesson, and teacher levels as appropriate; analysis (2) controlled only for school fixed effects. (See Exhibit A.17 in Appendix A for details about the main impact analyses of spring 2014 classroom practice.)

Analysis (3) excluded District 2 where there were substantial differences in teacher knowledge and student achievement between the treatment and control groups at baseline.

The means for the treatment group are unadjusted. The means for the control group were computed based on the unadjusted treatment group means and estimated group differences.

* Difference between the average treatment teacher score and the average control teacher score is statistically significant at the 0.05 level, two-tailed test.

Source: MQI scores of video-recorded lessons from spring 2014 (two per teacher).

**Exhibit D.4. Differential Impact of PD on Teacher Knowledge and Practice Outcomes for Teachers with Different Baseline Math Knowledge**

| Outcome Measures | Estimate | P value |
|---|---|---|
| Teacher Knowledge | | |
| Fall Teacher Knowledge Standardized RIT Score | 0.03 | 0.761 |
| Spring Teacher Knowledge Standardized RIT Score | 0.43* | 0.003 |
| Classroom Practice (MQI Scores) | | |
| Fall *Richness of Mathematics* | 0.05 | 0.707 |
| Spring *Richness of Mathematics* | −0.10 | 0.280 |
| Fall *Student Participation in Mathematics* | −0.01 | 0.973 |
| Spring *Student Participation in Mathematics* | 0.02 | 0.874 |
| Fall *Errors and Imprecision* | −0.11 | 0.434 |
| Spring *Errors and Imprecision* | 0.01 | 0.961 |

Note: Sample size = 73 schools; 79 treatment teachers and 86 control teachers.

The analyses for teacher knowledge outcomes are based on a teacher-level regression controlling for school fixed effects and teacher background characteristics (see Exhibit A.19 in Appendix A). The fall analyses for classroom practice outcomes are based on a two-level model for *Richness of Mathematics* and *Student Participation in Mathematics* and a teacher-level regression for *Errors and Imprecision*. The spring analyses for classroom practice outcomes are based on a three-level model for *Richness of Mathematics* and *Student Participation in Mathematics* and a two-level model for *Errors and Imprecision*. All analyses of classroom practice outcomes controlled for school fixed effects and covariates at the segment, lesson, and student levels as appropriate.

The estimate of differential impact represents the difference in the impact of the PD between teachers whose baseline math knowledge scores differed by 1 standard deviation.

* Difference between the average treatment teacher score and the average control teacher score is statistically significant at the 0.05 level, two-tailed test.

Source: Baseline, Fall 2013, and Spring 2014 Teacher Knowledge Tests; MQI scores of video-recorded lessons from the 2013–14 school year.

**Exhibit D.5. Differential Impact of PD on Teacher Practice Outcomes for Teachers with Different Average Classroom Prior Math Achievement**

| MQI Dimensions | Estimate | P value |
|---|---|---|
| *Richness of Mathematics* | | |
| Fall | −0.69* | 0.003 |
| Spring | −0.22 | 0.255 |
| *Student Participation in Mathematics* | | |
| Fall | −0.72* | 0.032 |
| Spring | −0.39 | 0.145 |
| *Errors and Imprecision* | | |
| Fall | 0.53 | 0.067 |
| Spring | 0.36 | 0.321 |

Note: Sample size = 73 schools, 79 treatment teachers, and 86 control teachers.

The fall analyses for classroom practice outcomes are based on a two-level model for *Richness of Mathematics* and *Student Participation in Mathematics* and a teacher-level regression for *Errors and Imprecision*. The spring analyses for classroom practice outcomes are based on a three-level model for *Richness of Mathematics* and *Student Participation in Mathematics* and a two-level model for *Errors and Imprecision*. All analyses of classroom practice outcomes controlled for school fixed effects and covariates at the segment, lesson, and student levels as appropriate.

The estimate of differential impact represents the difference in the impact of the PD between teachers whose classroom average prior math achievement scores differed by 1 standard deviation.

* Difference between the average treatment teacher score and the average control teacher score is statistically significant at the 0.05 level, two-tailed test.

Source: District administrative records; MQI scores of video-recorded lessons from the 2013–14 school year.

## Exhibit D.6. Impact of the PD on Students' Grade 4 Mathematics Achievement Based on the Main Impact Analyses and Sensitivity Analyses

| | Sample | Covariates | Outcome | Treatment Group Mean (SD) | Control Group Mean (SD) | Estimated Difference | P value |
|---|---|---|---|---|---|---|---|
| Main Impact Analyses | | | | | | | |
| (1) | Impact Sample | All | NWEA Test | −0.09<br>(0.94) | −0.04<br>(1.00) | −0.05 | 0.217 |
| | | | State Assessment | −0.10<br>(0.95) | −0.04<br>(1.00) | −0.06* | 0.024 |
| Sensitivity Analyses | | | | | | | |
| (2) | Impact Sample | School fixed effects | NWEA Test | −0.09<br>(0.94) | −0.04<br>(1.00) | −0.05 | 0.396 |
| | | | State Assessment | −0.10<br>(0.95) | −0.01<br>(1.00) | −0.09 | 0.105 |
| (3) | All students with data | All | NWEA Test | −0.02<br>(0.95) | 0.03<br>(1.00) | −0.05 | 0.195 |
| | | | State Assessment | −0.11<br>(1.01) | −0.07<br>(1.06) | −0.04 | 0.155 |
| (4) | Impact Sample, excluding District 2 | All | NWEA Test | −0.03<br>(0.99) | 0.00<br>(1.00) | −0.03 | 0.521 |
| | | | State Assessment | −0.04<br>(1.02) | 0.01<br>(1.00) | −0.05 | 0.124 |

Note: Size of Samples (1) and (2) for the NWEA Test = 79 teachers and 806 students in the treatment group; 86 teachers and 891 students in the control group. Size of Samples (1) and (2) for the State Assessment = 79 teachers and 1,760 students in the treatment group; 86 teachers and 1,917 students in the control group. Size of Sample (3) for the NWEA Test = 97 teachers and 978 students in the treatment group; 109 teachers and 1,147 students in the control group. Size of Sample (3) for the State Assessment = 97 teachers and 2,161 students in the treatment group; 110 teachers and 2,419 students in the control group. Size of Sample (4) for the NWEA Test = 75 teachers and 695 students in the treatment group; 69 teachers and 764 students in the control group. Size of Sample (4) for the State Assessment = 69 teachers and 1,513 students in the treatment group; 75 teachers and 1,659 students in the control group.

All analyses are based on a two-level model controlling for school fixed effects. Analyses (1), (3), and (4) further controlled for student characteristics. (See Exhibit A.18 in Appendix A for details about the main impact analyses of student achievement.)

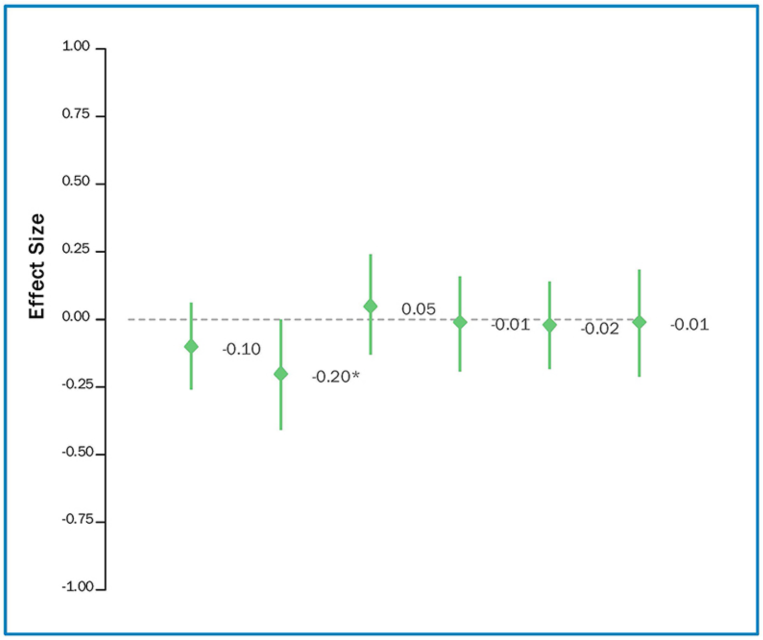
The first sensitivity analysis is based on an impact model that included school fixed effects but no other covariates. The second sensitivity analysis is based on an expanded sample that included all grade 4 students with math achievement data in the classrooms of teachers who were randomly assigned and still taught grade 4 in the study year, not just students in the classrooms of teachers in the impact sample. The third sensitivity analysis excluded District 2 where there were substantial differences in teacher knowledge and student achievement between the treatment and control groups at baseline.

The means for the treatment group are unadjusted. The means for the control group were computed based on the unadjusted treatment group means and estimated group differences.

* Difference between the average treatment student score and the average control student score is statistically significant at the 0.05 level, two-tailed test.

Source: Spring 2014 NWEA Test; District administrative records.

**Exhibit D.7. Impact of the PD on Student Mathematics Achievement on the NWEA Test, by District**



Note: Sample size = 73 schools; 79 teachers and 806 students in the treatment group; 86 teachers and 891 students in the control group.

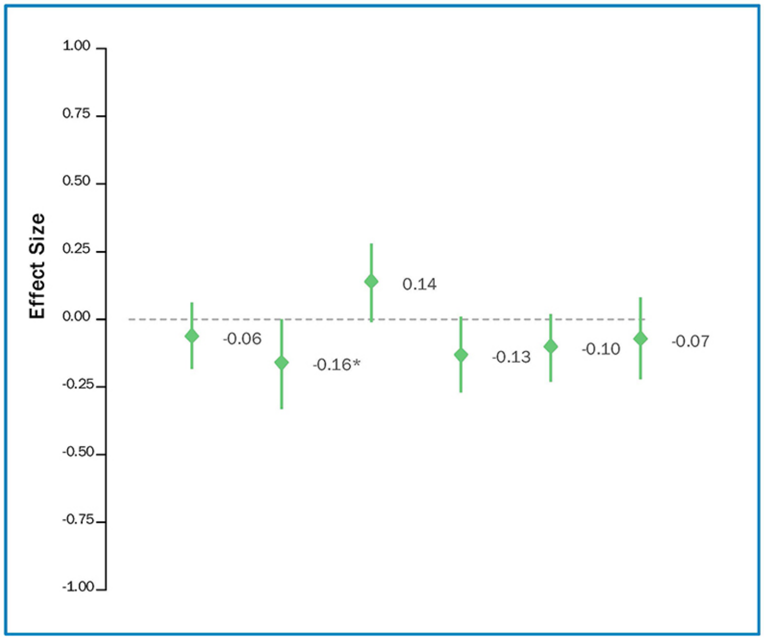
The analyses are based on a two-level model controlling for school fixed effects and student characteristics (see Exhibit A.18 in Appendix A).

The impact estimates are expressed as effect sizes, representing the mean differences between treatment and control students in the control group standard deviation unit of the outcome. The vertical bars represent the 95 percent confidence intervals for the impact estimates.

* Effect size is statistically significant at the 0.05 level, two-tailed test.

Source: Spring 2014 NWEA Test.

**Exhibit D.8. Impact of the PD on Student Mathematics Achievement on the State Assessment, by District**



Note: Sample size = 73 schools; 79 teachers and 1,760 students in the treatment group; 86 teachers and 1,917 students in the control group.

The analyses are based on a two-level model controlling for school fixed effects and student characteristics (see Exhibit A.18 in Appendix A).

The impact estimates are expressed as effect sizes, representing the mean differences between treatment and control students in the control group standard deviation unit of the outcome. The vertical bars represent the 95 percent confidence intervals for the impact estimates.

* Effect size is statistically significant at the 0.05 level, two-tailed test.

Source: District administrative records.

### Exhibit D.9. Differential Impact of PD on Student Mathematics Achievement

| Student Achievement Outcome | Moderators | Estimate | P value |
|---|---|---|---|
| NWEA Test | Teacher baseline knowledge | −0.01 | 0.902 |
| | Teacher experience | 0.06 | 0.601 |
| | Classroom average prior achievement | −0.08 | 0.376 |
| | Student prior (grade 3) math achievement | −0.04 | 0.303 |
| State Assessment | Teacher baseline knowledge | −0.03 | 0.514 |
| | Teacher experience | 0.10 | 0.220 |
| | Classroom average prior achievement | 0.03 | 0.661 |
| | Student prior (grade 3) math achievement | 0.01 | 0.679 |

Note: NWEA test sample size = 73 schools; 79 treatment teachers and 86 control teachers; 806 treatment students and 891 control students. State assessment sample size = 73 schools; 79 treatment teachers and 86 control teachers; 1,760 treatment students and 1,917 control students.

The analyses are based on a two-level model controlling for school fixed effects and student characteristics.

The estimate of differential impact represents the difference in the impact of the PD between students whose values on the moderator differed by 1 unit.

None of the differences between students whose value on the moderator differed by 1 unit is statistically significant at the 0.05 level, two-tailed test.

Source: Spring 2014 NWEA Test; District administrative records; Baseline Teacher Knowledge Test; Spring 2014 Teacher Survey.

# Appendix E. Supporting Exhibits for Correlational Analyses

This appendix includes additional results from the correlational analyses to supplement the results presented in chapter V.

**Exhibit E.1. Estimated Relationships Between Measures of Fall 2013 Teacher Knowledge and Spring 2014 Instructional Practice**

| Measure of Instructional Practice | Estimate | P value |
|---|---|---|
| Richness of Mathematics | 0.35* | <0.001 |
| Student Participation in Mathematics | 0.19* | 0.025 |
| Errors and Imprecision | −0.31* | <0.001 |

Note: Sample size = 73 schools; 79 teachers and 1,760 students in the treatment group; 86 teachers and 1,917 students in the control group.

The *Richness of Mathematics* and *Student Participation in Mathematics* analyses are based on a three-level model (segments within lessons within teachers), controlling for school fixed effects and covariates at the segment, lesson, and teacher levels. The *Errors and Imprecision* analysis is based on a two-level model (lessons within teachers), controlling for school fixed effects and covariates at the lesson and teacher levels. The estimate from each analysis represents the change in the standardized teacher score for a given practice measure per 1 standard deviation increase in the fall teacher knowledge score.

* Association is statistically significant at the 0.05 level, two-tailed test.

Source: Fall 2013 Teacher Knowledge Tests; MQI scores of video-recorded lessons from the 2013–14 school year.

**Exhibit E.2. Estimated Relationships Between Measures of Teacher Knowledge and Instructional Practice and Grade 4 Math Score on Study-Administered Assessment**

| Teacher Outcome | | Model | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 |
| Teacher Knowledge | | | | | | | |
| Teacher Knowledge | *Estimate* | 0.02 | | | | | 0.00 |
| | *P-value* | 0.466 | | | | | 0.863 |
| Instructional Practice | | | | | | | |
| Richness of Mathematics | *Estimate* | | −0.04 | | | −0.04 | −0.05 |
| | *P-value* | | 0.460 | | | 0.545 | 0.469 |
| Student Participation in Mathematics | *Estimate* | | | 0.01 | | −0.00 | −0.02 |
| | *P-value* | | | 0.871 | | 0.974 | 0.727 |
| Errors and Imprecision | *Estimate* | | | | −0.14* | −0.18* | −0.20* |
| | *P-value* | | | | 0.027 | 0.007 | 0.004 |

Note: Sample size = 73 schools; 79 teachers and 806 students in the treatment group; 86 teachers and 891 students in the control group.

The analyses predict students' grade 4 math scores on the study-administered Northwest Evaluation Association (NWEA) assessment with teacher outcomes based on a two-level model controlling for school fixed effects and student characteristics. The main predictor is the average of a teacher's fall and spring knowledge scores for Model 1, and a teacher-level measure of instructional practice for Models 2–4. Model 5 includes all three measures of instructional practice as predictors, and Model 6 further includes the average of a teacher's fall and spring knowledge score as a predictor. All teacher-level measures of knowledge and practice were standardized using the control group mean and standard deviation. The estimate from each analysis represents the change in the standardized student achievement scores per 1 standard deviation increase in the predictor.

* Association is statistically significant at the 0.05 level, two-tailed test.

Source: Spring 2014 NWEA Test; Fall 2013 and Spring 2014 Teacher Knowledge Tests; MQI scores of video-recorded lessons from the 2013–14 school year.

## Exhibit E.3. Estimated Relationships Between Measures of Teacher Knowledge and Instructional Practice and Grade 4 Math Score on State Assessment

| | | Model | | | | | |
|---|---|---|---|---|---|---|---|
| Teacher Outcome | | 1 | 2 | 3 | 4 | 5 | 6 |
| Teacher Knowledge | | | | | | | |
| Teacher Knowledge | *Estimate* | −0.01 | | | | | −0.02 |
| | *P-value* | 0.716 | | | | | 0.277 |
| Instructional Practice | | | | | | | |
| Richness of Mathematics | *Estimate* | | −0.06 | | | −0.06 | −0.04 |
| | *P-value* | | 0.169 | | | 0.337 | 0.489 |
| Student Participation in Mathematics | *Estimate* | | | −0.04 | | −0.02 | −0.04 |
| | *P-value* | | | 0.290 | | 0.639 | 0.456 |
| Errors and Imprecision | *Estimate* | | | | −0.14* | −0.18* | −0.21* |
| | *P-value* | | | | 0.013 | 0.002 | 0.001 |

Note: Sample size = 73 schools; 79 teachers and 1,760 students in the treatment group; 86 teachers and 1,917 students in the control group.

The analyses predict students' grade 4 math scores on state assessment with teacher outcomes based on a two-level model controlling for school fixed effects and student characteristics. The main predictor is the average of a teacher's fall and spring knowledge scores for Model 1, and a teacher-level measure of instructional practice for Models 2–4. Model 5 includes all three measures of instructional practice as predictors, and Model 6 further includes the average of a teacher's fall and spring knowledge score as a predictor. All teacher-level measures of knowledge and practice were standardized using the control group mean and standard deviation. The estimate from each analysis represents the change in the standardized student achievement scores per 1 standard deviation increase in the predictor.

* Association is statistically significant at the 0.05 level, two-tailed test.

Source: District administrative records; Fall 2013 and Spring 2014 Teacher Knowledge Tests; MQI scores of video-recorded lessons from the 2013–14 school year.

# References

Agodini, R., Harris, B., Thomas, M., Murphy, R., & Gallagher, L. (2010). *Achievement effects of four early elementary school math curricula: Findings for first and second graders* (NCEE 2011-4001). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Allen, J., Hafen, C., Gregory, A., Mikami, A. Y., & Pianta, R. (2015). Enhancing secondary school instruction and student achievement: Replication and extension of the My Teaching Partner-Secondary intervention. *Journal of Research on Educational Effectiveness, 8*(2), 1–15.

Allen, J. P., Pianta, R. C., Gregory, A., Mikami, A. Y., & Lun, J. (2011). An interaction-based approach to enhancing secondary school instruction and student achievement. *Science, 333,* 1034–1037.

Ball, D. L., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education, 59*(5), 389–407.

Birman, B., Boyle, A., Le Floch, K. C., Elledge, A., Holtzman, D., Song, M., et al. (2009). *State and local implementation of the No Child Left Behind Act, Volume VIII—Teacher quality under* NCLB: *Final Report*. Washington, DC: Policy and Program Studies Service, Office of Planning, Evaluation, and Policy Development, U.S. Department of Education.

Blank, R. K., & de las Alas, N. (2009). *Effects of teacher professional development on gains in student achievement: How meta-analysis provides scientific evidence useful to education leaders.* Washington, DC: Council of Chief State School Officers.

Blazar, D. (2015). Effective teaching in elementary mathematics: Identifying classroom practices that support student achievement. *Economics of Education Review, 48,* 16–29.

Borko, H. (2004). Professional development and teacher learning: Mapping the terrain. *Educational Researcher, 33*(8), 3–15.

Burmester, M., & Wu, H. H. (2004). *Some lessons from California.* Berkeley, CA: University of California.

Carpenter, T., Fennema, E., Peterson, P., 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), 499–531.

Cheong, Y. F., & Raudenbush, S. W. (2000). Measurement and structural models for children's problem behaviors. *Psychological Methods, 5*(4), 477–495.

Colby, G. T., Boston, M., & Smith, T. (2011). *Examining relationships between instructional quality and student achievement in middle-grades mathematics.* Paper presented at the fall conference of the Society for Research on Educational Effectiveness (SREE), Washington, DC.

Conference Board of the Mathematical Sciences. (2012). *The mathematical education of teachers II*. Providence, RI, & Washington, DC: American Mathematical Society and Mathematical Association of America.

Desimone, L., & Stuckey, D. (2014). Sustaining professional development. In. L. Martin, S. Kragler, D. Quatroche, & K. Bauserman (Eds.), *Handbook of professional development in education: Successful models and practices, prek-12* (pp. 467–482). New York, NY: Guilford Publications.

Garet, M., Wayne, A., Stancavage, F., Taylor, J., Eaton, M., Walters, K., et al. (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.

Garet, M., Wayne, A., Stancavage, F., Taylor, J., Walters, K., Song, M., et al. (2010). *Middle school mathematics professional development impact study: Findings after the first year of implementation* (NCEE 2010-4009). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

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* (REL 2014–010). Washington, DC: Regional Educational Laboratory Southeast, National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Glancy, E., Fulton, M., Anderson, L., Dounay Zinth, J., Millard, M., & Delander, B. (2014). *Blueprint for college readiness.* Denver, CO: Education Commission of the States.

Greenberg, J., & Walsh, K. (2008). *No common denominator: The preparation of elementary school teachers in mathematics by America's education schools*. Washington, DC: National Council on Teacher Quality.

Hammerman, J. K. L., Demers, L. B., & Higgins, T. L. (2015). *Measuring the impact on teaching practice of an innovative elementary math professional development program*. Paper presented at the American Educational Research Association (AERA), Chicago, IL.

Hill, H. C., Blunk, M., Charalambous, C., Lewis, J., Phelps, G. C., Sleep, L., & Ball, D.L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction, 26*(4), 430–511.

Hill, H. C., Kapitula, L., & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. *American Education Research Journal, 48*(3), 794–831.

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

Hodges, J. L., & Lehmann, E. L. (1963). Estimates of location based on rank tests. *The Annals of Mathematical Statistics, 34*(2), 598–611.

Intel Foundation. (2009). *Intel Math, Version 2.5*. Santa Clara, CA: Author.

Jacob, R., Hill, H. C., & Corey, D. (2015). *Investigating the effect of professional development on teachers' knowledge for teaching, instruction, and student achievement.* Arlington, VA: National Science Foundation.

Jacobs, V., Franke, M., Carpenter, T., 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.

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains.* Seattle, WA: Measures of Effective Teaching project, Bill & Melinda Gates Foundation.

Kennedy, M. M. (2016). How does professional development improve teaching? *Review of Educational Research.* Advance online publication.

Kraft, M. A., & Blazar, D. L. (2016). Individualized coaching to improve teacher practice across grades and subjects: New experimental evidence. *Educational Policy*. Advance online publication.

Linacre, J. M. (2014). Facets 3.71.4 [Computer software]. Retrieved from http://www.winsteps.com/facets.htm.

Martin, L., & Umland, K. (2008). Mathematics for middle school teachers: Choices, successes, and challenges. *The Mathematics Enthusiast, 5*(2), 305–314.

Mathematics Instrument Development Group. (2013). *Mathematical quality of instruction.* Cambridge, MA: Author.

May, H., Perez-Johnson, I., Haimson, J., Sattar, S., & Gleason, P. (2009). *Using state tests in education experiments: A discussion of the issues* (NCEE 2009-013). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Mullis, I. V. S., Martin, M. O., Foy, P., & Arora, A. (2012). *TIMSS 2011 international results in mathematics.* Chestnut Hill, MA: Boston College.

Muraki, E., & Bock, R. D. (2003). PARSCALE 4 for Windows [Computer software]. Lincolnwood, IL: Scientific Software International, Inc.

National Center for Education Statistics. (2014). *The nation's report card, a first look: 2013 mathematics and reading* (NCES 2014–451). Washington, DC: Institute of Education Sciences, U.S. Department of Education.

National Center for Education Statistics. (2015). *The nation's report card.* Washington, DC: Institute of Education Sciences, U.S. Department of Education. Retrieved November 4, 2015, from http://www.nationsreportcard.gov/reading_math_2015/#?grade=4.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common core state standards for mathematics.* Washington, DC: Authors.

National Mathematics Advisory Panel. (2008). *Foundations for success: The final report of the National Mathematics Advisory Panel.* Washington, DC: U.S. Department of Education.

National Research Council. (2001). Adding it up: Helping children learn mathematics. In J. Kilpatrick, J. Swafford, & B. Findell (Eds.), *Mathematics learning study committee, center for education, division of behavioral and social sciences, and education.* Washington, DC: National Academies Press.

Organization for Economic Cooperation and Development. (2014). *PISA 2012 results: What students know and can do. Student performance in mathematics, reading, and science.* Paris, France: Author.

Ottmar, R., Rimm-Kaufman, S. E., Larsen, R., & Merritt, E. G. (2011). *Relations between mathematical knowledge for teaching, mathematics instructional quality, and student achievement in the context of responsive (RC) approach.* Paper presented at the fall conference of the Society for Research on Educational Effectiveness (SREE), Washington, DC.

Perry, R. R., & Lewis, C. C. (2011). *Improving the mathematical content base of lesson study: Summary of results.* Oakland, CA: Mills College.

Puma, M., Bell, S., Olsen, R., & Price, C. (2009). *What to do when data are missing in group randomized control trials* (NCEE 2009-0049). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Regional Science Resource Center at the University of Massachusetts Medical School. (2011). *Mathematics Learning Community*. Malden, MA: Commonwealth of Massachusetts, Department of Elementary and Secondary Education.

Rockoff, J. E., Jacob, B. A., Kane, T. J., & Staiger, D. O. (2011). Can you recognize an effective teacher when you recruit one? *Education Finance and Policy,* 6(1), 43–74.

Scher, L. S., & O'Reilly, F. E. (2009). Professional development for K–12 math and science teachers: What do we really know? *Journal of Research on Educational Effectiveness,* 2(3), 209–249.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, *15*(2), 4–14.

Supovitz, J. (2013). *The linking study: An experiment to strengthen teachers' engagement with data on teaching and learning.* Paper presented at the annual meeting of the American Educational Research Association (AERA), San Francisco, CA.

U.S. Department of Education. (2010). *ESEA blueprint for reform*. Washington, DC: Office of Planning, Evaluation, and Policy Development, U.S. Department of Education.

U.S. Department of Education. (2014). *Fiscal year 2014 budget summary and background information.* Washington, DC: U.S. Department of Education.

What Works Clearinghouse. (2014). *WWC procedures and standards handbook* (Version 3.0). Washington, DC: Author.

Wu, H.-H. (2009). What's so sophisticated about elementary mathematics: Plenty—that's why elementary schools need math teachers. *American Educator, 3*(3), 4–14.

Wu, H.-H. (2011). *Understanding numbers in elementary school mathematics*. Providence, RI: American Mathematics Society.

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: Regional Educational Laboratory Southwest, National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.