NCEE 2013-4000

U. S. DEPARTMENT OF EDUCATION

The Impact of the Measures of Academic Progress (MAP) Program on Student Reading Achievement

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

The Impact of the Measures of Academic Progress (MAP) Program on Student Reading Achievement

Final Report

December 2012

Authors:

David Cordray
Vanderbilt University

Georgine Pion
Vanderbilt University

Chris Brandt
REL Midwest

Ayrin Molefe
REL Midwest

Megan Toby
Empirical Education

Project Officer:

Sandra Garcia
Institute of Education Sciences

NCEE 2013–4000
U.S. Department of Education

U.S. Department of Education
Arne Duncan
Secretary

Institute of Education Sciences
John Q. Easton
Director

National Center for Education Evaluation and Regional Assistance
Ruth Curran Neild
Commissioner

December 2012

This report was prepared for the National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, under contract ED-06C0-0019 with Regional Educational Laboratory Midwest administered by Learning Point Associates, an affiliate of the American Institutes for Research.

Disclosure of Potential Conflict of Interest

The research team for this study was based at Regional Educational Laboratory Midwest, administered by Learning Point Associates, an affiliate of the American Institutes for Research. Neither the authors nor Learning Point Associates and its key staff have financial interests that could be affected by the findings of this study.[1]

[1]Contractors carrying out research and evaluation projects for IES frequently need to obtain expert advice and technical assistance from individuals and entities whose other professional work may not be entirely independent of or separable from the tasks they are carrying out for the IES contractor. Contractors endeavor not to put such individuals or entities in positions in which they could bias the analysis and reporting of results, and their potential conflicts of interest are disclosed.

Acknowledgments

The authors thank all the school districts, schools, principals, teachers, and students who participated in this study. This study was only possible because of their sustained commitment to the project.

The authors thank members of the Regional Educational Laboratory- Midwest technical working group for their insights and feedback on the evaluation design: Ellen Goldring, James Heckman, Larry Hedges, Rebecca Maynard, Brian Rowan, and Russ Whitehurst. We also thank Larry Hedges for taking the time to meet individually with the research team several times over the course of this study to provide feedback on our analysis plans and procedures. The authors acknowledge contributions from Matt Dawson for his expert advice during the project planning stage; and Jiuping Chen and Shaheen Khan, who served in key roles with data collection and data cleaning activities. We thank Megan Toby at Empirical Education for her exceptional management of all data collection activities and responsiveness to the needs and requests of the authors. We thank Vicki McCoy from NWEA for her leadership and implementation support during all phases of this project; and John Cronin and the leadership from NWEA for their willingness to allow us to conduct this study.

Contents

Figures

Tables

Box

Executive Summary

During the past decade, the use of standardized benchmark measures to differentiate and individualize instruction for students received renewed attention from educators (Bennett 2002; Public Agenda 2008; Russo 2002). Although teachers may use their own assessments (tests, quizzes, homework, problem sets) for monitoring learning, it is challenging for them to equate performance on classroom measures with likely performance on external measures, such as statewide tests or nationally normed standardized tests. Benchmark measures reflective of such external tests may be more useful in helping teachers make decisions about differentiating instruction, which in turn can lead to gains in student learning, higher scores on state standardized tests, and improvements in schoolwide achievement (Baenen et al. 2006; Baker and Linn 2003).

One of the most widely used commercially available systems incorporating benchmark assessment and training in differentiated instruction is the Northwest Evaluation Association's (NWEA) Measures of Academic Progress (MAP) program. The MAP program includes (1) computer-adaptive assessments administered to students three or four times a year and (2) teacher training and access to MAP resources on how to use data from these assessments to differentiate instruction. MAP tests and training are currently in use in nearly 20 percent of K–12 school districts nationwide and more than a third of districts in the Midwest (http://www.nwea.org/support/article/1339). Although the technical merits and popularity of MAP assessments have been widely referenced in practitioner-oriented journals and teacher magazines (Ash 2008; Clarke 2006; Olson 2007; Russo 2002; Woodfield 2003), few studies have investigated the effects of MAP or other benchmark assessment programs on student outcomes. This study was designed to address questions from Midwestern states and districts about the extent to which benchmark assessment may affect teachers' differentiated instructional practices and student achievement.

Thirty-two elementary schools in five districts in Illinois participated in a two-year randomized controlled trial to assess the effectiveness of the MAP program. Half the schools were randomly assigned to implement the MAP program in grade 4, and the other half were randomly assigned to implement MAP in grade 5. Schools assigned to grade 4 treatment served as the grade 5 control condition, and schools assigned to grade 5 treatment served as the grade 4 control.

The study investigated one primary and two secondary confirmatory research questions:

1. Did the MAP program (that is, training plus formative testing feedback) affect the reading achievement of grade 4 students after Year 2 of implementation, as measured by the Illinois Standards Achievement Test (ISAT) reading scale scores or the MAP composite test scores in reading and language use?
2. Were MAP resources (training, consultation, web-based materials) delivered by NWEA and received and used by teachers as planned?
3. Did MAP teachers apply differentiated instructional practices in their classes to a greater extent than their control counterparts?

The report also addressed one exploratory question:

4. Did the MAP program affect the reading achievement of grade 5 students after Year 2 of implementation, as measured by the Illinois Standards Achievement Test (ISAT) reading scale scores or the MAP composite test scores in reading and language use?

The results of the study indicate that the MAP program was implemented with moderate fidelity but that MAP teachers were not more likely than control group teachers to have applied differentiated instructional practices in their classes. Overall, the MAP program did not have a statistically significant impact on students' reading achievement in either grade 4 or grade 5.

Chapter 1: Introduction and Study Overview

During the past decade, the use of standardized benchmark measures to differentiate and individualize instruction for students has received renewed attention from educators (Bennett 2002; Public Agenda 2008; Russo 2002). Effective differentiation based on prior readiness, interests, and learning profiles requires a valid descriptive dataset at the classroom level (Decker 2003). Although teachers may use their own student-level assessments (for example, tests, quizzes, homework, problem sets) to monitor learning, it is challenging for them to equate performance on classroom measures with likely performance on external measures such as statewide tests or nationally normed standardized tests. Benchmark assessments reflective of such external tests are potentially more useful in helping teachers make decisions about differentiating instruction, which in turn can lead to student learning gains, higher scores on state standardized tests, and improvements in schoolwide achievement (Baenen et al. 2006; Baker and Linn 2003).

Another educational innovation representing a noticeable effort on the part of educators in recent years, often in conjunction with benchmark assessment, is differentiated instruction (McTighe and Brown 2005). In differentiated instruction, individual teachers provide a more personalized instructional experience for students within their classroom (Tomlinson and McTighe 2006). This differentiation is valuable in addressing variations in both ability and preparedness among students within a single classroom group (Tomlinson and McTighe 2006).

Tomlinson (2001) defines differentiated instruction as "A flexible approach to teaching in which the teacher plans and carries out varied approaches to content, process, and product in anticipation of and in response to student differences in readiness, interests, and learning needs" (p. 10). Hall (2002) elaborates on this definition by offering the following characterization: "To differentiate instruction is to recognize students' varying background knowledge, readiness, language, preferences in learning, interests, and to react responsively. Differentiated instruction is a process to approach teaching and learning for students of differing abilities in the same class. The intent of differentiating instruction is to maximize each student's growth and individual success by meeting each student where he or she is, and assisting in the learning process" (p. 2). Beyond these general definitions, in practice, differentiation of instruction has relied on a vague set of techniques that are undefined and situational and that depend heavily on the teacher, the students, and the resources available for responding to intended instructional outcomes and student needs. As a result, differentiated instruction has seen very little research either supporting or refuting the approach.

Differentiation, as commonly instituted, directs teachers to make choices about the specific *content* of what is being taught, the *processes* or instructional strategies (procedures and techniques) that are used, and the nature of the *product* by which students demonstrate their proficiency. These choices are to be based upon student characteristics such as readiness (e.g., prior experience and knowledge), interests, and learning profile (e.g., ability, learning style, cognitive development) (Hall, 2002). These choices result in student grouping or, where necessary, individualized instruction for small numbers of students.

Benchmark, or interim, assessments are tests administered at scheduled times during the year. Teachers can use benchmark tests to evaluate students' progress on a specific set of standards or

benchmarks that students must master to be on track to reach end-of-year learning goals. One of the most widely used commercially available systems incorporating benchmark assessment and training in assessment data use is the Northwest Evaluation Association's (NWEA) Measures of Academic Progress (MAP) program. The MAP program consists of two components: (1) computer-adaptive tests administered three or four times per year and (2) training and online resources for administrators and teachers to understand and use results to differentiate instruction.[2]

MAP tests and training are currently used in nearly 20 percent of K–12 school districts nationwide and more than a third of districts in the Midwest (http://www.nwea.org/support/article/1339). NWEA has produced numerous technical reports describing the reliability and validity of its portfolio of MAP assessments;[3] it also maintains the country's largest repository of data on student growth (Cronin et al. 2007). These features have influenced partnerships between NWEA and education researchers in which MAP assessments have been used as a key data source in studies of educational initiatives.

States across the Midwest Region have considered adopting benchmark assessment to provide schools and teachers with predictive measures for improving instruction and state test results. In 2008 Wisconsin created a task force to examine the utility of introducing formative and benchmark assessment as part of a comprehensive statewide assessment program. In 2008 Indiana introduced the Diagnostic Assessment program, in which teachers from more than 500 schools teaching some 220,000 K–8 students were trained to implement and use results from benchmark tests. Indiana expects that virtually all schools will adopt benchmark assessment programs by 2013/14.

Midwest districts have witnessed increased demand for benchmark assessment, in particular for the MAP program. In 2009 more than 30 percent of districts in the Midwest used MAP assessment, and the number of districts adopting the program continues to increase. The increasing interest in benchmark assessment from Midwest states and the wide use of the MAP program among Midwest districts prompted the REL Midwest to propose an experimental trial of the MAP program in 2007.

Although the technical merits and popularity of MAP assessments have been widely referenced in practitioner-oriented journals and teacher magazines (Ash 2008; Clarke 2006; Olson 2007;

[2] Representatives of NWEA provided comments on a draft of this report. They noted that the MAP program serves multiple purposes within schools and that a "relatively small minority of the partners have implemented the full training program" (as was done in this study). They argue that the MAP program implemented in this study is "but one particular form of MAP implementation" (memo from NWEA, dated February 15, 2011). Researchers believe that although alternative forms of MAP implementation may affect other types of outcomes relevant to school leaders (for example, more consistent assessment practices), the MAP program as implemented in this study is most likely to produce the largest impact on student outcomes, and is therefore more aligned to the main purpose of this study, namely, to assess the impact of the MAP program on student achievement.

[3] NWEA reports that test-retest correlations as well as test correlations between different item pool structures are generally high. The reported range of test-retest correlations with common item pool structures is between .628 and .915 across mathematics, reading, and language usage tests in grades 2–10. The range of test correlations between different item pool structures is between .678 and .920 for correlations reported across these same subjects and grade levels. Both sets of correlations report values that most generally fall between .7 and .9 (NWEA, 2009). The marginal reliability estimates (a measure of internal consistency) for these subject area tests are similarly high. The range of marginal reliabilities is between .614 and .918 with most values ranging between .7 and .9. Concurrent and predictive validity estimates range between .366 and .859 with most values ranging between .65 and .85 (NWEA, 2009).

Russo 2002; Woodfield 2003), few studies have investigated the effects of MAP or other benchmark assessment programs on student outcomes. Research on the effects of formative assessment suggests that it is associated with improvements in student learning (Black and Wiliam 1998; Kingston and Nash 2009; Meisels et al. 2003; Nyquist 2003), particularly among low achievers (Black and Wiliam 1998) and students with learning disabilities (Fuchs and Fuchs 1986).[4]

The formative assessment literature is frequently cited to support the effectiveness of benchmark assessments (Perie, Marion, and Gong 2007). However, the evidence from formative assessment research is limited in its ability to demonstrate the effectiveness of benchmark assessment in three primary ways. First, although a substantial number of these studies used experimental or quasi-experimental designs, many confounded treatments, compared nonequivalent groups, or assigned participants to treatment groups in nonrandom ways. Such constraints jeopardize the validity of their findings (Dunn and Mulvenon 2009; Fuchs and Fuchs 1986). Second, a clear definition and commonly used models of formative assessment have only recently begun to emerge in the literature. Differing and often complex conceptions of the nature of formative assessment have yielded wide variations in the reported effects of formative assessment on student outcomes across studies (Dunn and Mulvenon 2009; Hattie and Timperley 2007). Third, the vast majority of formative assessment practices investigated in these studies focus on classroom-based assessment practices, which are administered much more frequently than benchmark assessments and used to guide classroom instruction on a day-to-day basis (Torgesen and Miller 2009).

Empirical studies investigating the effects of benchmark assessment on student achievement have recently begun to emerge. The results are mixed. Borman, Carlson, and Robinson (2010) report the results of a multistate district-level cluster randomized trial investigating the impact on student achievement of benchmark assessment and consulting services to assist in the interpretation of results. The study collected data through the Center for Data-Driven Reform in Education (CDDRE). The analytic sample included 509 schools across 56 districts in 7 states (Alabama, Arizona, Indiana, Mississippi, Ohio, Pennsylvania, and Tennessee). Results show significant positive effects of the intervention on students' state test scores in mathematics (d = 0.21) but not in reading (d = 0.14; p-value = .10).

The Regional Educational Laboratory Northeast and Islands published two studies that investigate the impact of benchmark assessments on student outcomes (Henderson et al. 2007a, 2007b). The studies find no significant differences in gains in mathematics achievement between schools that used quarterly benchmark exams and schools that did not. Although similar in focus, the two studies differ from the current investigation of the MAP program in at least two important ways. First, these studies focus on the impact of benchmark testing, whereas the

[4] The recent research literature on formative assessment distinguishes between formative assessment and benchmark assessment (Perie, Marion, and Gong 2007; Torgesen and Miller 2009). For the purposes of this report, researchers use the term *formative assessment* to denote "a process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning" (Council of Chief State School Officers [CCSSO] 2007, p. 2]). Benchmark assessment is used much less frequently (two to four times a year). It is designed primarily to predict a student's academic success, monitor progress, and provide information about a student's performance on a specific set of standards or skills that teachers can use to differentiate instruction. Although formative assessment is conducted unobtrusively as part of normal classroom activity, benchmark assessment is administered as an interrupted event that occurs outside the context of normal instruction (Hunt and Pellegrino 2002).

current study focuses on the impact of a program that relies on training to understand and use MAP assessment results to differentiate instruction for students. Second, both benchmark studies used a quasi-experimental design to create a set of comparison schools that were similar to treatment schools across several observable characteristics, leaving open the possibility that other known or unknown factors could have influenced the study's findings.

Although the MAP program is used extensively in school districts across the United States, there is no experimental evidence on its impact on student outcomes. Given that the number of schools investing in MAP and similar programs is projected to increase, evidence on the effectiveness of such programs is critical.

In this study, the study team focused on the effects of MAP on reading outcomes. Reading outcomes were selected for this study for two primary reasons. First, reading proficiency in elementary school is fundamental to students' ongoing success in school, and, in the current era of accountability, differentiating reading instruction has become a primary focus among elementary schools and teachers. Second, observation and survey instruments designed to measure classroom reading instruction were prevalent and well tested at the time this study began. Access to these measures enhanced the study team's ability to develop valid scales to index the extent to which teachers differentiated reading instruction.

Description of intervention

The MAP program has two main components: an extensive portfolio of tests and training and on-demand support in the use of test results to guide instructional practice. Each component is described below.

MAP assessments

The MAP assessments are a collection of computer-adaptive tests in reading, language usage, mathematics, and science that place individual students on a continuum of learning from grade 3 to grade 10 in each discipline. Each MAP assessment uses a continuous interval scale, called the Rasch (RIT) unit scale score, to evaluate student growth and student mastery of various strand-defined skills within disciplines.[5] NWEA has conducted scale alignment studies linking the MAP assessment's RIT scale to proficiency levels from standardized assessments in all 50 states and the District of Columbia. These studies provide evidence of an association between the MAP assessments and each state's standardized test (Brown and Coughlin 2007; Northwest Evaluation Association 2005). In addition, studies provide evidence that MAP assessments predict performance on assessments in at least five states (Cronin et al. 2007; Northwest Evaluation Association 2008; Steering Committee of the Delaware Statewide Academic Growth Assessment Pilot 2007). Relying on this evidence, schools and teachers use MAP results to monitor their students' progress toward state proficiency standards. NWEA recommends that schools administer each MAP subject area test to students three times during the school year (in the fall, winter, and spring), with a fourth administration suggested during summer school. Because the

[5] RIT uses individual test item difficulty to estimate student achievement level. RIT scores are reported on an equal interval scale, so that differences between scores have the same meaning regardless of whether a student is at the top, bottom, or middle of the RIT scale and regardless of grade level.

tests are computer-adaptive, students are given their overall score immediately after the test ends, and teachers can generate a series of customized reports on students' performance on key subject domains and goal strands within 24 hours of administration.

For this study, the researchers employed the MAP tests in reading and language usage for grades 4 and 5 and administered the tests three times a year (in the fall, winter, and spring) to treatment students and once (in the spring) to control students.

MAP training

To support the administration and use of the MAP assessments, NWEA provides training sessions and face-to-face consultative services. MAP training consists of four one-day sessions, along with on-demand consultation through conference calls and on-site visits from an NWEA MAP coach throughout the school year. The primary objectives of the training are to equip teachers with the knowledge and skills to administer the tests; generate and interpret outcome reports at the individual, group, and classroom level; use report results and other MAP online resources to determine student readiness and differentiate instruction; and use MAP data over time to set student growth goals and evaluate instructional programs and practices. The MAP data reports (which include a student's Lexile range score) allow teachers to group students appropriately on the basis of their skill needs, to identify books and learning resources that are appropriate for students at different reading levels, and to differentiate, or individualize, instruction in order to more effectively address students learning needs. In each of the four one-day sessions, a certified MAP trainer lectures and facilitates a structured set of activities on one of the four major topic areas (table 1.1) corresponding to the objectives of the training. Schools have the option of scheduling three to four consultative sessions throughout the school year with a MAP trainer to provide further training on specific areas of need (for instance, teachers may request assistance generating reports or understanding how to use the results to group students for reading instruction or to target individual student skill needs). Visits typically last one to two hours and may occur before, during, or after school.

A key underlying assumption embedded throughout the training continuum maintains that differentiated instruction relies on the availability of periodic assessment data and that effective use of the data requires a clear and functional understanding of techniques in differentiation. The theory underlying the MAP program is that, as teachers become more adept at interpreting MAP data reports and utilizing available resources to differentiate instruction, student achievement will improve. MAP testing is spaced out across the school year, and teachers have unrestricted access to student-level MAP results obtained from the multiple test administrations. They also have access to online resources to assist them in interpreting results, reconfiguring instructional strategies, and tailoring instruction to the needs of students. These resources provide opportunities for teachers to alter their instructional approaches between MAP test administrations.

Table 1.1. Prototypical MAP testing and training timeline

Component	August	September	October	November	December	January	February	March	April	May
MAP Testing	X →	X			X →	X		X →	X	
Training Session 1: MAP Administration	X									
Training Session 2: Using MAP Data			X							
Training Session 3: Differentiated Instruction						X				
Training Session 4: Growth and Goals										X
Consultative on-site school visits		X		X		X				

Source: NWEA Certified Training Manuals 2008.

Research questions

This study used an experimental design to assess the effectiveness of NWEA's MAP benchmark testing system and teacher training on grade 4 students' reading performance in five districts in Illinois. The study investigated one primary and two secondary confirmatory research questions:[6]

1. Did the MAP program (that is, training plus benchmark testing feedback) affect the reading achievement of grade 4 students after Year 2 of implementation, as measured by the Illinois Standards Achievement Test (ISAT) reading scale scores or the MAP composite test scores in reading and language use?
2. Were MAP resources (training, consultation, web-based materials) delivered by NWEA and received and used by teachers as planned?
3. Did MAP teachers apply differentiated instructional practices in their classes to a greater extent than their control counterparts?

[6] The Year 2 grade 5 cohort is not treated as a confirmatory test of the effects of MAP because some of the students in the Year 2 control classes had been enrolled in grade 4 classes in which teachers were exposed to MAP training and resources. For this reason, in Year 2 only the grade 4 cohort was used in the confirmatory intent-to-treat analysis. Investigation of the Year 2 grade 5 cohort is treated as an exploratory analysis. It uses the same analytic methods as the confirmatory analyses for the Year 2 grade 4 cohort. Although the Year 2 grade 5 cohort is treated as an exploratory analysis, appendix L provides supplemental analyses that suggest little or no between-condition contamination.

The report also addressed one exploratory question:

4. Did the MAP program affect the reading achievement of grade 5 students after Year 2 of implementation, as measured by the Illinois Standards Achievement Test (ISAT) reading scale scores or the MAP composite test scores in reading and language use?

The most critical MAP training component—Session 3, on using MAP to differentiate instruction—was not delivered until January 2009 (see table 1.1). This left at most two and a half months for treatment teachers to apply what they learned to their classroom instruction before the ISAT was administered in March 2009. Because of the short interval between the delivery of Session 3 and state testing, there were limited opportunities for teachers to implement instructional practices embodied in the training and consultation components during Year 1. Thus, Year 1 (2008/09) of this study is regarded as an "implementation-process year." This report analyzes results using student outcome data collected in spring 2010, allowing a full year (2009/10) for teachers to implement the full MAP program in their classes.

Roadmap to this report

Chapter 2 provides details on the study's research design, sample recruitment and characteristics, data collection and outcome measures, data analytic methods, and limitations of the study design. Chapter 3 addresses implementation fidelity. It discusses the MAP intervention as implemented for this study and presents findings on the fidelity of implementation, including the extent of program delivery by NWEA and teacher participation in MAP training and consultation services, use of MAP resources, and modification of instructional practices in keeping with the principles of differentiated instruction. Changes in teacher practices are viewed as part of the causal process that affects student achievement. Specifically, the study team regards the implementation process as entailing the delivery of services and resources, the receipt of these resources and services by teachers, and the adoption of desired instructional practices by teachers. Of course, the adoption of instructional practices can be viewed as an intermediate outcome. But because this report assesses the impact of MAP implementation on student outcomes, differences in teacher practices are conceptualized as belonging to the causal chain. Chapter 4 presents confirmatory and exploratory results on student achievement for the Year 2 grade 4 cohort. Chapter 5 presents exploratory results for the Year 2 grade 5 cohort.

Chapter 2: Study Design and Methodology

This study employed a two-year cluster-randomized design to obtain unbiased estimates of the impact of the MAP program on student reading achievement. A cluster-randomized design randomly assigns clusters of units to either a treatment or a control condition. Randomization ensures that the treatment and control groups are, in expectation, equivalent on baseline characteristics, and therefore yields unbiased estimates of the causal effects of being randomized to the intervention. This chapter describes the research design, recruitment of districts and schools, randomization of schools to treatment or control condition, analysis sample, and baseline characteristics of participating schools, teachers, and students. It also discusses attrition, data collection and measures, methods used for impact estimation, and study limitations.

Recruitment

Sample eligibility

Districts employ various types of reading assessments for a variety of purposes. Examples include summative tests to measure end-of-year performance; screening assessments to identify students who may need intensive reading assistance; diagnostic assessments to identify specific instructional needs; classroom-based formative assessments for more immediate and individualized instructional adjustments; and benchmark assessments to monitor student progress, make adjustments in how students are grouped for instruction, and provide targeted instructional assistance (Torgesen and Miller 2009).

Districts and schools were eligible for this study if they implemented any of these assessment types except benchmark assessments similar to those used in the MAP program. Districts were not eligible if they had previously adopted or used MAP or similar computer-adaptive benchmark testing programs in any of their schools. To participate in the study, districts had to agree to delay schoolwide implementation of MAP or similar programs in the study schools for two years, starting in fall 2008. Districts were also asked to assign a point of contact to act as a liaison between the study team and the school community, to facilitate formal district approval for the study, and to assist the study team in gathering data on teachers and students.

The study focused on schools in Illinois because it was the Midwest state with the largest number of interested and potentially eligible districts and schools. Districts and their eligible schools were required to agree to school-level random assignment. To be eligible, schools needed to have at least one full-time regular classroom teacher who taught reading in a self-contained classroom in grade 4 and one full-time regular classroom teacher who taught reading in a self-contained classroom in grade 5.

Grade 4 and 5 reading teachers were eligible provided they had not previously been exposed to MAP or MAP-like products or training (box 2.1). The study population was restricted to regular education classroom teachers (special education and gifted education teachers were not eligible). Participating teachers agreed to carry out the requirements associated with their school's assignment to the treatment or control condition. For teachers assigned to treatment,

requirements included administering the MAP test three times a year and, at minimum, participating in four day-long training sessions during the year. Teachers were also encouraged to participate in consultative sessions throughout the school year. During these sessions, a MAP trainer provided on-site technical assistance and individualized support to MAP teachers. Teachers in the control condition were asked to conduct business as usual and to agree not to review or use any MAP program materials or resources. Control group teachers also agreed to administer the MAP assessment once a year, in the spring of each of the two study years. The total score function displayed at the end of the test was turned off for students of control teachers in order to eliminate any potential influence the final test result may have had on their instructional practices or students' future test performance.

These participation requirements were established so that, at the study's conclusion, the study team could rigorously assess what the outcomes would have been for the treatment group had it not been exposed to the MAP program and continued in a business-as-usual fashion. Business as usual did not preclude control group schools or teachers from testing their students or using results from a variety of assessments available for making instructional decisions. It did prohibit teachers from administering MAP or similar computer-adaptive assessments and from attending MAP or a similar training program during the two-year study period.

Box 2.1 Eligibility criteria for participating in the study

Districts must…

- Assign a district point of contact to support and assist the study team with all data collection activities.
- Obtain study approval from the district's board of education or institutional review board.
- Facilitate provision of data on teachers and students for all grade 4 and 5 reading/English language arts classrooms between fall 2007 and spring 2010.
- Delay schoolwide implementation of MAP in study schools for two years (2008–10).

Schools must…

- Include at least one grade 4 and one grade 5 self-contained classroom.
- Not have used MAP or associated training in prior years.
- Not be implementing a benchmark assessment program with features similar to the MAP program.
- Agree to school-level random assignment to the control or treatment group.

Teachers must…

- Teach grade 4 or grade 5 students reading and English language arts in a self-contained regular education classroom.
- Not have used MAP or associated training in prior years.
- Agree to carry out requirements associated with their school's assignment to the control or treatment group.

Sample size requirements

In fall 2007, the study team developed a plan, based on a power analysis, to recruit a minimum of 30 eligible schools to detect an effect size of at least 0.20 standard deviation on a statewide accountability measure and the MAP assessments. The choice of a minimum detectable effect size of 0.20 was based on Nyquist's (2003) meta-analytic study of the effects of formative assessment on learning outcomes, which indicated that the effects of feedback on achievement were about 0.15–0.50 standard deviation. The largest effects were observed when feedback was immediate and detailed (for example, provided directions for improvement, explained why an answer was incorrect, provided a goal).

The studies reviewed in this meta-analysis were conducted mainly in laboratory settings or classrooms, where the researcher had greater control over the delivery of the feedback and other important elements of formative assessment (for example, use of meta-cognitive strategies to improve performance, goal specification) than in public school classrooms (see Hulleman and Cordray 2009 for evidence of differences in effects between laboratory studies and regular classroom sessions). The research team expected that teachers would vary in the fidelity with which they used the MAP assessments in their classes. In addition, the MAP program uses interim assessments that are administered less frequently (three times during the school year) and employs feedback from these assessments for grouping students and responding to their individual instructional needs in less immediate ways than the feedback processes included in Nyquist's (2003) report. Given these differences, the study team chose a more conservative detectable effect size of 0.20.

Sample recruitment

Schools were recruited for the study beginning in spring 2008. Initially, the study team collected district and school demographic information on all schools in the Midwest Region (Illinois, Indiana, Iowa, Michigan, Minnesota, Ohio, and Wisconsin) from the 2007/08 National Center for Education Statistics Common Core of Data. The demographic data were merged with administrative files from NWEA, which included data on district clients and districts that had contacted NWEA about potentially implementing the MAP program. Using this information, the study team and NWEA identified Illinois as the state that had the largest number of interested and potentially eligible districts and schools.

After narrowing the list of potentially eligible schools, the study team sent a letter to each district's superintendent introducing the study. The study team held at least one phone conference with key staff from each interested district to explain the study and request additional information with which to determine eligibility. After the phone conferences, representatives from NWEA and REL Midwest visited school sites to present the study to principals and teachers, answer questions, and confirm school eligibility.[7] Once confirmed, districts and schools

[7] This stage of the recruitment process was handled differently in the largest district (District 1) because of the large number of participating schools. During the initial site visit, in spring 2008, researchers presented the study to administrators and teachers from all the study schools at one time in the auditorium of the local high school. They conducted a follow-up site visit shortly before the 2008–09 school year began in order to gather administrators' consent, distribute information to teachers, and describe the process for working with administrators to gather individual teacher consent forms.

that elected to participate signed a Memorandum of Understanding to confirm their commitment. In total, 32 schools were randomly assigned to conditions.[8] Table 2.1 indicates the sample sizes at each stage of the recruitment process.

Table 2.1. Recruitment stages and sample sizes

Stage	Number of districts	Number of schools
Initial district contact (spring 2008)		
Sent introduction letter and made follow-up call	88	553
Conducted initial web-based conference call	14	93
Approved school eligibility and verified interest	7	54
District site visits (spring/summer 2008)		
Presented MAP program and study	7	54
Collected memorandum of understanding	5	32
Conducted school-level random assignment	5	32
District and school follow-up site visits (fall 2008)		
Presented study/confirmed teacher eligibility	5	32
Conducted random assignment	5	32

Source: Authors' compilation based on data from the study districts and the Northwest Evaluation Association.

Teacher consent

In late summer 2008, a member of the study team visited the study schools and presented information about study participation to school administrators and eligible grade 4 and 5 teachers. The presentation included information about the study's purpose; possible risks and discomforts to participants; benefits; confidentiality; and whom to contact with questions throughout the study period. Teachers reviewed a packet of information that provided study details along with a teacher consent form. All eligible staff were given time to ask questions about the study and the consent process, review the information packet carefully, and sign and submit the consent form if and when they were ready. Teachers who were still uncertain about participating in the study after this meeting were invited to e-mail or fax their signed consent at a later time. Appendix K provides detailed information about the recruitment process.

Random assignment of schools to treatment

Once the 32 schools were identified for study participation, schools were randomly assigned to one of two conditions, receiving the MAP program in grade 4 or grade 5. If grade 5 classrooms in School A were assigned to the treatment condition, grade 4 classrooms in the school were assigned to the control condition. If grade 5 classrooms in School B were assigned to the control condition, grade 4 classrooms were assigned to the treatment condition. The control group for grade 4 classes consisted of grade 4 classes in schools in which MAP was randomly assigned to

[8] Less than four schools dropped out of the study immediately after randomization.

grade 5, and the control group for grade 5 classes consisted of grade 5 classes in schools in which MAP was randomly assigned to grade 4.

This randomization technique resulted in two experiments, one at grade 4 and one at grade 5, and produced a valid counterfactual for the treatment group within each grade (see Borman et al. 2007 for a similar randomization design).[9] It had the added advantage of being more appealing to schools because it guaranteed a more equitable distribution of the intervention (leaving no school totally deprived of the intervention during the two study years). One potential drawback to this approach, however, is the increased chance of contamination, in the form of teachers (or school leaders) at the treatment grade influencing the instructional practices of teachers in the control grades within the same school. Results of interim fidelity studies in Year 1 found no evidence of contamination.[10] If contamination occurred in Year 2 (but was not detected because it introduced hidden or unmeasured bias), it would attenuate the magnitude of the estimated impacts presented in this report.

A second potential route of contamination that springs from this method of randomization is the exposure of students to MAP teachers in Year 1 and then to control teachers in Year 2. This would be the case for grade 5 students in Year 2 who attended (and stayed in) a school that was randomized to implement the MAP program in grade 4. The study team found no evidence of this form of contamination. If exposure to MAP teachers attempting to implement the intervention in grade 4 in Year 1 affected the students' grade 5 performance in Year 2, this exposure could result in the underestimation of the magnitude of estimated impacts for grade 5 students. Erring on the side of caution, only the core intent-to-treat analyses of overall impacts on grade 4 outcomes in Year 2 are considered confirmatory. Analyses of Year 2 grade 5 outcomes are considered exploratory.

To minimize extraneous sources of variation caused by district differences and to improve the power to detect impact, the study team blocked random assignment of grade levels within schools by district. This block randomization resulted in treatment and control schools being roughly equally represented within each district, with 16 schools randomized to MAP in each grade.[11] Blocking was advantageous for this study because the five participating districts varied considerably in size and student composition. One of the five participating districts (District 1) was considerably larger than the other four (table 2.2).[12] This district also had a much higher proportion of economically disadvantaged students, and a more ethnically diverse student population.

---

[9] The counterfactual condition included schools that implemented a variety of assessment types but had never implemented benchmark assessment or conducted training to help teachers interpret and use benchmark data to inform their instruction.

[10] See appendix L for a detailed discussion on the issue of control group contamination.

[11] The randomization of schools to treatment or control condition was carried out as follows. The 32 schools were arranged in a list stratified by district. Thirty-two five-digit (uniformly distributed) random numbers between 0 and 1 were then generated using Excel. Within each district, schools were assigned to implement the MAP program in grade 4 if the fifth digit was even; they were assigned to implement MAP in grade 5 if the fifth digit was odd. This process resulted in no even digits for one district that had four schools (District 3). Allocation of the four schools in this district was determined by a flip of a coin, resulting in two schools assigned to the MAP program in grade 4 and two schools assigned to the MAP program in grade 5.

[12] For the district characteristics in Year 1 (2008/09), see table A.1, appendix A. The results exhibit patterns similar to the ones observed here.

Table 2.2. Characteristics of the five study districts, 2007/08

Characteristic	District				
	1	**2[a]**	**3**	**4**	**5**
Number of schools	20	2	4	3	3
Socioeconomic status					
Percentage of Title I schools in district	100.0	–	100.0	33.3	33.3
Percentage of students in district eligible for free or reduced-price	70.3	–	21.0	18.7	0.0
Race/ethnicity (percentage of students in district)					
Hispanic	1.7	–	2.2	24.3	13.1
Black	37.0	–	2.0	0.5	5.2
White	50.8	–	92.7	67.8	74.6
Other	10.5	–	3.1	7.5	7.0
Enrollment and number of teachers					
Total district enrollment	6,151	–	1,471	1,917	1,623
Total number of full-time teachers in each district	389	–	72	103	102

a. The characteristics of District 2 have been suppressed to prevent a disclosure risk.

Source: Authors' analysis based on data from the National Center for Education Statistics Common Core of Data 2007/08.

Analytic sample

In spring 2007, before the first year of the study, 32 schools were block-randomized to adopt the MAP intervention in either grade 4 or grade 5. Figure 2.1 describes the construction of the Year 2 (2009/10) analytic sample from the 184 teachers and 3,787 grade 4 and grade 5 students present at the start of Year 2. These 184 teachers were composed of 147 "two-year" teachers (teachers in either grade 4 or grade 5 during both years of the study) and 37 "one-year" teachers (new to the study in Year 2). Of the 184 teachers, 12 did not satisfy the eligibility criteria because they were either special education or gifted education teachers. These teachers and their 67 students were excluded from the analyses, leaving 172 teachers and 3,720 students in the final analytic sample.

Of the 172 teachers in the analytic sample, 145 were "two-year" teachers and 27 were "one - year".[13] Of the 145 "two-year" teachers, 140 taught in the same school and the same grade during both years of the study, and five teachers changed positions leading to a change in treatment condition.[14] The intent-to-treat analysis of the Year 2 sample is a school-level intent-to-treat analysis, in the sense that the treatment assignment of these 172 teachers (and their students) is determined by the original random assignment of the grade and the school to which they belonged in Year 2.[15]

Of the 32 schools randomized, less than four schools withdrew from the study immediately after randomization. This sample attrition contributed to the data attrition on the MAP 2010 composite test scores but not the ISAT 2010 test scores, because all students in the schools that left the study took the ISAT 2010. These schools are included in all tables, figures, and analyses in this report unless otherwise noted. Because these schools were included in random assignment, they are included in the intent-to-treat analysis of the Year 2 grade 4 sample presented in chapter 4 and the intent-to-treat analysis of the Year 2 grade 5 sample in chapter 5. In addition to the teachers (from the schools that withdrew) that did not implement the MAP program, the intent-to-treat analysis includes students of other teachers who either did not consent to participate or were temporarily away from their school.[16] In sum, students of a total of nine "inactive" teachers were included in the intent-to-treat analysis.

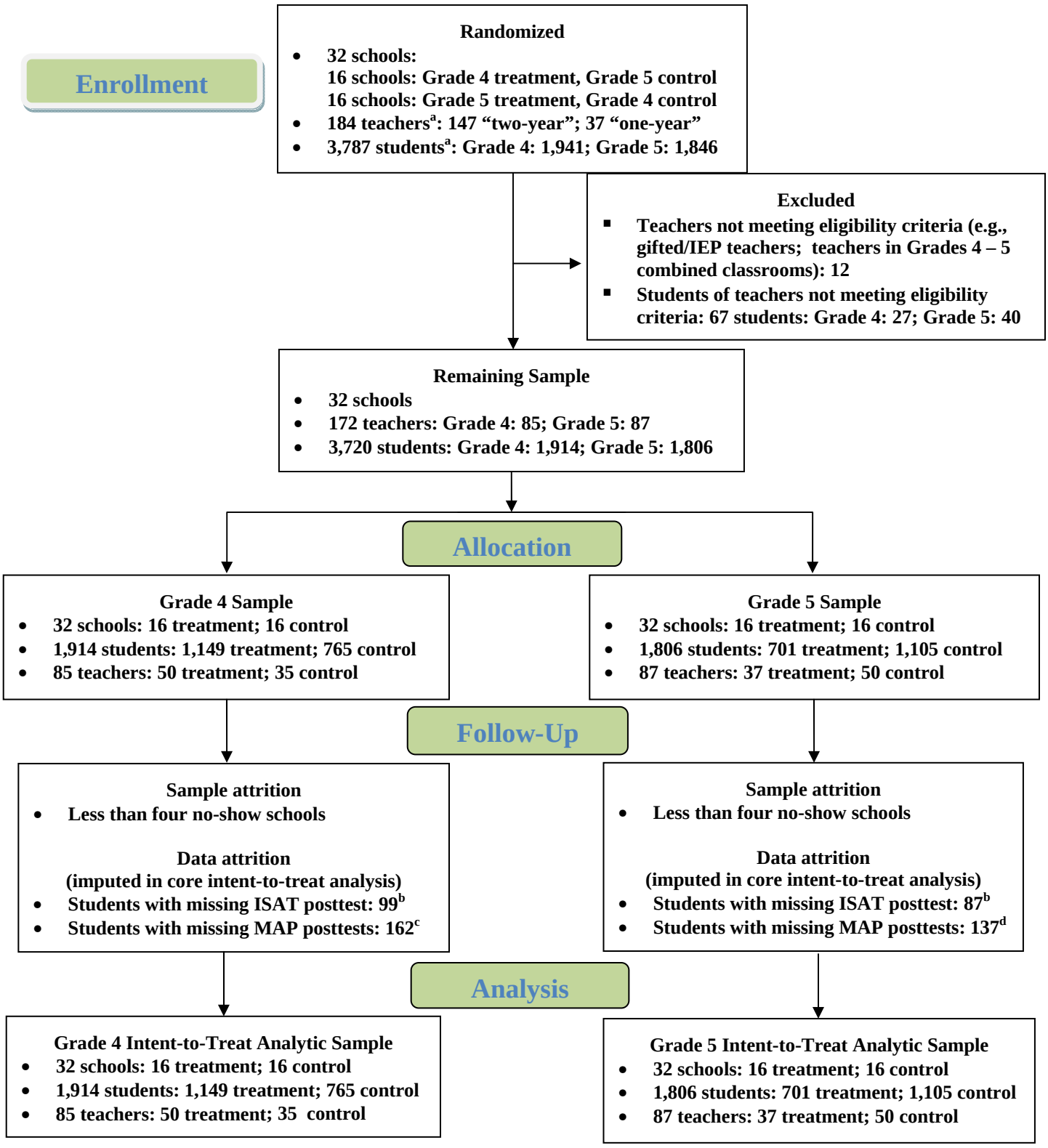
[13] Of the 178 eligible teachers in Year 1, 33 teachers were not in the study in Year 2 for various reasons (moved to another district, taught a grade other than grades 4 and 5, retired). The remaining 145 teachers consisted of 70 grade 4 teachers (42 treatment, 28 control) and 75 grade 5 teachers (29 treatment, 46 control). The 27 "one-year" teachers included 15 grade 4 teachers (8 treatment, 7 control) and 12 grade 5 teachers (8 treatment, 4 control) in Year 2

[14] Of the five teachers that changed treatment condition, some crossed over from treatment to control, while some crossed over from control to treatment.

[15] For example, if a teacher taught in the same grade 4 treatment, grade 5 control school during the two-year study period but taught grade 4 in Year 1 and grade 5 in Year 2, she would be assigned to the control group in Year 2 (even though she was a MAP teacher in Year 1), because the treatment condition is based on the school's grade-level assignment.

[16] Data on classroom observations, logs, and surveys were not collected from teachers who declined to participate in the study, and these teachers did not participate in the MAP training. Their students, however, were administered the MAP assessments.

Figure 2.1 CONSORT flow diagram for 2009/10



a. Grade 4 and grade 5 teachers and students who were present in the 32 study schools at the start of the 2009/10 school year.
b. This count does not include students from the less than four no-show schools, because they all had nonmissing ISAT 2010 posttest scores.
c. Students with missing scores on both the spring 2010 MAP reading and spring 2010 MAP language tests.
d. Students with missing scores on both the spring 2010 MAP reading and spring 2010 MAP language tests.
Source: Authors' analysis based on data from the study districts and the Northwest Evaluation Association.

Sample distribution

The distribution of the final analytic sample appears in tables 2.3, 2.4, and 2.5. Table 2.3 gives the sample distribution by treatment group separately for grades 4 and 5; tables 2.4 and 2.5 present the distributions disaggregated by district and treatment group. Table 2.3 shows that in both grades, there is imbalance in the teacher and student sample sizes between the MAP and control groups. For example, in grade 4, there were 1,149 MAP students and 765 control students, and in grade 5 there were 701 MAP students and 1,261 control students.

This imbalance can be attributed to the fact that four of the schools with the largest Year 2 total enrollment in grade 4 and grade 5 were randomly assigned to the "grade 4 MAP/grade 5 control" condition, making the sample sizes for the MAP group in grade 4 and the control group in grade 5 larger than their respective counterparts in each grade. For the same reason, there is an imbalance between the samples sizes of teachers in the MAP and control groups: 50 MAP teachers in contrast with 35 control teachers in grade 4, and 50 control teachers in contrast with 37 MAP teachers in grade 5.

Table 2.3. Sample distribution in Year 2 (2009/10)

Sample	Grade 4 MAP, grade 5 control schools	Grade 5 MAP, grade 4 control schools	Total
Schools	16	16	32
Teachers			
Grade 4	50 (MAP)	35 (control)	85
Grade 5	50 (control)	37 (MAP)	87
Total	100	72	172
Students			
Grade 4	1,149 (MAP)	765 (control)	1,914
Grade 5	1,105 (control)	701 (MAP)	1,806
Total	2,254	1,466	3,720

Source: Authors' analysis based on data from the study districts and the Northwest Evaluation Association.

The distributions of the Year 2 analytic samples are broken down by districts in table 2.4 (grade 4) and table 2.5 (grade 5). In each grade, the imbalance in teacher and student sample sizes between the MAP and control groups in the overall distribution is also observed for teachers in Districts 4 and 5 and for students in all districts except District 3. Moreover, the number of participating schools (and consequently the total number of participating teachers and students) is disproportionately distributed across the five districts, with the first district having larger samples of schools, teachers, and students.

Table 2.4. Grade 4 sample distribution in Year 2 (2009/10), by district

District	Number of teachers			Number of students		
	Total	MAP	Control	Total	MAP	Control
1	39	20	19	869	477	392
2	7	3	4	150	59	91
3	11	6	5	262	129	133
4	13	9	4	300	218	82
5	15	12	3	333	266	67
Total	85	50	35	1,914	1,149	765

Source: Authors' analysis based on data from the study districts.

Table 2.5. Grade 5 sample distribution in Year 2 (2009/10), by district

District	Number of teachers			Number of students		
	Total	MAP	Control	Total	MAP	Control
1	41	20	21	781	346	435
2	7	4	3	147	74	73
3	10	5	5	252	115	137
4	13	4	9	304	90	214
5	16	4	12	322	76	246
Total	87	37	50	1,806	701	1,105

Source: Authors' analysis based on data from the study districts.

Characteristics of participating schools

Table 2.6 summarizes the characteristics of the 32 schools that participated in the study in the year before the start of the study.[17] Twenty-eight of the 32 schools (87.5 percent) were eligible for Title 1 services, and 78.1 percent were located in either a city or a suburb.[18] On average, about half the students in the participating schools were eligible for free or reduced-price lunch (range: 0–95 percent), and about 62 percent were White (range: 8–97 percent). Total enrollment in the study schools ranged from 162 to 701, with an average of 385 students (including about 60 students in grade 4 and 60 in grade 5) taught by about 23 full-time teachers in each school.

To provide context to the types of schools included in the study, table 2.6. also shows the characteristics of all eligible schools in Illinois; in seven REL Midwest states (Illinois, Indiana,

[17] The characteristics of the study schools in the first year of the study (2008/09) are summarized in table A.2 in appendix A. Comparison of the school characteristic across the two years indicates that three schools changed Title I status; less than four schools changed locale (reclassified from suburb to rural); and the average number of full-time teachers increased from 23 to 27, even though total enrollment stayed about the same (the increase probably reflected the fact that one pre-K–grade 2 school became a pre-K–grade 5 school and one grade 3–5 school became a pre-K–grade 5).

[18] These classifications are based on the National Center for Educations Statistics revised (2006) typology of locale codes, in which city, suburb, town, and rural were subclassified into three categories, resulting in 12 urban locale codes (http://nces.ed.gov/ccd/rural_locales.asp).

Iowa, Michigan, Minnesota, Ohio, and Wisconsin); and in the United States, as well as all regular schools in the United States.[19] Relative to other eligible schools in Illinois and in the United States, the study schools have higher rates of Title 1 eligibility and higher percentages of White students, and they are more likely to be located in a city. They have about the same percentages of students eligible for free or reduced-price lunch, and have lower total and grades 4 and 5 enrollments.

Table 2.6. Characteristics of study schools and eligible schools in Illinois, the Midwest, and the United States the year before random assignment (2007/08)

Characteristic	All study schools	Eligible schools in Illinois[a]	Eligible schools in Midwest[b]	Eligible schools in United States[c]	All U.S. schools[d]
Number of schools	32	1,962	7,994	38,022	43,873
Socioeconomic status					
Percentage of Title I schools	87.5	81.3 (*n* = 1,748)	84.3 (*n* = 7,780)	75.7 (*n* = 37,609)	75.7 (*n* = 43,393)
Average percentage of students eligible for free or reduced-price lunch	50.3	48.5 (*n* = 1,905)	44.2 (*n* = 6,486)	49.4 (*n* = 35,939)	49.3 (*n* = 41,389)
Race/ethnicity and gender (average percentage of students)					
Hispanic	5.0	18.9	9.2	21.7	20.2
Black	24.4	23.0	16.8	17.4	17.3
White	62.3	50.9	68.0	53.4	55.0
Other	8.3	7.2	6.0	7.5	7.4
Male	48.2	49.6	50.2	50.9	50.8
Enrollment and number of teachers					
Average total school enrollment	385	461	415	488	475
Average number of students in grade 4	65 (*n* = 31)	69	66	77	74
Average number of students in grade 5	63 (*n* = 31)	70	66	76	73
Average number of full-time teachers	23	25 (*n* = 1,960)	24 (*n* = 7,991)	31 (*n* = 37,994)	30 (*n* = 43,609)
School setting (percentage of schools)					
City	50.0	35.3	29.2	31.5	31.3

[19] The National Center for Education Statistics defines a "regular" school as a public elementary or secondary school that does not focus primarily on vocational, special, or alternative education.

Characteristic	All study schools	Eligible schools in Illinois[a]	Eligible schools in Midwest[b]	Eligible schools in United States[c]	All U.S. schools[d]
Suburb	28.1	38.5	29.8	32.4	30.7
Town	3.1	7.5	12.3	10.4	10.7
Rural	18.8	18.7	28.7	25.7	27.2

Note: Averages are unweighted means across schools. When data are missing on some schools, *n* is the actual number of schools used for calculating the average characteristic across schools.

a. Schools located in Illinois that had at least 10 students in grade 4 and at least 10 students in grade 5, were noncharter schools, were defined as "regular" schools by the Common Core of Data, and were operational at the time of the Common Core of Data report.

b. Schools that met the same eligibility criteria but were located in the seven states served by the REL Midwest (Illinois, Indiana, Iowa, Michigan, Minnesota, Ohio, and Wisconsin).

c. Schools that met the same eligibility criteria but were located in the 50 states and the District of Columbia.

d. All schools in the 50 states and the District of Columbia that had at least 10 students in grade 4 and at least 10 students in grade 5 during 2007/08, were defined as regular schools by the Common Core of Data, and were operational at the time of the Common Core of Data report.

Source: Authors' analysis based on data from the National Center for Education Statistics Common Core of Data 2007/08.

Baseline comparisons

The purpose of randomization is to create groups that are, on expectation, equivalent on all observable and unobservable characteristics so that any observed differences in the outcomes between the treatment and control groups can be attributed to the intervention. Although it is not possible to test equivalence on unobservable characteristics, baseline equivalence of the groups can be assessed for variables on which data are available.[20] To assess whether the randomization of schools within districts (blocks) yielded groups that have, on average, similar baseline characteristics, the study team compared treatment and control schools, teachers, and students.

Because randomization was blocked by district, the study team accounted for the design by conducting district-specific comparisons and pooling the district-specific estimates into an average, weighted by the number of schools within each district. For school characteristics, the study team conducted comparisons for the year before randomization (2007/08) and the year before the Year 2 implementation (2008/09). For teacher characteristics, the study team compared MAP and control teachers in the Year 2 analytic sample based on their characteristics in the year before the Year 2 implementation. They compared the characteristics of students in the Year 2 analytic sample using their characteristics before the Year 2 implementation.

Table 2.7 and table A.3, in appendix A, show the results of the school comparisons. Tables 2.8 and 2.9 present the grade 4 teacher and student comparisons. Tables A.4 and A.5, in appendix A, present the corresponding grade 5 comparisons. These comparisons entailed numerous hypothesis tests (in general, one for each baseline characteristic compared), which increased the

[20] As Bloom (2006) underscores, the randomization process yields intervention and control groups that are equivalent on all observable and unobservable characteristics on average. Randomization applied to a specific sample does not guarantee group equivalence, because it is possible to obtain groups that differ simply by chance.

chances of concluding that the groups were significantly different in one characteristic when in fact they were not (that is, inflating the Type I error). To protect against spurious significant findings, the study team also conducted a joint test of the overall difference between the groups using a chi-square test or an *F*-test. The results of these tests are presented at the bottom of the tables. A significant omnibus test indicates that the groups differed in at least one of the characteristics in the table. A nonsignificant omnibus test indicates that any significant difference for a single baseline characteristic may have been caused by chance.

School characteristics

Before Year 1 implementation, there were no statistically significant differences in school characteristics of MAP and control schools, with the exception of two variables that measure school size (table 2.7). Specifically, grade 4 MAP schools enrolled significantly more students and (consequently) had more full-time equivalent teachers. MAP and control schools were not systematically different in terms of their characteristics, however. A similar pattern was found in the comparison of school characteristics before Year 2 implementation (table A.3, in appendix A).[21]

Table 2.7. Characteristics of study schools the year before random assignment (2007/08)

Characteristic	Mean		Estimated difference	*p*-value
	Grade 4 MAP/ grade 5 control schools	Grade 5 MAP/ grade 4 control schools		
Number of schools	16	16		
Title I and school composition				
Percentage of Title I schools	85.9	90.6	–4.7	.384
Average percentage of students eligible for free or reduced-price lunch	46.0	54.1	–8.1	.066
Average percentage of White students	62.4	62.1	0.3	.955
Average percentage of male students	47.2	48.8	–1.5	.284
Enrollment and number of teachers				
Average total school enrollment	435	340	95	.007*

[21] As shown in table A.3, in addition to differences in total enrollment, researchers found that in the year before the Year 2 implementation, the grade 4 control schools had statistically significantly higher percentages of students eligible for free or reduced-price lunch and smaller numbers of students in grade 4 or grade 5 (although the percentage of students in grade 4 and 5 [not shown] were about the same between the two groups). However, as in the year before the Year 1 implementation, the joint test was nonsignificant (*p*-value = .323), indicating that there were no systematic differences between the two groups.

Characteristic	Mean		Estimated difference	*p*-value
	Grade 4 MAP/ grade 5 control schools	Grade 5 MAP/ grade 4 control schools		
Average number of students in grade 4	69 (n = 15)	58	11	.361
Average number of students in grade 5	67 (n = 15)	56	11	.340
Average number of full-time teachers	25	20	5	.009*
School locale (percentage of schools)[a]				
City	56.3	43.8	12.5	.480
Suburb	18.8	37.5	–18.8	.238
Town	0	6.2	–6.2	.310
Rural	25.0	12.5	12.5	.365
Joint test of difference in school characteristics between MAP and control groups[b] (χ^2 = 10.2, df = 9)				.331

* Difference statistically significantly different from zero at the .05 level.

Note: Means and differences were regression adjusted using ordinary least squares to account for district effects and weighted by the number of schools in each district.

a. The chi-squared test of homogeneity of distributions is not statistically significant (χ^2 = 2.92, *p*-value =.405).

b. An overall test of the difference between MAP and control groups based on all school characteristics in the table was conducted using a chi-square test. The chi-square test is from a logistic regression model with the binary treatment indicator as outcome and the school characteristics as covariates (school locale was included in the model as the combined percentage of city or suburb, because no schools in the grade 4 MAP/grade 5 control sample were located in towns).

Source: Authors' analysis based on data from the National Center for Education Statistics Common Core of Data 2007/08.

Teacher characteristics

Demographic characteristics of the 85 participating grade 4 teachers are compared in table 2.8.[22] The omnibus test reveals that there were no systematic differences between MAP and control teachers (*p*-value = .286) despite the fact that individual hypothesis tests show that, on average, MAP teachers were more experienced (by about three years) in teaching English language arts and more likely to be White (almost 100 percent versus about 90 percent). Both groups of teachers are predominantly White and female, more than three out of four have graduate degrees, and about four out of five have permanent teaching licenses. To increase precision and minimize bias, the study team used all these characteristics as covariates in the core impact models for student achievement (see equation B.1, in appendix B).

[22] A similar comparison for the 87 grade 5 participating teachers, given table A.4, appendix A, shows that there are no significant differences between MAP and control teachers on any characteristic based on both the omnibus test and the individual tests.

Table 2.8. Characteristics of grade 4 teachers, 2008/09 (before Year 2 implementation)

Characteristic	MAP	Control	Estimated difference	*p*-value
Number of teachers	50	35		
Percent female	87.2	87.0	.2	.98
Percent with graduate degree	83.3	76.4	7.0	.480
Years teaching English language arts	10.4	7.2	3.2	.048*
Percent with permanent license	81.3	78.7	2.5	.796
Percent White	99.2	87.8	11.4	.038*
Joint test of difference in student characteristics between MAP and control groups[a] ($\chi^2 = 6.2$, df = 5)				.286

* Difference statistically significantly different from zero at the .05 level.

Note: Means and differences were regression adjusted to account for district effects and weighted by the number of schools in each district. *p*-values are from a two-tailed test of the null hypothesis of equality of MAP and control means.

a. An overall test of the difference between the MAP and control groups based on the teacher characteristics in this table was conducted using a chi-square test. The chi-square test is from a logistic regression model with the binary treatment indicator as outcome and the teacher characteristics in this table as covariates.

Source: Authors' analysis based on the Year 2 teacher survey and district records.

Student characteristics

Table 2.9 compares grade 4 MAP and control students in Year 2.[23] Although the individual test results indicate that the control group had a significantly higher proportion of students who were eligible for free or reduced-price lunch, the overall test of difference between the two groups shows that there was no systematic difference in demographic characteristics or prior achievement, indicating that the randomization successfully created equivalent groups of grade 4 students at baseline. Nevertheless, all these characteristics were used as covariates in the core impact model.

[23] A similar comparison for the Year 2 grade 5 students (table A.5, in appendix A) shows that there were no significant differences between MAP and control students on any characteristic based on either the omnibus test or the individual tests.

Table 2.9. Characteristics of grade 4 students, 2008/09 (before Year 2 implementation)

Characteristic	MAP	Control	Estimated difference	*p*-value
Number of students	1,149	765		
Mean ISAT 2009 reading scale score	202.0 (*n* = 1,068)	203.1 (*n* = 697)	–1.12	.724
Percent eligible for free or reduced-price lunch	49.8 (*n* = 1,143)	63.3 (*n* =759)	–13.50	.012*
Percent White	60.5	62.2	–1.70	.777
Percent with disability	16.9 (*n* = 1,127)	15.0 (*n* = 764)	1.90	.465
Percent proficient in English	96.5 (*n* = 1,135)	97.5 (*n* = 761)	–1.00	.557
Percent male	50.2	52.2	–2.00	.490
Joint test of difference in student characteristics between MAP and control groups[a] *F* = 0.100, df = (11, 27)				1.000

* Difference statistically significantly different from zero at the .05 level.

Note: Means and differences were regression adjusted to account for district effects and clustering of students within schools and weighted by the number of schools in each district. When data are missing data, *n* is the actual number of students used to calculate the average characteristic in each treatment group. *p*-values are from a two-tailed test of the null hypothesis of equality of MAP and control means.

a. An overall test of the difference between the MAP and control groups based on all student characteristics in this table was conducted using an *F*-test adjusted for the randomization of blocks within districts and the clustering of students within schools. The *F*-test is from a two-level logistic regression model with the binary treatment indicator as outcome and the student characteristics in this table as covariates.

Source: Authors' analysis based on the Year 2 student baseline data collected from study districts in spring 2009, when students were in grade 3.

Attrition

Attrition occurs when the outcome data are not measured for all the participants initially randomized to treatment and control groups (What Works Clearinghouse 2008).[24] Differential attrition, or the difference between the treatment and control groups in the proportions of the original sample included in the analysis, can violate the critical assumption of baseline equivalence in experimental designs. If severe enough, it can result in seriously biased impact estimates, threatening the internal and external validity of the study.

There was no differential attrition on the grade 4 ISAT posttest scores (about 5 percent of data were missing for each group) (table 2.10). In contrast, on the grade 4 MAP spring 2010 scores,[25]

[24] In this report, attrition encompasses missing outcome data caused by both sample attrition (for example, a school dropping out of the study) and data attrition (for example, students missing the posttest because they were absent the day the test was administered).

[25] MAP spring 2010 scores were considered missing only if both components (reading and language use scores) were missing.

there was a statistically significantly higher attrition rate in the control group (14.8 percent) than the MAP group (4.2 percent).[26] Attrition on the ISAT scores occurred because some students who were present in the study at the start of Year 2 moved out of the schools before ISAT 2010 testing or were absent the day the test was administered. Attrition on the MAP posttest scores occurred for the same reasons as on the ISAT. Attrition on MAP testing also occurred because any school that dropped out of the study immediately after randomization did not administer any of the MAP tests, resulting in missing MAP data for the entire school, and because some classrooms could not administer the MAP posttests due to technical problems.[27]

Table 2.10. Grade 4 attrition rates on 2010 posttest scores

	Illinois Standards Achievement Test (ISAT) 2010			MAP spring 2010		
Status	**Overall**	**MAP**	**Control**	**Overall**	**MAP**	**Control**
Observed	1,815	1,088	727	1,752	1,100	652
Missing	99	61	38	162	49	113
Total number of students	1,914	1,149	765	1,914	1,149	765
Attrition rate (percent)	5.2	5.3	5.0	8.5	4.2	14.8
Chi-square test of equality of proportions	$\chi^2 = 0.1$, df = 1, p-value = .741			$\chi^2 = 65.4$, df = 1, p-value < .0001*		

* Difference statistically significantly different from zero at the .05 level.

Source: Authors' analysis based on data from the study districts.

When data can be assumed to be missing completely at random (MCAR) or missing at random conditional on the characteristics observed in the sample (MAR), differential attrition is not a problem, because the participants who dropped out can be assumed to be representative of the original sample of the population.[28] Because ISAT testing is a federal and state requirement and each participating district provided a complete student file that included all eligible students and their available data, the study team believes it is reasonable to assume that any missing posttest ISAT results were missing at random. Although (from conversations with district and school administrators on MAP administration) it is reasonable to assume that missing MAP posttest data

[26] The corresponding attrition rates for grade 5 (see table A.6, in appendix A) show a similar pattern: there was no differential attrition on the grade 5 ISAT posttest scores (with 4.3 percent missing for the MAP group and 5.2 percent missing data for the control group) but a statistically significantly higher attrition rate for the grade 5 control group (12.7 percent) relative to the MAP group (4.3 percent) on the MAP spring 2010 scores.

[27] Fewer than four grade 4 control classrooms and fewer than four grade 5 treatment classrooms could not administer the MAP tests because of technical problems.

[28] Under these conditions, the missing data problem is considered ignorable (that is, the factors that led to missingness are unrelated or weakly related to the estimated program impact), and the estimate will be unbiased (Puma et al. 2009).

on other schools were missing at random, this assumption may not be plausible for the missing MAP scores in the school that dropped out and in the school that could not administer the test.[29]

The approach used in the core impact analysis of grade 4 achievement data in chapter 4 (and the core exploratory analysis of grade 5 achievement data in chapter 5) was to impute missing data on both outcomes using multiple imputation.[30] To assess the robustness of the confirmatory findings to other ways of treating missing data, the study team also conducted several sensitivity analyses in which they listwise deleted missing outcomes (see appendix C). Although there was relatively low overall attrition and no differential on ISAT scores, given the moderately high overall attrition rate and differential attrition on the MAP outcomes, it is important to assess the potential bias that missing data may have caused in the estimation of intervention impacts.

The study team investigated the effects of attrition on internal and external validity (using the approach suggested by Hansen et al. 1985)[31] separately for each outcome and grade. To assess internal validity, they compared the ISAT 2009 pretest scores of MAP students with missing outcomes and those of control students with missing outcomes in the same grade. Table 2.11 shows that for grade 4 students with missing ISAT 2010 scores, the MAP and control groups were statistically equivalent on their baseline achievement levels, indicating that attrition on grade 4 ISAT 2010 scores did not pose a threat to the internal validity of the study. For the grade 4 MAP 2010 posttest, however, the control students with missing outcome data had significantly higher baseline mean achievement than the MAP students with missing outcome data, causing higher-achieving students to be underrepresented among control students with observed MAP 2010 scores.[32] This means that listwise deletion in the analysis of grade 4 MAP scores could result in upwardly biased estimated impacts (suggesting that the intervention was more beneficial than it actually was). Although the study team concedes this possibility, the results of the sensitivity analyses (see appendix C) show that listwise deletion resulted in findings that were consistent with those of the core analysis in which missing outcomes were imputed.

To address external validity, the study team compared the average ISAT 2009 pretest score of students with missing outcomes ("dropouts") with the average score of students with nonmissing outcomes ("stayers"), separately for each grade and separately for the ISAT 2010 and MAP 2010 scores. In grade 4 (table 2.12), dropouts had statistically significantly lower prior achievement levels than stayers on both the ISAT and MAP tests.[33] Thus, if dropouts were deleted from the

[29] As Schafer and Graham (2002, p. 152) point out, the missing at random assumption is untestable, because testing it requires "obtaining follow-up data from nonrespondents" or "imposing an unverifiable model." They note that "when the missingness is beyond the researcher's control, its distribution is unknown and MAR is only an assumption."

[30] Multiple imputation rests on the assumption that the missing at random assumption (MAR) holds. In a simulation study, Collins, Schafer, and Kam (2001) show that "in many realistic cases, erroneous assumption of MAR (for example, failing to take into account a cause or correlate of missingness) may often have only a minor impact on estimates and standard errors" (Schafer and Graham 2002, p. 152).

[31] See Borman et al. (2007) for application of this approach in checking the internal and external validity of their experimental evaluation of the Success for All program.

[32] Table A.7, in appendix A, presents analogous comparisons for grade 5 students. Results show that for both grade 5 students with missing ISAT 2010 scores and grade 5 students with missing MAP 2010 scores, the MAP and control groups were statistically equivalent on their baseline achievement levels, indicating that the attrition on ISAT 2010 scores and on MAP 2010 scores in grade 5 did not pose a threat to the internal validity of the study.

[33] In grade 5, a similar analysis (table A.8, in appendix A) shows that for both the ISAT and MAP outcomes, dropouts and stayers had statistically equivalent prior achievement levels, suggesting that in contrast to grade 4, the grade 5 attrition on both outcomes did not pose a potential threat to the external validity of the study.

analyses, higher achieving students would be overrepresented in both groups relative to the original samples of the two groups. This poses a potential threat to the external validity of the study, in that it limits its generalizability to students who are higher achieving than the population from which they were sampled. Although this is a possibility, the sensitivity analyses indicated that listwise deletion of dropouts in each grade yielded results that were similar to the findings from the core impact analysis.

Table 2.11. Illinois Standards Achievement Test (ISAT) pretest scores of grade 4 students with missing 2010 posttest scores in Year 2

	Missing ISAT 2010 scores			Missing MAP spring 2010 scores		
Characteristic	**MAP**	**Control**	**Difference (*p*-value)**	**MAP**	**Control**	**Difference (*p*-value)**
Number of students	61	38		49	113	
Mean ISAT 2009 reading scale score [a]	190.5 (*n* = 35)	186.6 (*n* = 25)	3.9 (.602)	184.6 (*n* = 38)	197.3 (*n* = 96)	–12.7 (.027*)

* Difference statistically significantly different from zero at the .05 level.

Note: n includes only students with nonmissing ISAT 2009 scores. A two-tailed *t*-test for equality of means was used.

a. Results show average scores on the 2009 ISAT assessment administered in the spring before the Year 2 implementation (pretest scores), when grade 4 students in Year 2 were in grade 3.

Source: Authors' analysis based on test scores from the study districts.

Table 2.12. Illinois Standards Achievement Test (ISAT) pretest scores of grade 4 "dropouts" and "stayers" in Year 2

	ISAT 2010 scores			MAP spring 2010 scores		
Characteristic	**Dropouts**	**Stayers**	**Difference (*p*-value)**	**Dropouts**	**Stayers**	**Difference (*p*-value)**
Number of students	99	1,815		162	1,752	
Mean ISAT 2009 reading scale[a]	186.8 (*n* = 60)	203.0 (*n* = 1,705)	–16.2 (.000*)	193.3 (*n* = 134)	203.5 (*n* = 1,631)	–10.2 (.001*)

* Difference statistically significantly different from zero at the .05 level.

Note: Means were weighted by the number of schools in each district. *n* includes only students with nonmissing ISAT 2009 scores. A two-tailed *t*-test for equality of means was used.

a. Results show average scores on the 2009 ISAT assessment administered in the spring before the Year 2 implementation (pretest scores), when grade 4 students were in grade 3.

Source: Authors' analysis based on test scores from the study districts.

Data collection and outcome measures

Table 2.13 summarizes the study's data collection plan for the 2008/09 and 2009/10 school years. Sources of information on fidelity included instructional logs and student engagement surveys on a sample of eight students in each classroom, observations of teachers' instruction, and principal and teacher surveys.[34] Data collection on outcomes included annual student assessment results on the ISAT reading scale and the MAP tests in reading and language usage. Data on principals and teachers were also collected, to measure fidelity of implementation by teachers assigned to the treatment condition.

Table 2.13. Data collection schedule for the Measures of Academic Progress (MAP) impact study, 2008/09 and 2009/10

Data collection element	August	September	October	November	December	January	February	March	April	May
Implementation fidelity										
MAP administrative records[a]	X					X				
Classroom observations		X				X			X	
Instructional logs			X	X	X	X	X	X	X	
Teacher surveys									X	
Student reading performance										
Illinois State Achievement Test reading scale score								X		
MAP assessment composite score[b]		X			X			X		

a. MAP administrative records were collected twice a year, at the conclusion of each semester, to determine the extent to which teachers used MAP data reports and other resources to support classroom differentiation. Year 1 (2008/09) data were collected in January and August 2009. Year 2 data were collected in January and August 2010.

b. MAP assessments in the fall and in winter were administered to students in treatment classrooms only. MAP assessments were administered to both treatment and control students in the spring.

Source: Authors' compilation.

[34] School leader and student engagement surveys were ultimately not used in the fidelity analysis because they did not contain items that directly related to implementation of the MAP core components. Data about support from school leaders (for example, principals and assistant principals) suggested potential reasons for differences in teacher-level implementation fidelity. However, leadership support for MAP does not measure MAP implementation fidelity. The engagement survey contained items that would allow researchers to examine differences in outcomes that were not related to the amount of exposure to differentiated instruction. In hindsight, this measure was not an adequate index of implementation at the student level.

Data on implementation fidelity

Multiple data collection methods were used to describe and assess MAP implementation fidelity. MAP administrative records (such as training attendance data) and web-based computerized reports were used to describe the extent to which NWEA delivered the program to the study schools as intended. Teacher surveys, instructional logs, and classroom observations were used to assess whether teachers in the treatment group implemented core components underlying the MAP training (for example, differentiated instruction practices) to a greater extent than their control group counterparts. Chapter 3 discusses the methods used to describe and assess implementation of the MAP program and presents the findings on fidelity of implementation.

Data on student performance

Students' reading performance was assessed with the ISAT in spring 2010. The ISAT is administered to all Illinois students in grades 3–8 during the spring of each school year. In addition, results of the MAP tests in reading and language usage, administered in spring 2010, were used as a composite measure to assess students' reading and literacy achievement. MAP assessments in reading and language usage were administered in the fall and winter to students in treatment classrooms only. MAP assessments were administered to both treatment and control students in the spring to provide a post-only outcome measure on which to compare students' achievement.

A concern with using the MAP tests to assess the MAP program is overalignment of the tests to the content of the intervention. Overalignment could occur as a result of more frequent administration of the MAP tests to the treatment group than to the control group, MAP teachers' use of terminology or concepts specifically learned from the MAP training program but not ordinarily used in classrooms, and differing testing conditions for treatment and control groups. The MAP assessments include several features to ensure that the tests provide an unbiased measure of students' ability. For instance, the tests are not timed, teachers do not have access to test items, and individual items are not readministered to the same student for two consecutive years. These features limit any advantage a student or teacher might otherwise gain by becoming familiar with the tests over time. In addition, NWEA incorporates procedures to align MAP test items with state content standards and maintain the test's high reliability and validity for predicting state achievement test performance. NWEA trains school-based MAP test proctors to achieve consistency across testing events. As an additional measure to mitigate contamination, NWEA turned off the scoring function on the MAP test for the control group to prevent control teachers and students from seeing their MAP scores and to prevent control teachers from generating MAP reports.

Analytic methods

This section provides an overview of the analytic strategy used to examine fidelity of implementation and the methods used to estimate impacts on student achievement. It describes the analyses conducted, the estimation models used, and the presentation of impact findings and discusses statistical power and adjustments for multiple comparisons. Appendix B provides a detailed description of the statistical models used to estimate impact. Appendix D describes the

imputation procedures used for the implementation fidelity analysis and the impact analysis on student outcomes.

Analysis of implementation fidelity

To assess implementation fidelity, the study team created behavioral indexes of differentiated instruction for each teacher in the MAP and control conditions. To assess the magnitude of the difference between the treatment and control conditions, they divided the difference in group averages on the fidelity indexes by the pooled standard deviation for each index. Cordray and Pion (2006) and Hulleman and Cordray (2009) refer to these standardized values as indexes of the *achieved relative strength* of the contrast. The Achieved Relative Strength Index (ARSI) accounts for deviations from the original treatment model (adherence, when applicable) and treatment–control differences in the delivery and receipt of core MAP components. Hulleman and Cordray (2009) provide formulas for accounting for clustering in deriving these summary measures of implementation fidelity. Direct comparisons between the treatment and control groups were conducted separately for grades 4 and 5.

Analysis of impacts on student achievement

This section describes the confirmatory and exploratory analyses for grade 4 and 5 student outcomes. Appendix C describes the exploratory sensitivity analyses for both grades.

Estimation of overall impacts. Overall impacts were estimated by first conducting a core analysis that included the full analytic sample of eligible grade 4 students and teachers from the 32 participating schools (including the school that dropped out of the study shortly after randomization) and employed a full model that controlled for six baseline student characteristics, five teacher characteristics, the grade 4 school mean prior ISAT reading scale score, and district fixed effects (equation B.1, in appendix B). The six student characteristics were prior ISAT reading achievement, gender, eligibility for free or reduced-price lunch, racial/ethnic minority status, English proficiency status, and disability status.[35] The five teacher characteristics were gender, graduate degree status, teaching experience in English language arts, licensure status, and racial/ethnic minority status).

The covariates used in the core estimation model were chosen before conducting the analysis. They were selected because they are commonly used in evaluation studies in education and are known to be correlated to some degree with student performance. Although tests for baseline equivalence revealed that there were no systematic differences between the MAP and control groups, the study team included these covariates in order to increase the precision of the estimates. To assess how the choice of covariates influenced the estimated impacts from the core analysis, the study team also explored three alternative covariate specifications:

[35] For both the MAP and the ISAT outcomes, the 2009 ISAT reading scale score was used as a pretest measure because no pretest on the MAP assessment was available. (The MAP tests were administered to the treatment group on two other occasions—fall 2009 and winter 2010—before the spring 2010 testing, but they were not administered to the control group. Furthermore, these tests were administered after the study was already underway and, for that reason, were deemed inappropriate to use as pretests even for the treatment group.) Although the ISAT pretest and the MAP assessments are different instruments, they share a common content domain (reading), and the MAP language usage test is in a related domain (language usage). Moreover, the ISAT pretest scores and the MAP scores are highly correlated, as shown in table A.9 (grade 4) and table A.10 (grade 5), in appendix A.

- An "unadjusted" model that included only district fixed effects
- A "pretest" model that included only the student ISAT pretest score and the school mean ISAT pretest score for the grade under analysis
- A "student + school covariates" model that included the covariates in the pretest model plus all student characteristics included in the full model.

Appendix B describes these models and presents the results obtained from each.

Inclusion of the full analytic sample (including the no-show school) in the core analysis was made possible by filling in missing outcome and covariate data through multiple imputation.[36] To investigate the sensitivity of the core analysis results to the handling of missing data, the study team conducted three sets of sensitivity analyses that used subsets of the full analytic sample and employed either multiple imputation or listwise deletion. These sensitivity analyses are described in appendix C and summarized in tables D.2 and D.3. They include analysis of the subset of students with observed posttest scores from the 32 study schools, analysis of all students (with and without posttest scores) from all schools except the no-show school, and analysis of students with observed posttest scores from the all schools except the no-show school. The three sets of analyses used only the full model (equation B.1, in appendix B).

Several important features of the models used in estimating the overall impacts of the MAP program on student achievement are noteworthy:

- The models were two-level models[37] (students nested within schools) that accounted for the dependencies among students in the same school (thereby producing correct standard errors and more efficient estimates than those of ordinary least squares models) and allowed for the examination of variation in student performance separately at the student and school levels.
- By including district indicators as a fixed effect, the models controlled for variations in student performance attributable to both observable and unobservable differences across districts. As a consequence of using districts as fixed effects, the generalizability of the study findings is limited to the districts included in this study.
- The models incorporated interactions between the district indicators and the treatment indicator, thereby taking advantage of the block randomization and providing estimates of district-specific impacts, which were then pooled into a weighted average (using the number of study schools in each district as weights) to produce overall impacts.[38]
- Although the treatment effect (that is, the coefficient of the treatment indicator in equation B.1, in appendix B) was allowed to differ across districts, the association between the outcome and each of the baseline student, teacher, and school characteristics was assumed to

[36] Details of the imputation procedure are in appendix D.

[37] Although students are nested within classrooms that are nested within schools, most schools had very few classrooms, making it difficult to assess classroom variability within schools with sufficient power. In each grade, 78 percent of schools had three or fewer classrooms in grade 4, and 75 percent had three or fewer classrooms in grade 5 (figures A.1 and A.2, in appendix A). Therefore, in the model as estimated, variability in achievement between classrooms within schools is confounded with variability between schools.

[38] Appendix B presents district-specific impacts. These estimates should be interpreted with caution because of the lack of power to detect true impacts with the relatively small sample sizes (four of the five districts had no more than four schools). These district estimates are shown in tables B.2 and B.3 for grade 4 and tables B.4 and B.5 for grade 5.

be homogeneous across all districts (that is, no interactions between the district indicator and baseline characteristics were included in the models). This assumption was necessary to obtain estimates of these associations because in some districts there were empty (or almost empty) cells for categories of some of these covariates.[39]

- Wherever baseline characteristics were used, they were centered on their corresponding grand means across the sample (including level-1 teacher covariates, which were centered on their means across all students in the sample, and the school mean ISAT pretest, which was centered on the mean of the school means). Grand-mean centering was used because the substantive interest lies in estimating the effect of the level-2 treatment indicator while controlling for differences in level-1 covariates (see Enders and Tofighi 2007). The choice of centering has implications for the interpretation of the impact estimates. Specifically, the impact estimates represent the effect of the MAP intervention adjusted for student-level covariates (namely, the baseline student and teacher characteristics).[40] Moreover, the adjusted (MAP or control) mean is the achievement level of an average student who attends an average school (assigned to the MAP or control condition) and is taught by the teacher of an average student.

Presentation of impact findings

The overall impacts (table 4.1 for grade 4 and table 5.1 for grade 5) include the regression-adjusted mean for the MAP group, the regression-adjusted mean for the control group, and the overall impact (the difference between the two means). These estimates are averages of corresponding district-specific estimates, weighted by the number of schools in each district. Also reported are the standard error of the impact estimate, the *p*-value for testing the equality of the MAP and control means, and the effect size of the impact obtained by dividing the impact by the standard deviation of the outcome for the control group of grade 4 (or grade 5) students.[41]

Statistical power

A statistical power analysis conducted during the design phase of the study (geared toward the analysis of Year 2 outcomes) showed that 30 schools[42] were needed for a minimum detectable effect size of 0.20 using a two-tailed test with 80 percent power and 5 percent significance

[39] For example, in grade 4, in one district there were no male teachers, and in some districts all or almost all students were English proficient. Such cases precluded the estimation of the district-specific effects of teacher gender and race/ethnicity, and they led to either no estimate or unstable estimates of student English proficiency status.

[40] Group-mean centering would have yielded impact estimates that were not adjusted for the level-1 covariates (Raudenbush and Bryk 2002, p. 142).

[41] Use of the control group standard deviation kept effect size calculations free of any effects of the intervention on variation of outcomes. The standard deviations of the outcomes for the grade 4 treatment and control groups were very close to each other (25.8 for the MAP group and 27.5 for the control group on the ISAT 2010 scores and 14.6 for the MAP group and 14.5 for the control group on the MAP 2010 composites scores), so that using a pooled estimate of the treatment and control standard deviations instead of the control standard deviation resulted in very similar effect sizes.

[42] Although only 30 schools were needed to achieve 80 percent power (based on the parameters assumed for the power calculation), the study had 32 (instead of 30) schools randomized to treatment conditions. Of the 32 schools, less than four schools withdrew from the study immediately after randomization.

level.[43] This minimum detectable effect size was based on the following additional assumptions: a two-level cluster randomized design, an intraclass correlation coefficient (ICC) of 0.13,[44] a level-1 covariate (student pretest) and level-2 covariate (school mean pretest) that explain 75 percent of the variability in achievement at their respective levels, and an average cluster (school) size of 80 students (that is, four classes with 20 students each).

The study team calculated, separately for grades 4 and 5, the study's actual minimum detectable effect size for the overall impact of the MAP program on the two achievement outcomes (ISAT score and MAP composite score) by replacing these values by the values observed in the study (keeping the power and significance level the same).[45] In grade 4, the study team used 32 schools; an average cluster size of 60 students a school;[46] an ICC of .09 for the ISAT score and .09 for the MAP score;[47] and student-level variation of 57 percent for the ISAT outcomes and 45 percent for the MAP outcomes and school-level variation of 82 percent for ISAT outcomes and 77 percent for MAP outcomes, explained by a student pretest and a school-level pretest.[48] From these data, the study was able to detect impacts of 0.16 standard deviation on the ISAT outcome and 0.18 standard deviation on the MAP outcome. Thus, in grade 4 the actual detectable effect sizes were slightly smaller than the planned minimum detectable effect size of 0.20.

In grade 5 the study team computed the study's actual minimum detectable effect size using 32 schools, an average cluster size of 56 students a school, and ICC of .07 for the ISAT score and .09 for the MAP score, and within-school variation of 49 percent on the ISAT outcome and 41 percent on the MAP outcome and between-school variation of 86 percent on the ISAT outcome and 59 percent on the MAP outcome, explained by student and school-level pretests. From these values, the study team found that the actual minimum detectable effect size was 0.14 for the

---

[43] This means that 0.20 standard deviation is the smallest true effect that the study can detect with 80 percent power at the 5 percent significance level.

[44] Our selection of .13 as the estimated ICC for the power calculation was based on Hedges and Hedberg's (2007) compilation of ICC values for academic achievement that can be used for planning group-randomized experiments. The article found that when a pretest was used in the impact model, the ICC values for reading achievement in grades 3–5 ranged from 0.113 to 0.135. When a pretest and demographic covariates were included in the impact model, ICC values for reading achievement in grades 3–5 ranged from 0.083 to 0.101. We selected .13 because it was a more conservative estimate.

[45] The minimum detectable effect size was calculated using equation 4 from Bloom, Richburg-Hays, and Black (2007, p. 34): $MDES = M_{J-K}\sqrt{\frac{\rho(1-R_2^2)}{P(1-P)J} + \frac{\rho(1-\rho)(1-R_1^2)}{P(1-P)Jn}}$, where J is the number of schools randomized; n is the number of students sampled in each school; K is the number of school-level covariates included in the model; ρ is the unconditional ICC (which represents the proportion of the total variance in the outcome that lies between schools when no covariates are used in the model); P is the proportion of schools randomized to treatment; R_1^2 is the proportion of within-school variance explained by the student- and school-level covariates; R_2^2, is the proportion of between-school variance explained by the student- and school-level covariates; and M_{J-K} is a multiplier of the standard error of the impact estimate that accounts for the degrees of freedom $(J-K)$.

[46] The average cluster size (the number of students sampled in each school) was 1,914 students divided by 32 schools for the Year 2 cohort of grader 4 students, and 1,806 students divided by 32 schools for the Year 2 cohort of grade 5 students.

[47] These ICC values are from modeling each of the two outcomes using unadjusted two-level models with a random intercept and district fixed effects but no covariates (see the unadjusted models in tables B.2–B.5 in appendix B).

[48] In grade 4, 57 percent of within-school variability on the ISAT scores and 45 percent of within-school variability on the MAP posttest scores was explained by the student-level and school-level pretests, the treatment indicator, and the district indicators. Eighty-two percent of between-school variability on the ISAT posttest scores and 77 percent of between-school variability on the MAP posttest scores was explained by these covariates. (See the pretest model in tables B.2 and B.3, in appendix B. Tables B.4 and B.5 show the proportions for grade 5 students.)

ISAT outcome, which was smaller than the prespecified minimum detectable effect size of 0.20, and 0.22 for the MAP outcome, which was slightly larger than the planned minimum detectable effect size.

These findings indicate that the study achieved its goal of detecting true overall impacts of at least 0.20 standard deviation for the confirmatory analysis of the overall impacts on the ISAT and MAP outcomes in grade 4. In grade 5, the study achieved the desired precision for the overall impact estimate on the ISAT outcome and was able to detect a minimum effect size that was slightly larger than the prespecified threshold of 0.20 for the MAP outcome.

Adjustment for multiple hypothesis testing

For the confirmatory analyses of grade 4 achievement, testing the impact of the MAP program resulted in two comparisons, one for the ISAT reading scores and one for the MAP composite scores. With two hypothesis tests each tested at the 5 percent significance level, the probability of declaring at least one of the tests significant when in fact it is not (that is, at least one false positive) is roughly 9.8 percent ($= 1 - [1 - 0.05])^2$), assuming independence. Because only two comparisons were made, the study team planned on using a simple multiplicity adjustment procedure, such as the Bonferroni or Sidak method.[49] None of the estimated overall impacts for the confirmatory analyses for grade 4 turned out to be statistically significant at the 5 percent level, however. It was therefore not necessary to adjust for multiple testing. Neither of the statistical tests of the overall impacts for grade 5 yielded statistically significant results and therefore could not have yielded any spurious findings, making adjustment for multiple testing unnecessary.

Study limitations

The key questions addressed in this report pertain to implementation fidelity and program impact on grade 4 reading achievement after Year 2 of the cluster randomized trial. The study's greatest strength lies in the randomization of schools to treatment at each grade level, which allows causal inferences to be drawn. However, it is important to point out some caveats on the design and analysis of Year 2 outcomes that restrict the conclusions that can be drawn from this report.

First, the schools that were recruited and volunteered to be part of the study are not necessarily representative of the schools that are currently using or are intending to implement the MAP program. Because participation was voluntary, the observed effects in this study could be different from what might be observed in actual use. In actual use, districts may not always find it feasible or desirable to fund the comprehensive package of four one-day sessions of MAP training, on-site visits, and intermittent conference calls with NWEA trainers. Districts may not have the internal capacity to administer three tests a year or may simply prefer not to do so. These types of district and school decisions could alter effectiveness. The conclusions drawn

[49] These procedures control for the familywise error rate but tend to be conservative when the number of comparisons is large, which is not the case in the confirmatory analysis. The two methods result in roughly the same adjustments. The Bonferroni method results in a significance level of .025 (=.05/2); the Sidak method set the significance level at .0253 ($=1 - (1–.05)^{1/2}$).

from this report apply only to the schools in the study. No attempt is made to generalize the findings to a larger group of schools and districts.

Second, provision of the treatment to one grade but not another within the same school may have increased the potential for control group teachers and students to be exposed to the MAP program. School administrators were encouraged to attend MAP training sessions and support their teachers in MAP implementation and use of MAP results. Although administrators were asked to refrain from discussing the program with control group teachers, it is possible that they may have shared general knowledge gained through the MAP training with control group teachers, thereby influencing changes in control teachers' instructional practices. Exposure to treatment, or treatment contamination, could reduce the magnitude of MAP program impacts between the two study groups. Although it is not possible to test for all possible sources of contamination, supplemental analyses found no evidence of between-group contamination. These analyses are included in appendix L.

Chapter 3: Implementation

This chapter examines the extent to which (a) core components of the MAP model were implemented as planned by NWEA staff and (b) teachers participated in MAP training and consultation, used MAP data and resources, and used core aspects of the MAP program model in their classes. Because the outcome analyses examine the relative effects of MAP on achievement outcomes using an intent-to-treat model, the implementation analyses focus on the average level of implementation across all schools within a given grade. When there is variability in the implementation of MAP components at the teacher level, the degree of variability is reported. The analyses in this chapter describe what happened when NWEA delivered the MAP program components and teachers attempted to implement and use these program elements. It does not attempt to explain variation in the extent to which schools and teachers implemented various components of the program.

The chapter is divided into four sections. The first section presents information on the extent to which the MAP program was implemented by NWEA and by MAP teachers. These descriptive analyses refer to program-specific implementation fidelity. The second section examines the extent to which MAP classes differed from control classes on a key construct underlying the MAP program—differentiated instructional practices.[50] The third section briefly discusses the exploratory analysis of the effects of teacher experience and the academic composition of the classes. The last section summarizes the chapter's main findings.

This study entailed separate experiments for grade 4 or grade 5. The implementation analyses for teachers are presented separately for each experiment.

This study addresses two questions about intervention implementation fidelity:

- Were MAP resources (training, consultation, web-based materials) delivered by NWEA and received and used by teachers as planned?
- Did MAP teachers apply differentiated instructional practices in their classes to a greater extent than their control counterparts?

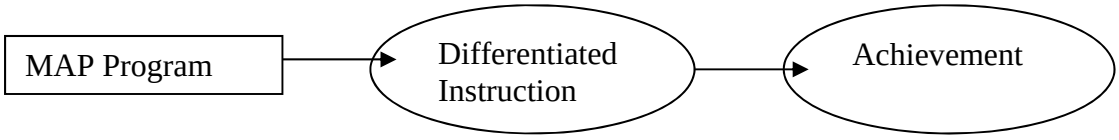
The first question entails a program-specific implementation assessment; the second question entails between-group comparisons regarding the core components of the intervention model. Specifically, in addition to assessing if the MAP program was implemented as planned, the study team broadened the definition of intervention fidelity by assessing the extent to which MAP teachers engaged in key behaviors (core components) to a greater extent than their non-MAP counterparts. The study team assessed treatment contrast between the two study conditions. Treatment contrast measures the extent to which treatment group teachers engage in practices more than, less than, or the same as teachers in the control group. The model of causality acknowledges that the control or business-as-usual condition can exhibit MAP-like instructional

[50] Several implementation variables that were initially specified did not properly represent the idea of implementation fidelity. For example, data about support from school leaders (for example, principals and assistant principals) suggested potential reasons for differences in teacher-level implementation fidelity. However, leadership support for MAP does not measure MAP implementation fidelity. For this reason, researchers focused on the extent to which teachers actually implemented the MAP program within their classrooms.

practices that are not the result of contamination but the result of generalized diffusion of innovations (see Shadish, Cook, and Campbell 2002). Thus, the causal effect of the treatment condition on outcomes must be considered relative to the causal components embedded in the control condition associated with control group outcomes. An Achieved Relative Strength Index (ARSI) was used to index this difference (see Cordray and Jacobs 2005; Cordray and Pion 2006; Hulleman and Cordray 2009). Fidelity measures and indexes of achieved relative strength are described in more detail below.

Figure 3.1 depicts the model of change underlying the MAP program. As depicted, the MAP intervention—composed of teacher training, consultation services, multiple computer-adaptive benchmark assessments, and online instructional resources—is supposed to enhance teachers' use of differentiated instructional practices, use of which is supposed to enhance student achievement.

Figure 3.1. Measures of Academic Progress (MAP): model of change



The logic (or operational) model underlying the MAP program (figure 3.2) specifies that complete implementation requires that NWEA deliver specific services (training, consultation, computer-adaptive testing) and online instructional resources to teachers and schools. For their part, teachers are required to attend the MAP-based training sessions and to access additional NWEA services and resources. Teachers' use of periodic formative assessment reports is supposed to guide their formation of subgroups of students based on homogeneous levels of reading readiness (reading ability). NWEA provides online resources (for example, information on Lexiles, goal setting, and booklists) to assist teachers in tailoring instructional materials to meet the needs of these subgroups. In addition to attending training sessions and using, as needed, follow-up consultation, teachers are expected to access and use these resources.

To ensure that teachers (and school leaders) are equipped with the knowledge and skills needed to use data and differentiate instruction, NWEA provides multiple services and resources. During the two-year implementation period for this study, teachers could engage in up to 12 MAP-relevant activities and resources. The sequencing of these activities is displayed in table 3.1. The next section describes the 12 program components.

Figure 3.2. Logic model for Measures of Academic Progress (MAP)

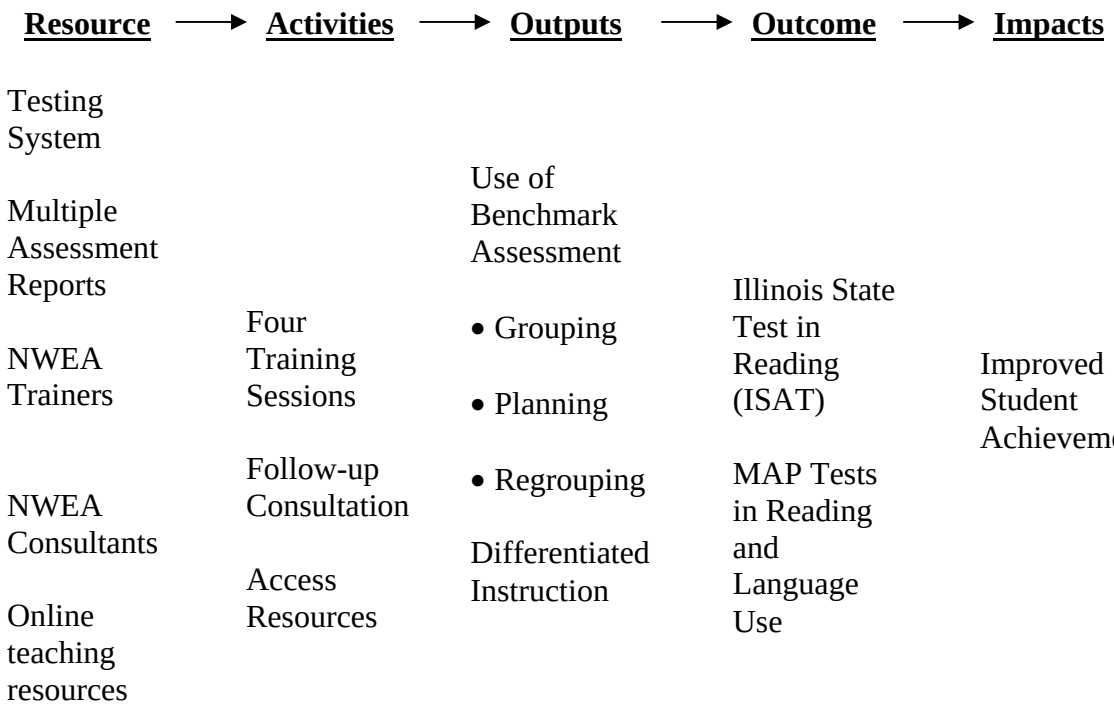


Table 3.1. Sequencing of Measures of Academic Progress (MAP) program components

Component	2008					2009					2009					2010				
	8	9	10	11	12	1	2	3	4	5	8	9	10	11	12	1	2	3	4	5
MAP training sessions																				
Data Administration	1	1									a	a								
Stepping Stones			2	2									a	a						
Climbing the Data Ladder					3	3									a	a				
Growth and Goals										4										a
On-site consultation	5										9									
MAP data use: grouping	6										10									
MAP resource use: data meaning	7										11									
MAP resource use: lesson planning	8										12									

Note: Numbers in body of table refer to activity numbers. Numbers in boxhead indicate months.

a. Training for new Year 2 teachers.

Source: Authors' compilation.

During Year 1 there were eight opportunities for teachers to implement aspects of the MAP program. Teachers were supposed to attend four training sessions (Activities 1–4). They could engage NWEA staff in on-site consultation (Activity 5), use MAP resources for grouping students (Activity 6), use MAP resources to align instruction with test results (Activity 7), and use MAP resources to tailor their lesson plans (Activity 8). The same eight activities were available to teachers who joined the MAP treatment group in Year 2. Teachers who remained in the study both years had four additional opportunities to use MAP resources (see chapter 5). For the majority of MAP teachers, full implementation entailed participation in 12 activities.

Were MAP resources delivered by NWEA and received and used by teachers as planned?

To assess the extent to which NWEA met its programmatic responsibilities and determine whether teachers engaged in MAP-relevant activities, the study team used NWEA administrative and web-based computerized records to document the delivery and receipt of MAP training, consultation, and teachers' use of MAP materials and resources. These records included teacher-level attendance logs for training and consultations and records of individual teachers' use of MAP resources. These records list all individuals—including non-MAP individuals—who received training or consultation or used MAP resources and materials (Lexiles, goal setting, and booklists). In addition, questions on the annual teacher survey provided data on the extent to which teachers used MAP resources for grouping and regrouping students and whether they used MAP resources in planning their lessons.

Implementation by Northwest Evaluation Association

Implementation of the MAP program at the classroom level requires NWEA to provide essential resources (for example, computer-adaptive testing in each school, web-based teacher resources); schedule and deliver the four training sessions; and provide consultation services, on request of school leaders or teachers. NWEA's role in implementing the MAP program began in August 2008. The bulk of NWEA's responsibilities for implementing the MAP program were undertaken in Year 1. In Year 2 NWEA provided supplemental training of new teachers and continued to provide consultation services. This section summarizes NWEA's implementation performance in Year 1 and describes its activities in Year 2.

Year 1. NWEA was successful in providing the equipment needed for computer-adaptive benchmark testing as planned in all participating MAP schools. Testing was completed on schedule, with minor departures from the plan, and test results made available to teachers. Web-based resources (described later in this report), designed to supplement training and facilitate alterations in instructional practices, were continuously available throughout the implementation period. Through the scheduled training sessions and consultative visits, participating teachers had multiple contacts with NWEA training staff during the school year.

For this study, NWEA trainers provided all the training and consultative sessions for the participating schools. Each NWEA trainer was assigned to deliver training and consultation to all the study participants within a particular district. Before delivering MAP training to the schools, the NWEA trainers underwent extensive training and received NWEA MAP training

certification. In addition, these trainers were given access to extensive facilitator notes and materials to support consistent implementation across schools.

In Year 1 NWEA conducted the intended training sessions and provided consultative services. As planned, three of the training sessions (Administering MAP, Stepping Stones to Data, and Climbing the Data Ladder) were offered between August and December 2008. The fourth session, on assessing growth and goals, was held, as planned, in May and June 2009. NWEA training staff conducted 28 days of training. At the request of school officials and teachers, they provided 43 days of consultation, most of it (32 sessions) between January and June 2009.

During Year 1, 98 percent of MAP teachers received at least one training session from NWEA, and 90 percent received at least one consultation session. Overall, 988 training or consultation contacts were recorded for teachers and school leaders at the participating schools, two-thirds of them with teachers. About half of the contacts with teachers (46 percent) were associated with one of the four scheduled training sessions; the remaining contacts (54 percent) were the result of requests by school personnel for consultation services. The content of this consultation was not evenly distributed across the topics covered by the formal training sessions. Of the 358 consultation contacts, 303 (85 percent) occurred after the third training session (Climbing the Data Ladder), which was directed at concepts and practices associated with differentiating instruction. The remaining contacts occurred following the training session on data use and interpretation (Stepping Stones).

Year 2. Having established the MAP testing procedures within schools during Year 1 and provided at least some training to all MAP teachers, NWEA's presence in the schools was reduced in Year 2. NWEA focused on training new MAP participant teachers and, in some districts, individuals not participating in the study (for example, grade 3 teachers and support staff). NWEA scheduled at least 21 days of MAP training and 37 consultation sessions.[51] For MAP program teachers, 140 training and consulting contacts were recorded. Unlike in Year 1, when the balance between training and consultation was approximately equal, in Year 2 training accounted for 7 (5 percent) of the 140 contacts, with the balance (95 percent) devoted to consultations. Because most teachers received MAP training the previous year, it is not surprising that only four teachers (5 percent) received one or more training sessions in Year 2. Of the 16 new MAP teachers, 3 (19 percent) received no MAP training. With respect to consultations, 54 (62 percent) of MAP teachers in Year 2 received one or more consultation sessions; 10 (63 percent) of new MAP teachers received one or more consultations.

Teacher-level implementation

At the heart of the MAP program is the classroom teacher. For the program to be effective, NWEA has to implement it properly and teachers have to use the MAP components and resources. The training sessions and consultation services are intended to prepare teachers to use MAP resources to make data-based decisions on content, processes, and products in tailoring their instruction to the needs of their students.

[51] Participation by specific teachers and administrators in training was indicated for 7 of the 21 scheduled sessions. Planned training sessions for District 1 were cancelled for the Stepping Stones, Climbing the Data Ladder and Growth and Goal Setting sessions.

Table 3.2 summarizes participation rates for each of the 12 MAP components in Year 2. The MAP components are conceptualized as opportunities to participate, allowing participation across the 12 components (and across both program years) to be characterized as a "dose" of MAP services and resources. For this reason, the 16 teachers who joined the study in Year 2 are included in calculating all rates. A dose index is presented following this discussion of component-wise participation rates.

Table 3.2. Teacher participation rates in Measures of Academic Progress (MAP) activities in Year 2 (percent)

Component	Activity	Grade 4 (*n* = 50)	Grade 5 (*n* = 37)
NWEA training	Session 1: Administrative Data System	78	70
	Session 2: Stepping Stones: Using Data	72	62
	Session 3: Climbing the Data Ladder: Differentiating Instruction	66	68
	Session 4: Growth and Planning	70	65
	Attended all training sessions	56	43
	Attended no training sessions	22	19
NWEA consultation	Any consultation in Year 1	66	62
	Any consultation in Year 2	60	65
MAP web-based resources (Lexiles, goal setting, and booklists)	At least three uses of online resources: Year 1	60	54
	At least three uses of online resources: Year 2	34	46
Grouping students	At least some use of MAP data for grouping students: Year 1	48	49
	At least some use of MAP data for grouping students: Year 2	60	68
Planning lessons	At least some use of MAP resources for planning lessons: Year 1	36	51
	At least some use of MAP resources for planning lessons: Year 2	90	81

Source: Authors' analysis based on Year 2 (2009–10) data from the study districts and the Northwest Evaluation Association.

MAP training. Teacher training entails four one-day training sessions offered throughout the school year by NWEA. The four sessions include:

- Information on the administration of MAP testing (called MAP Administration)
- Guidance on interpreting the results of MAP testing (called Stepping Stones to Data)
- Information, guidance, and practice in applying the data to alter instructional practices. (called Climbing the Data Ladder)
- Use of data for assessing growth and goals (called Growth and Planning).

Of the 87 total MAP teachers included in this study, 71 (82 percent) were eligible to receive training in Year 1; the other 16 teachers (18 percent) were new to the study in Year 2. To index the overall participation rates, training in Year 1 and Year 2 were considered equivalent. Table 3.2 indicates that participation was fairly consistent across the four training session: more than half (56 percent) of grade 4 teachers and less than half (43 percent) of grade 5 MAP teachers completed all four training sessions. Twenty-two percent of grade 4 and 19 percent of grade 5 MAP teachers received no MAP training.

Consultation. Teachers can receive follow-up consultation with NWEA staff on each of the four training sessions on demand. The extent to which a teacher uses consultation services is left to the discretion of teachers and school leaders. Consultation is available throughout the school year, in both years of the study. NWEA does not specify how many times teachers should use these consultation services.[52] In this study, 68 percent of grade 4 and 62 percent of grade 5 MAP teachers received at least one consultation session in Year 1. In Year 2, 60 percent of grade 4 and 65 percent of grade 5 MAP teachers received at least one consultation.

Use of web-based resources. To help teachers align instructional materials with test results, NWEA provides online resources that are available only to MAP teachers. These resources include information on Lexiles, goal setting, and booklists. Sixty percent of grade 4 and 54 percent of grade 5 MAP teachers used these web-based resources in Year 1 (see table 3.2). These rates dropped to 43 percent for grade 4 teachers and 46 percent for grade 5 teachers in Year 2.

Use of MAP data to group and regroup students. The NWEA computer-adaptive assessment allows teachers to monitor the progress of students throughout the school year. The assessment is intended to serve as a vehicle for data-based formation of subgroups of students with similar reading levels. Using these data to group students is a key element in the logic model underpinning the MAP program. Data are supposed to be used to group and regroup students throughout the year.

During the two-year study period, teachers had multiple opportunities to use data to group students. In Year 1 about half of teachers (48 percent in grade 4 and 50 percent in grade 5) made at least some use of MAP data for grouping students. In Year 2 these rates rose to 60 percent for grade 4 teachers and 68 percent for grade 5 teachers.

Use of MAP data for lesson planning. Modification of instructional practices may be needed to meet the needs of various subgroups of students. The extent to which teachers used MAP resources to guide the planning of their lessons represents an important program activity. In Year 1, 72 percent of grade 4 teachers and 51 percent of grade 5 MAP teachers reported using MAP data in planning lessons. These rates rose to 90 percent among grade 4 and 81 percent among grade 5 MAP teachers in Year 2.

Dose levels. Because the MAP program has several program components, the program-specific implementation dose is indexed as the extent to which teachers participated in each of these components in Years 1 and 2. The teacher-level index of implementation dose is a unit-weighted

[52] Teachers exhibited variability in the number of consultation services they used. They also varied in the number of times they accessed online resources and assessment reports. Because NWEA does not specify how many times teachers should use these resources, researchers defined participation in a component as participating in it at least once.

count variable, ranging from no participation (0) to full participation (12).[53] For consistency across indexes of program implementation and use variables, the study team divided the total count by 12, creating a new range of 0 (no dose) to 1 (full dose). The reliability estimate (alpha) for the MAP dose index is 0.853.

Despite the relatively consistent levels of participation seen in table 3.2 (generally 50–70) across the 12 MAP components, table 3.2 shows considerable variability in the proportion of the 12 components in which teachers participated. Table 3.3 displays the distributions of MAP doses for teachers in grades 4 and 5.

Table 3.3. Measures of Academic Progress (MAP) dose levels for MAP teachers in Year 2

Dose level	Grade 4 teachers (*n* = 50)			Grade 5 teachers (*n* = 37)		
	Frequency	Cumulative frequency	Cumulative percentage	Frequency	Cumulative frequency	Cumulative percentage
0 to .2	5	5	10	6	6	16
>.2 to .4	5	10	20	4	10	27
>.4 to .6	13	23	46	6	16	43
>.6 to .8	10	33	66	6	22	59
>.8 to1.0	17	50	100	15	37	100
Mean	0.618			0.617		
Median	0.663			0.655		
Standard deviation	0.267			0.324		

Note: Dose range is 0 (no dose) to 1(full dose).

Source: Authors' analysis based on Year 2 (2009–10) data from the study districts and the Northwest Evaluation Association.

Some MAP teachers from each grade participated in no MAP-relevant activities. Twenty percent of teachers in grade 4 and 27 percent of teachers in grade 5 participated in 40 percent or less of program components. In contrast, 54 percent of grade 4 teachers and 57 percent of grade 5 teachers participated in 60 percent or more of the program. On average, across both grades, the dose level for MAP teachers was 0.62 (standard deviation of 0.27 for grade 4 and 0.32 for grade 5).

Participating school districts differed in the extent to which schools and their teachers implemented MAP components. Table 3.4 summarizes the average teacher dose for each of the five participating districts. To protect the identity of each district, the districts are labeled from 1 to 5, consistent with district labels used to report intervention effects.

[53] Except for the explicit expectation that teachers attend all four MAP training sessions, the logic model underlying the MAP program is not specific enough to weight the importance of the other components (consultation, use of resources for grouping, planning and aligning instruction to reading level). Researchers assumed a unit (1.0) weighted approach to scaling program-specific implementation dose. The index records any use of consultative services (one or more times), assessment reports, and online MAP resources.

Table 3.4. Measures of Academic Progress (MAP) dose by school district

Grade	Summary statistic	District				
		1	2	3	4	5
4	Average dose	.525	.472	.764	.769	.625
	Standard deviation	0.251	0.048	0.243	0.235	0.300
	Number of teachers	20	3	6	9	12
5	Average dose	.558	.395	.933	.771	.583
	Standard deviation	0.351	0.185	0.091	0.315	0.226
	Number of teachers	20	4	5	4	4

Source: Authors' analysis based on Year 2 (2009–10) data from the study districts and the Northwest Evaluation Association.

The average MAP dose for teachers within districts ranged from .472 to .769 in grade 4 and .395 to .933 in grade 5. Overall, district accounted for 16.3 percent of the variance in MAP dose in grade 4 and 22.8 percent of the variance in grade 5. The average dose levels are similar across grades (except in District 3).

MAP training crossovers. In implementing educational interventions within an experimental context it is possible for teachers in the control condition to access key aspects of the intervention. It is conceivable that teachers from the control group in a school could have attended MAP training sessions or participated in consultation sessions. In experimental studies, these individuals are called crossovers. NWEA records of attendance at training and consultation sessions revealed that no more than three control teachers (out of 79) received some MAP training in Year 1. In Year 2 there were no instances of control teachers receiving either MAP training or consultation.

Because this study was conducted over two years, it was possible for teachers to be reassigned to a different group. Of particular interest are MAP teachers in Year 1 who were reassigned to the control condition in Year 2. In the intent-to-treat sample for Year 2, at most three MAP teachers were assigned to the control condition. The dose scores for these teachers were between 0.650 and 0.850.

Did MAP teachers apply differentiated instructional practices to a greater extent than their control counterparts?

As depicted in figure 3.2, the logic model for MAP specifies that student achievement will be affected by the outputs of the program-specific MAP activities—namely, changes in the extent to which teachers implement aspects of differentiated instruction. This gives rise to the second major question about intervention implementation fidelity, which examines whether differentiated instruction was used more by MAP teachers than by their control group counterparts. MAP-induced changes in instructional practices can be considered a primary outcome for this randomized controlled trial; they can also be regarded as part of the causal chain embodied in the MAP program. As shown in the logic model for the MAP program, instructional practices are regarded as outputs of the delivery and receipt of MAP activities, resources, and processes. In the context of an intent-to-treat model for student outcomes, these outputs (altered

instructional practices) represent the part of the causal chain embodied in the MAP program. However, because the primary objective of this randomized controlled trial is to assess the effects of MAP on student achievement, we did not include changes in MAP-prescribed teacher practices as part of the intent-to-treat model of analysis for student outcomes.

Because teacher practices can be influenced by factors other than the MAP program, the study team assumed that control teachers could acquire and exhibit MAP-like skills. Therefore, the extent to which teachers adopted the core components (for example, differentiated instruction as represented by student grouping and regrouping, tailored instructional strategies, use of alternative content domains) needs to be assessed in both MAP and control conditions (comparative assessment of MAP implementation). Assessing teacher use of differentiated instruction using variables that are common and applicable across conditions allows the differences between conditions on the key MAP (causal) variables to be summarized. This difference is quantified as the Achieved Relative Strength Index (ARSI) of the intervention contrast.

By itself, program-specific treatment fidelity does not indicate the strength of the intervention: It is only in comparison with the control condition that the achieved relative strength of the intervention can be determined. The study team quantified achieved relative strength as the standardized difference in the adoption composites in the treatment and control conditions. As with conventional effect sizes, the effect size measure of achieved relative strength is expressed in standard deviation units. The ARSI is based on Hedges's *g*, with a correction for clustering in the classroom (Hedges 2007):

$$g = \left(\frac{\bar{X}_1 - \bar{X}_2}{S_T}\right) \times \left(1 - \frac{3}{4(n_1 + n_2) - 9}\right) \times \sqrt{1 - \frac{2(n-1)p}{N-2}}$$

where

$\bar{X}_1$ = mean composite for MAP teachers

$\bar{X}_2$ = mean composite for non-MAP teachers

S_T = pooled within-group standard deviation

n = average cluster size

ρ = intraclass correlation[54]

N = total sample size.

The ARSI was calculated for each of the scales and three composites that were constructed to measure differentiated instruction (see the later discussion of data sources).[55]

[54] Because there were relatively few teachers in each group and each grade, researchers assumed that the ICC was 0.

[55] The ARSI represents a new use for the classic effect size indicator of the magnitude of outcome effects. Benchmarks for small, medium, and large ARSI values are not available. The meaning of the size of the ARSI can be understood simply as standard deviation units. A value of 1.0 means that the causal difference in conditions is one standard deviation; the larger the value, the greater the separation of the groups, on average. Cohen's U3 measure of nonoverlap could be invoked to specify the percentile of the null distribution in which the treatment group mean is located.

Multiple operationalizations of differentiated instruction

Although interest in differentiated instruction dates back to the early 1950s, the explicit measurement of differentiated instructional practices lags well behind interest in the topic. For that reason, empirical guidance on measuring aspects of differentiation (for example, question and response formats) and guidance on combining scale items into composite indexes was limited at the start of this study. Measures of differentiated instruction were developed in Year 1 as part of this study. Our general approach was to rely on instrumentation developed by others in large-scale studies of curriculum and instruction. In particular, the end-of- year survey of teachers included selected items from the *Study of Instructional Improvement: Teacher Questionnaire 2000–2001* (Regents of the University of Michigan, 2001) and Section III of the *Surveys of Enacted Curriculum* (Blank, Porter, and Smithson 2001)[56]; the classroom observations for the full sample of classes obtained three times a year (fall, winter, spring) were based on a modified version of the Center for the Improvement of Early Reading Achievement (CIERA) observation protocol (Taylor et al. 2003); and the teacher logs obtained on a sample of 8 students on 10 occasions across the school year followed Rowan, Camburn, and Correnti (2004). To the extent possible, the wording of items and response formats used in these prior research studies were retained. When modifications were made, they entailed adaptations to enhance the relevance of the topic to the MAP program.

Differentiated instruction is a multifaceted construct. According to the logic model for the MAP program, it is the end product of several procedures. Successful professional development training and consultation should produce an increase in the extent to which teachers use data to determine a student's reading readiness. Teachers are then supposed to use data to group students by common readiness levels or interest and to alter the content, materials, or instructional strategies they use for students within each group.

Three aspects of differentiated instruction—instructional grouping, content coverage, and instructional strategies—were measured using composite indexes derived from each of the three data collection methods.[57] Each composite index—the survey-based composite index, the observation-based composite index, and the log-based composite index—measures the diversity with which teachers group students, cover content, and instruct students.[58] This means that each aspect of differentiated instruction was measured using data from the three data collection methods (see table 3.5).

[56] See appendix F for the observation protocol, appendix G for the log protocol, and appendix H for the teacher survey.

[57] The teacher survey was administered in spring 2009 and spring 2010. The indexes described in this section were developed using the 2009 data from 170 MAP and control teachers in Year 1.

[58] The full definition of differentiated instruction would include measures of data use, assessment methods, and use of materials. Common measures were not available for these three variables. The data from the survey produced scales with marginal reliability. The correlation between composite measures with and without scales for data use and use of materials was 0.91. The loss of the two scales should not result in a composite that underrepresents the differentiated instruction construct.

Table 3.5. Measures and data sources used to assess differentiated instruction, by component

Aspect of differentiation	Scale/measure	Data collection method
Instructional grouping/use of multiple instructional groups	Proportion of ability-grouping activities (includes any type of grouping, grouping frequency, and change in groupings)	Teacher survey
	Average proportion of segments with subgroup instructional modalities	Classroom observations
	Average proportion of logs with multiple differentiation for subgroup instructional modalities	Instructional logs
Content coverage/ diversity of instructional topics	Proportion of literacy topics covered three or more days a week	Teacher survey
	Any type of differentiated instruction in multiple content areas per observation segment	Classroom observations
	Average proportion of log events with differentiation for focal topics	Instructional logs
Instructional strategies/ diversity of instructional strategies	Proportion of instructional strategies used in a week	Teacher survey
	Any use of differentiated instructional strategies by teachers or their students in comprehension or writing	Classroom observations
	Average proportion of log rounds with differentiation for comprehension, writing, and word analysis areas	Instructional logs

Source: Authors' compilation.

Scales and composite indexes based on teacher surveys, classroom observation, and teacher logs

Teacher survey–based composite index. The teacher survey on instructional practices includes a series of questions pertaining to the overall school environment; characteristics of the teacher's reading/English language arts class (amount of time spent on reading/English language arts on a typical day, instructional grouping); and general approaches to instruction. To capture the extent to which teachers differentiate instruction, the study team asked teachers to report separately on their instructional practices for students with high and low reading readiness. The survey included a parallel set of questions about grouping students, content coverage (word analysis, reading comprehension), instructional strategies (activating prior knowledge), and instructional materials (informational text, narrative text without control of vocabulary) used when working with high-achieving and low-achieving students. For both high-achieving and low-achieving students, teachers indicated the frequency with which topics, strategies, and materials were used. Responses to the survey items were frequency categories (never, 1–2 times a week, 3–4 times a week, every day). In constructing scales for the survey-based composite index, the study team collapsed the frequency categories into binary variables (0 = 2 or fewer times a week; 1 = 3–4 times a week or every day).

In measuring the extent to which teachers differentiated their instruction, the study team assumed that using more of the listed topics, strategies, materials, formal data sources, and instructional

groupings represented more differentiation. To account for different numbers of items used for each scale, the study team divided the sum of the binary responses for each teacher by the total number of survey items included for each scale. This created a proportion, ranging from 0 to 1.00. To assess whether MAP teachers differentiated more than control teachers based on each of the three scales, the study team compared the average proportion of items indicated by teachers in each condition. The main comparison between MAP and control teachers used a survey-based composite score, which is an average of the three scale scores, with equal weights given to scale scores.

Use of multiple instructional groups. Teachers were asked five questions about their grouping practices, and their responses were coded as 0 or 1:

- Whether they grouped students according to ability level (1 = yes)
- The size of the groups created for high-achieving students (1 = two- to five-person groups)
- The size of the groups created for low-achieving students (1 = two- to five-person groups)
- The frequency with which they grouped students (1 = at least once a week)
- The frequency with which they regrouped students (1 = at least once a month)

Responses were summed and divided by 5. The scale represents the proportion of ability grouping activities (any grouping, grouping frequency, or change in grouping students) engaged in by teachers. The alpha coefficient was 0.78.

Diversity of instructional topics. To measure the diversity of topics used by teachers for high-achieving and low-achieving students, the study team counted the number of topics that were reportedly used three or more times a week for high- and low-achieving students combined. The binary variables were coded as 1 = 3–4 times a week and every day and 0 = 0–2 times a week.

The seven literacy topics that could serve as a primary focus of instruction by the teacher (word analysis, reading fluency, listening comprehension, reading comprehension, grammar, spelling, written composition) for high- and low-achieving students (combined) made up the diversity of instructional topics scale. Diversity of instructional topics was represented as the proportion of these literacy topics reportedly covered three or more days a week. The alpha coefficient was 0.78.

Diversity of instructional strategies. The approach to measuring the use of different instructional strategies, based on responses to the teacher survey, was similar to that for instructional topics. For each of 19 instructional strategies (for example, basic worksheets, learning centers, interest groups), the use of this strategy 3–4 times a week or more was coded as 1, 0–2 times a week was coded as 0, for high- and low-achieving students as a whole. The proportion of affirmative answers to questions on the use (3 or more times per week) of these 19 instructional strategies made up the diversity of instructional strategies scale. The alpha coefficient was 0.83.

Survey-based composite index. A composite index of MAP instructional practices, as reported in the survey by teachers in both MAP and non-MAP conditions, was derived as an equal-weighted additive combination of the three scales just described. The survey-based composite index was derived using the following equation:

Survey-based composite index = (0.33 * Use of multiple instructional groups +

0.33 * Diversity of instructional topics + 0.33 * Diversity of instructional strategies)

Observation-based composite index. Observations were conducted in the fall, winter, and spring in MAP and control classrooms. Using a modification of the CIERA observation protocol, within each observation the study team recorded data on class characteristics, teacher behavior, and student behavior in 10-minute segments. The number of segments varied across classes and observations but generally include six to nine segments.

The composite index, based on observational data (discussed later in this report), was constructed from three binary variables at the unit of the observation segment. The three variables were the use of instructional groups, the diversity of instructional topics, and the diversity of instructional strategies. The composite percentage was then averaged across the segments for each classroom observation: separately for fall, winter, and spring observation periods.

Use of multiple instructional groups. Up to three subgroup instructional modalities (small groups, pairs, individuals) were recorded during each observation segment. If any of the three modalities was observed in a segment, the segment was coded as 1, otherwise the segment was coded as 0.

Diversity of instructional topics. The classroom observation scheme provides a direct measure of differentiated instructional practices. Within each 10-minute observational segment, observers recorded the content area that was the focus of instruction by the teacher (vocabulary, spelling, fluency, reading comprehension, writing, and speaking or listening). In addition, they recorded whether instructional content, processes, or products were differentiated within each of these areas. There were thus 18 possible types of differentiated instruction: 3 types of differentiation (content, process, and product) for 6 topical areas. The presence of any form of differentiation was summed across the 18 types of differentiated instruction within each 10-minute segment.

Diversity of instructional strategies. Twenty instructional strategies for reading comprehension could be recorded for teachers (10) or students (10) within each 10-minute observation segment. In addition, observers could record up to 18 writing-related instructional strategies for teachers (9) or students (9) in each segment. If any of the listed strategies (for comprehension or writing) were observed within a 10-minute segment, the segment was scored as 1; otherwise the segment was scored as 0. The presence of any of the listed instructional strategies was summed within each 10-minute segment.

Observation-based composite. Because of the generally low base rate for subgrouping, diverse content and differentiation, and use of listed instructional strategies within each 10-minute segment, in constructing the composite for observations, the study team dichotomized the totals for the three variables within a segment. If the sum of these three variables was greater than 0, the variable was recoded as 1, otherwise it was recoded as 0. The sum of these dichotomized variables was then divided by 3. Using the dichotomized version of each variable, the observation-based composite for each segment is:

Observation-based composite (segment) = (Use of multiple instructional groups +

Diversity of instructional topics with differentiation + Diversity of instructional strategies)/3

If teachers used subgroups, addressed more than one topic, and used more than one instructional practice within a 10-minute segment, their composite score would be 1.0 (3/3). If teachers did none of these, their composite score would be 0. The composite score is 0.33 if teachers engaged in any one of the above-mentioned practices and 0.67 if they engaged in two of the three practices. Averaging the segment composite across all segments within the full classroom observation period results in a score for the observation period (fall, winter, or spring):

Observation-level composite =(∑composite segments)/Number of segments per observation.

For the main analysis, the average observation composite across the three observation periods is

Observation-based composite =(∑observation-level composites)/3.

Teacher log-based composite index. All participating teachers used the log instrument originally developed for the Study of Instructional Improvement (Rowan, Camburn, and Correnti 2004).[59] The study team randomly selected from each classroom four students in the top quartile and four in the bottom quartile according to their prior year ISAT performance. Teachers used the logs to describe the reading/English language arts instruction provided to each student on each of 10 selected days throughout the year. The study team referred to the day on which a teacher is assigned to complete logs on up to eight students as a *log round*. By asking teachers to complete a log on each of the four highest and four lowest achievers in their classrooms, the study team attempted to optimize the chance that items across the logs and within a log round would detect differential instruction for different ability groups.

For each of the eight selected students, each teacher described the following:

- The instructional groupings used
- The extent to which nine topic areas (comprehension, writing, word analysis, concepts of print, reading fluency, vocabulary, grammar, spelling, and research strategies) were a focus of instruction
- For comprehension, writing, and word analysis, details on the nature of their instruction on the basis of 10–20 additional items describing specific instructional strategies within the specified topic area
- The strategies they used to assess students in comprehension, writing, and word analysis on the basis of 8–10 questions

For each student, teachers recorded the use (coded 1) or nonuse (coded 0) of a practice. A scale was constructed for each of the three aspects of differentiated instruction by considering instruction to be differentiated if the teacher used a specific practice that reflected MAP-relevant practices (for example, grouping students by interest) *and* applied the practice to at least one, but not all, students. Each scale expresses the degree of differentiation across categories of a specific practice (for example, grouping by achievement, by interest, etc.) relative to the total number of categorical practices enacted by the teacher. The process used to calculate scale scores was the

[59] The log instrument used for this study was originally developed as part of the Consortium for Policy Research in Education (CPRE) Study of Instructional Improvement, a large-scale, longitudinal study investigating the effects of three whole-school school reform programs. Information on this study and the instruments used is available at http://www.sii.soe.umich.edu/.

same for all three scales. Table 3.6 illustrates this process, using hypothetical data for the use of multiple instructional groups scale.

Table 3.6. Hypothetical data matrix for determining index of differentiation from single teacher log round

Student ID	Instructional grouping			
	Achievement	**Interest**	**Cooperative learning**	**Pairs**
01	1	1	0	0
02	1	1	0	0
03	1	1	0	0
04	1	1	0	0
05	1	0	1	0
06	1	0	1	0
07	1	0	1	0
08	1	0	1	0
Strategy enacted by the teacher? (1=yes, 0=no)	1	1	1	0
Strategy enacted for 0<p<1 students? (1=yes; 0=no)	0	1	1	0
Number of strategies differentiated relative to total number of strategies enacted	2/3= .67			

Source: Authors' compilation.

Use of multiple instructional groups. Within a teacher log round, the use of multiple instructional groups construct includes (1) grouping students by achievement, (2) grouping students by interest, (3) establishing cooperative learning groups, and (4) pairing students. The extent to which these four grouping strategies are differentially used across students within a log round is calculated in a three-step process. In step one, we summed the number of strategies that the teacher enacted for at least one student. For instance, in table 3.6, the data shows that, for this particular round, the teacher grouped all eight students by achievement, four of eight students by interest, and four of eight students into cooperative groups. The teacher did not group any of the eight students into pairs. Thus, the total number of enacted strategies equals 3 (1+1+1+0=3). In step two, we summed the strategies that the teacher differentially applied across students. The table above shows that the teacher grouped differentially in two of the four categories (interest and cooperative learning), making the total number of differentiated strategies equal to two (0+1+1+0=2). Finally, in step three, we divided the number of differentiated strategies by the total number of strategies enacted to obtain a value of 2/3=.67. These proportions were calculated for each log round and then averaged across rounds (8–10 logs per class per teacher) to obtain the use of multiple instructional groups scale for each individual teacher.

We followed this same process when we calculated the diversity of instructional topics scale and the diversity of instructional strategies scale. That is, the process would mirror that illustrated in table 3.6, with the exception that the columns would refer to different questions asked in the teacher logs, namely, those that target the topics covered by the teacher and those that focus on

the various instructional strategies that the teacher used for a specific student on a specific day. The items used to calculate these two scales are described below.

Diversity of instructional topics. As was previously described for how use of multiple instructional groups was constructed, each teacher was to complete a log for each of four high-achieving students and four low-achieving students on a specific day. The teacher logs allow teachers to record up to nine focal topics for each student (comprehension, writing, word analysis, reading fluency, vocabulary, grammar, spelling, concepts of print, and research strategies). For each of the nine topics, the study team coded the response as 1 if the teacher reported that the topic was a major focus of the day's lesson and 0 if the teacher reported that it was only "touched on" or was not a topic addressed in this day's lesson.

Diversity of instructional strategies. If teachers recorded a focus on comprehension, writing, or word analysis for each of the four high-achieving and four low-achieving students, they were asked to report on their use of 21 specific strategies or areas of comprehension (for example, making predictions, self-monitoring for meaning); 10 areas of writing (for example, organizing ideas for writing, editing, revision); and 14 areas of word analysis (for example, sound segmenting, sound blending, word recognition). It should be noted that the teacher log did not include questions on the instructional strategies used by the teacher for the topics concepts of print, reading fluency, vocabulary, grammar, spelling, and research strategies. The rationale for this was two-fold: (1) the teacher log instrument was that developed by Rowan and his colleagues, and these constructs were not the focus of their study; and (2) we believed that it was wise not to develop additional questions on instructional strategies for these topics so as to impose undue burden on our teacher respondents. The use index ($p > 0$) and the index of differential use ($0 < p < 1$) was calculated for each of the three areas separately, averaged within a log, and averaged across log rounds.

Log-based composite index. Each of the three scales just described produced values ranging from 0 to 1. The resulting differentiated instruction composite for data weights each of the scale values equally (0.33):

Log-based composite index = (0.33 * Use of multiple instructional groups +

0.33 * Diversity of instructional topics + 0.33 * Diversity of instructional strategies).

Results

The study team examined three aspects of differentiated instruction—use of multiple instructional groups, the diversity of instructional topics use by teachers, and the use of multiple instructional strategies—using an end-of-the-year teacher survey, classroom observations, and teacher logs. These assessment protocols generated three overall composite indexes (one for each protocol) and nine scales (three for each aspect of differentiated instruction). The results are presented separately below for grade 4 (table 3.7) and grade 5 (table 3.8).[60]

[60] The statistical results for the composites and the scales that compose them are based on data from all 172 teachers included in the intent-to-treat analyses of grades 4 and 5. To account for missing data, researchers generated five imputed datasets (see appendix D). The results in tables 3.7 and 3.8 are based on the average estimates across the five datasets.

Table 3.7. Achieved Relative Strength Index (ARSI) for differentiation composites for grade 4 teachers

Data source	Composite and scales	Group	Mean	BTX	SEB	*t*-test	Standard deviation	ARSI
Observation	Overall composite index	MAP	0.405	0.021	0.038	0.54	0.161	0.128
		Control	0.384					
	Use of multiple instructional groups scale	MAP	0.395	0.035	0.049	0.71	0.210	0.164
		Control	0.360					
	Diversity of instructional topics with differentiated instruction scale	MAP	0.218	–0.00	0.047	–0.00	0.195	–0.008
		Control	0.219					
	Diversity of instructional strategies scale	MAP	0.601	0.029	0.044	0.67	0.191	0.152
		Control	0.572					
Survey	Overall composite index	MAP	0.548	0.060	0.039	1.52	0.177	0.335
		Control	0.488					
	Use of multiple instructional groups scale	MAP	0.794	0.004	0.049	0.09	0.201	0.022
		Control	0.789					
	Diversity of instructional topics scale	MAP	0.573	0.110	0.076	1.45	0.343	0.318
		Control	0.463					
	Diversity of instructional strategies scale	MAP	0.276	0.065	0.045	1.45	0.197	0.327
		Control	0.210					
Teacher log	Overall composite index	MAP	0.350	0.046	0.046	1.00	0.206	0.220
		Control	0.304					
	Use of multiple instructional groups scale	MAP	0.384	0.079	0.063	1.26	0.279	0.280
		Control	0.305					
	Diversity of instructional topics scale	MAP	0.223	0.016	0.049	0.34	0.221	0.070
		Control	0.207					
	Diversity of instructional strategies scale	MAP	0.443	0.042	0.052	0.81	0.234	0.180
		Control	0.401					

Note: B_{TX} equals the difference between the MAP and control mean values. SE_B equals the standard error of the difference between the MAP and control mean scores.

Source: Authors' analysis based on Year 2 (2009–10) data from the study districts and the Northwest Evaluation Association.

Table 3.8. Achieved Relative Strength Index (ARSI) for differentiation composites for grade 5 teachers

Data source	Composite and scales	Group	Mean	BTX	SEB	*t*-test	Standard deviation	ARSI
Observation	Overall composite index	MAP	0.413	0.010	0.039	0.26	0.171	0.059
		Control	0.402					
	Use of multiple instructional groups scale	MAP	0.432	0.048	0.050	0.98	0.223	0.216
		Control	0.383					
	Diversity of instructional topics with differentiated instruction scale	MAP	0.234	0.024	0.048	0.50	0.214	0.110
		Control	0.210					
	Diversity of instructional strategies scale	MAP	0.572	–0.042	0.043	–0.97	0.189	–0.219
		Control	0.613					
Survey	Overall composite index	MAP	0.591	0.165	0.037	4.51***	0.183	0.894
		Control	0.426					
	Use of multiple instructional groups scale	MAP	0.777	0.087	0.049	1.79	0.218	0.396
		Control	0.690					
	Diversity of instructional topics scale	MAP	0.695	0.283	0.070	4.04***	0.327	0.856
		Control	0.413					
	Diversity of instructional strategies scale	MAP	0.299	0.125	0.045	2.79**	0.200	0.620
		Control	0.174					
Teacher log	Overall composite index	MAP	0.327	0.023	0.046	0.50	0.339	0.067
		Control	0.304					
	Use of multiple instructional groups scale	MAP	0.385	0.068	0.057	1.19	0.262	0.256
		Control	0.318					
	Diversity of instructional topics scale	MAP	0.217	0.007	0.043	0.16	0.191	0.035
		Control	0.210					
	Diversity of instructional strategies scale	MAP	0.381	–0.005	0.582	–0.19	0.256	–0.020
		Control	0.386					

** Difference statistically significantly different from zero at the .01 level.

*** Difference statistically significantly different from zero at the .001 level.

Note: B_{TX} equals the difference between the MAP and control mean values. SE_B equals the standard error of the difference between the MAP and control mean scores.

Source: Authors' analysis based on Year 2 (2009–10) data from the study districts and the Northwest Evaluation Association.

The overall composite scores for each method represent an equal weighting of the three scales. Each scale was constructed to represent the proportions of activities used by teachers (logs and survey) or the proportion of observation segments in which an activity or strategy was observed. Higher proportions are intended to reflect higher levels of differentiated instruction.

For grade 4 teachers, the average observation composite was 0.405 for MAP teachers and 0.384 for control teachers; the difference (B_{TX} = 0.021) was not statistically significant (t = 0.54). The ARSI (0.128) suggests very little difference between groups. The survey composite results reveal a larger difference between groups (ARSI = 0.335), but the mean difference (B_{TX} = 0.060) was not statistically significant (t = 1.52). The means were 0.548 for the MAP group and 0.488 for the control group. The log composite reveals no difference between the two conditions (B_{TX} = 0.046, t = 1.00, with means of 0.350 for MAP teachers and 0.304 for control teachers). The average ARSI across the three composites was 0.227.

For grade 5 teachers, the average observation composite was 0.413 for MAP teachers and 0.402 for control teachers; the difference (B_{TX} = 0.010) was not statistically significant (t = 0.26). The ARSI (0.059) suggests very little difference between groups. The survey composite results for grade 5 teachers reveal a relatively large difference between groups (ARSI = 0.894); the mean difference (B_{TX} = 0.165) was statistically significant (t = 4.51, $p < .001$, with means of 0.591 for MAP teachers and 0.426 for control teachers). The log composite did not reveal a difference between the two conditions (B_{TX} = 0.023, t = 1.19, with means of 0.327 for MAP teachers and 0.304 for control teachers). The average ARSI across the three composites was 0.340.

The composite scores for each method represent the average of the three indicators. The data in table 3.7 show no significant differences for any of the nine indicators: the ARSI values for the nine indicators vary around their composite ARSI values, with none greater than the largest ARSI composite value (0.335 for the survey composite). The ARSI results for specific indicators are consistent with the averages reported for the overall composites (see table 3.8). The variability that does exist (the observation-based results range from –0.219 to 0.216 and the log composite values range from –0.020 to 0.256) reflects chance-based fluctuations around 0. It is unlikely that the groups differed in important ways that are not reflected in the average composite values for each method.

Differences across districts. ARSI values varied across districts, ranging from –0.694 for District 5 to 1.188 for District 2 on the observation composite (table 3.9). They also varied within most districts across methods of data collection (for example, –0.554 for the log composite and 1.097 for the survey composite in District 3, grade 5). In just two districts were the ARSI values consistently positive across the three methods (averaging 0.869 for District 2/grade 4 and 0.876 for District 4/grade 5). The study team suspect that the ARSI values are too imprecise (as a result of small sample sizes within districts) to be meaningfully interpreted at the district and grade levels.

Table 3.9. Mean differentiated instruction composites and Achieved Relative Strength Index (ARSI) values in grades 4 and 5, by district

District	Composite	Group	Grade 4			Grade 5		
			n	Mean	ARSI	*n*	Mean	ARSI
1	Observation	MAP	20	0.337	0.047	20	0.324	–0.194
		Control	19	0.328		21	0.354	
	Survey	MAP	20	0.551	0.596	20	0.564	0.834
		Control	19	0.436		21	0.429	
	Logs	MAP	20	0.397	0.378	20	0.372	0.106
		Control	19	0.305		21	0.345	
2	Observation	MAP	3	0.377	1.188	4	0.313	0.190
		Control	4	0.259		3	0.281	
	Survey	MAP	3	0.559	0.605	4	0.552	0.063
		Control	4	0.521		3	0.538	
	Logs	MAP	3	0.304	0.814	4	0.187	–0.031
		Control	4	0.181		3	0.193	
3	Observation	MAP	6	0.565	0.380	5	0.628	–0.271
		Control	5	0.512		5	0.654	
	Survey	MAP	6	0.462	–0.846	5	0.564	1.097
		Control	5	0.580		5	0.388	
	Logs	MAP	6	0.332	0.462	5	0.312	–0.554
		Control	5	0.245		5	0.392	
4	Observation	MAP	9	0.435	–0.293	4	0.467	0.301
		Control	4	0.507		9	0.422	
	Survey	MAP	9	0.635	0.353	4	0.661	1.461
		Control	4	0.577		9	0.396	
	Logs	MAP	9	0.323	0.301	4	0.282	0.865
		Control	4	0.267		9	0.164	
5	Observation	MAP	12	0.422	–0.694	4	0.627	1.897
		Control	3	0.530		12	0.397	
	Survey	MAP	12	0.516	0.099	4	0.725	1.524
		Control	3	0.494		12	0.430	
	Logs	MAP	12	0.313	–1.890	4	0.311	–0.108
		Control	3	0.615		12	0.329	

Source: Authors' analysis based on Year 2 (2009–10) data from the study districts and the Northwest Evaluation Association.

Potential role of contamination. The ARSI does not address the possibility that treatment and control groups may be similar because of infidelity in the treatment condition or upgrading of the control condition (that is, contamination). In response to this issue, the study team assessed the extent to which implementation of MAP-like features in the control condition may have reflected contamination or preexisting instructional strategies of individual teachers (see appendix L). Based on a time-by-group ANOVA of classroom observation data (the only dataset with a true pretest measure of teacher behavior), the study team concluded that preexisting teacher dispositions were responsible for the presence of MAP-like features in the control condition. By extension, the presence of differentiated instructional practices in the control condition in Year 2 reflect preexisting individual differences in teacher practices, not contamination. This interpretation is consistent with the generally accepted notion of treatment diffusion (Shadish, Cook, and Campbell 2002).

Summary of results on implementation

Implementation of the MAP program

Implementation by NWEA. NWEA provided the resources needed to support the MAP program at the school and classroom levels. Throughout the study period, testing resources were fully available in all schools, web-based resources were continuously available, and MAP training and testing were scheduled and conducted in a timely fashion. During both years of the intervention NWEA trainers were available for follow-up consultations. Implementation of the MAP program unfolded without any notable problems.

As part of the plan, during Year 2, the presence of the NWEA staff in the schools was reduced. Fewer new teachers received MAP training (5 percent in Year 2 versus 99 percent in Year 1), and fewer MAP teachers received consultation services (62 percent in Year 2 versus 90 percent in Year 1).

Implementation by MAP teachers. The study team identified 12 MAP-relevant components that teachers could implement during the two-year period of this study. Only the MAP teachers included in the intent-to-treatment analyses were included in the program-specific implementation analysis for teachers.

The same implementation profile was observed for grade 4 and grade 5 MAP teachers. Participation rates varied across the 12 program components, ranging from 36 percent (use of MAP web-based resources) to 90 percent (use of MAP resources for planning lessons).

There was considerable variation in the dose level across teachers (ranging from 0 to 1). The average dose of MAP program components was .66 in both grades. About half of teachers participated at rates of .75 and higher. The dose data suggest that there was substantial variability in the extent to which MAP teachers implemented the program.

Use of differentiated instruction

Data from classroom observations and teacher logs show small, nonsignificant differences in the use of key aspects of differentiated instruction as measured in the current study. Teacher reports of differentiation in grade 5, however, reveal differences between conditions. The grade 5

differences were statistically significant for the survey composite measure ($p < 0.001$) and the ARSI was relatively large (0.894). The survey composite for grade 4 was not significant at $p < 0.05$, and the ARSI was modest (0.335). The best estimate of the ARSI for differences between conditions across the three measures was 0.227 for grade 4 and 0.340 for grade 5. By conventional standards for interpreting effect sizes, these estimates reflect small differences.

Although the MAP program was implemented with moderate fidelity, it did not translate into sizable differences in teacher practices. Only the survey-based results from teachers in grade 5 indicated a difference. The analyses conducted for this study indicate that the overall lack of difference between conditions on differentiation variables probably reflects the operation of two processes. First, MAP program teachers were variable in their implementation of the MAP training and use of resources. On average, teachers in both grades implemented about two-thirds of the twelve MAP activities in the two-year period (see table 3.3). Second, MAP was not the only resource available to all teachers, and the levels of differentiation observed in the control condition are likely to be the result of other forms of professional development and, more generally, preexisting instructional dispositions of teachers in both conditions, but not contamination of the control condition (see appendix L).

Chapter 4: Impacts on Grade 4 Student Achievement

This chapter presents the results of the intent-to-treat analysis of grade 4 student achievement in Year 2 of MAP implementation. It draws measures of achievement from students' test scores on the spring 2010 ISAT reading assessment and composite (average) scores on the spring 2010 MAP assessments in reading and language usage.

This chapter presents the evidence on the study's main confirmatory question:

- Did the MAP program (that is, training plus formative testing feedback) affect the reading achievement of grade 4 students after Year 2 of implementation, as measured by the Illinois Standards Achievement Test (ISAT) reading scale scores or the MAP composite test scores in reading and language use?

Confirmatory impact findings

The MAP program had no statistically significant overall impact on the reading achievement of grade 4 students as measured by the ISAT reading scale score or the MAP composite (reading and literacy) scores (table 4.1). The directions (positive) and magnitudes of the impacts were similar for the two outcomes: a 0.05 standard deviation for the ISAT reading score and a 0.07 standard deviation for the composite MAP score.

Table 4.1. Overall impact of Measures of Academic Progress (MAP) on grade 4 student achievement outcomes in Year 2

	Mean		Estimated impact			
Outcome	**MAP mean**	**Control mean**	**Impact**	**Standard error**	***p*-value**	**Effect size**
Illinois Standards Achievement Test (ISAT) reading scale score	215.6	214.3	1.29	1.570	.412	0.05
MAP composite score	202.5	201.5	0.96	0.891	.280	0.07
Sample size	1,149	765				

Note: Means and impacts were regression adjusted to account for clustering of students within schools, district effects, and baseline student, teacher, and school characteristics and weighted by the number of schools in each district. Effect sizes were computed by dividing the impact by the standard deviation of the outcome for the control group. *p*-values are from a two-tailed test of the null hypothesis of equality of MAP and control means. None of the estimated impacts was statistically significant at the .05 level. Because no test was found significant, it was not necessary to adjust for multiple testing. Sample sizes include all eligible students from the 32 participating schools. Missing outcome and covariate data were estimated using multiple model-based imputation.

Source: Authors' analysis based on Year 2 (2009–10) data from the study districts and the Northwest Evaluation Association.

These findings were obtained using two-level hierarchical regression models to adjust for the clustering of students within schools and district fixed effects to control for the randomization of schools within districts. The models also incorporated baseline student characteristics (prior reading achievement, gender, socioeconomic status, racial/ethnic minority status, English proficiency status, and disability status); teacher characteristics (gender, graduate degree status, teaching experience in English language arts, licensure status, racial/ethnic minority status); and school mean prior reading achievement on the ISAT.[61] The overall impacts presented in table 4.1 are averages of district-specific impacts obtained from the regression models, weighted by the number of schools in each district.[62] Analyses were conducted on the complete sample of 1,914 eligible grade 4 students (and 85 grade 4 teachers) from the 32 participating schools (including the grade 4 control school that withdrew from the study immediately after randomization[63]), using multiple imputation to fill in missing outcome and covariate values.[64]

[61] Similar results were found using models with other covariate specifications (see appendix B).

[62] Tables B.2 and B.3, in appendix B, give the district-specific impact estimates for grade 4. The study is not sufficiently powered to detect impacts at the district level; district-specific estimates must therefore be interpreted with caution.

[63] Analogous analyses that excluded the school that withdrew, as well as analyses that included only students with complete outcomes, were conducted as part of the sensitivity analyses (see appendix D). The conclusions presented here proved robust across different samples.

[64] Appendix D describes the imputation methods used and presents the missing rates on analysis variables. Table D.2 gives the proportions of missing grade 4 data.

Chapter 5: Impacts on Grade 5 Student Achievement

This chapter describes the results of the intent-to-treat analysis of grade 5 student achievement in Year 2 of MAP implementation. It provides evidence on the following exploratory research question:

- Did the MAP program affect the reading achievement of grade 5 students after Year 2 of implementation, as measured by the Illinois Standards Achievement Test (ISAT) reading scale scores or the MAP composite test scores in reading and language use?

Effect on reading achievement

The MAP program had no statistically significant impact on the reading achievement of grade 5 students as measured by the ISAT or the composite scores on the MAP reading and language use assessments (table 5.1). The magnitudes of the (nonsignificant) impacts were similar, but the directions were opposite: a negative effect size of 0.05 standard deviation for the ISAT reading score and a positive effect size of 0.01 standard deviations for the composite MAP score.

Table 5.1. Impacts on grade 5 student achievement outcomes in Year 2

	Mean		Estimated impact			
Outcome	**MAP**	**Control**	**Impact**	**Standard error**	***p*-value**	**Effect size**
Illinois Standards Achievement Test (ISAT) reading scale score	221.7	223.2	–1.48	1.366	.280	–0.05
MAP composite score	205.6	205.4	0.15	1.037	.889	0.01
Sample size	701	1,105				

Note: Means and impacts were regression adjusted to account for clustering of students within schools, district effects, and baseline student, teacher, and school characteristics and weighted by the number of schools in each district. Effect sizes were computed by dividing the impact by the standard deviation of the outcome for the control group. *p*-values are from a two-tailed test of the null hypothesis of equality of MAP and control means. None of the estimated impacts is statistically significant at the .05 level. Because no test was found significant, it was not necessary to adjust for multiple testing. Sample sizes include all eligible students from the 32 participating schools. Missing outcome and covariate data were estimated using multiple model-based imputation.

Source: Authors' analysis based on Year 2 (2009–10) data from the study districts and the Northwest Evaluation Association.

As in the core analysis for grade 4, the findings shown in table 5.1 were obtained using two-level hierarchical regression models to adjust for the clustering of students within schools, and district fixed effects to control for the randomization of schools within districts. The models also accounted for baseline student characteristics (prior reading achievement, gender, socioeconomic status, racial/ethnic minority status, English proficiency status, and disability status); teacher characteristics (gender, graduate degree status, teaching experience in English language arts, licensure status, racial/ethnic minority status); and school mean prior reading achievement on the ISAT.[65] The overall impacts presented in table 5.1 are weighted averages of district-specific impact estimates from the regression models, where the weights are the number of schools in each district.[66] The analyses included the complete sample of 1,806 eligible grade 5 students (and 87 grade 5 teachers) from the 32 participating schools (including the grade 5 MAP school that withdrew from the study immediately after randomization[67]), using multiple imputation to estimate missing outcome and covariate values.[68]

[65] Parallel to the grade 4 analysis, researchers also explored models with other covariate specifications (appendix B). The results were robust to the selection of covariates.

[66] Appendix B shows the district-specific impact estimates for grade 5. The small number of schools in four of the five study districts suggests that these estimates must be interpreted cautiously.

[67] Analogous analyses that excluded the school that withdrew, as well as analyses that included only students with complete outcomes, were conducted as part of the sensitivity analyses. The conclusions presented here proved robust across different samples used in the analyses.

[68] Appendix D describes the imputation methods used and presents the missing rates on analysis variables. Table D.3 gives the proportions of missing grade 5 data.

Appendix A. School and Student Characteristics

This appendix supplements the information on school and student characteristics presented in chapter 2.

Table A.1. Characteristics of study districts, 2008–09

Characteristic	District				
	1	**2[a]**	**3**	**4**	**5**
Number of schools	20	–	4	3	3
Socioeconomic status					
Percentage of Title I schools in district	100.0	–	75.0	100.0	33.3
Percentage of students in district eligible for free or reduced-price lunch	75.1	–	19.8	21.4	0.0
Race/ethnicity (percentage of students in district)					
White	50.1	–	93.5	67.0	72.8
Black	37.1	–	1.8	0.6	5.6
Hispanic	1.7	–	2.3	23.6	14.4
Other	11.0	–	2.4	8.9	7.2
Enrollment and number of teachers					
Total district enrollment	6,074	–	1,622	1,931	1,799
Total number of full-time teachers in each district	389	–	72	103	102

Note: This table includes only the study schools in each of the five participating districts.

a. The characteristics of District 2 have been suppressed to prevent a disclosure risk..

Source: Authors' analysis based on data from National Center for Education Statistics Common Core of Data 2008/09.

Table A.2. Characteristics of schools in study and eligible schools in Illinois, the Midwest, and the United States, 2008/09

Characteristic	Study schools	Eligible schools in Illinois[a]	Eligible schools in Midwest[b]	Eligible schools in United States[c]	All schools in United States[d]
Number of schools	32	1,960	7,869	37,646	43,498
Socioeconomic status					
Percentage of Title I schools	90.6	81.3 (*n* = 1,878)	84.7 (*n* = 7,787)	76.6 (*n* = 37,561)	76.5 (*n* = 43,070)
Average percentage of students eligible for free or reduced-price-lunch students	54.7	49.2 (*n* = 1,907)	45.9 (*n* = 7,816)	50.5 (*n* = 37,213)	50.5 (*n* = 43,830)
Race/ethnicity and gender (average percentage of students)					
White	61.4	50.4	67.5	52.6	54.3
Black	24.8	23.1	16.9	17.4	17.3
Hispanic	5.0	19.5	9.5	21.9	20.4
Other	8.8	7.0	6.2	8.1	8.0
Male	44.8	47.0	49.6	50.8	50.7
Enrollment and number of teachers					
Average total school enrollment	379	464	416	490	478
Average number of students in grade 4	64	69	66	77	74
Average number of students in grade 5	65	69	66	77	74
Average number of full-time teachers	27	28 (*n* = 1,958)	25	31 (*n* = 37,622)	30 (*n* = 43,241)
School setting (percentage of schools)					
City	50.0	35.7	29.3	31.4	31.2
Suburb	31.3	38.6	29.6	32.2	30.5
Town	3.1	7.2	12.2	10.2	10.5
Rural	15.6	18.5	28.9	26.2	27.7

Note: Averages are unweighted means across schools. Where data are missing on some schools, *n* is the actual number of schools used for calculating the average characteristic across schools.

a. Schools located in Illinois that had at least 10 students in grade 4 and at least 10 students in grade 5, were noncharter schools, were defined as "regular" schools by the Common Core of Data, and were operational at the time of the Common Core of Data report.

b. Schools that met the same eligibility criteria but were located in the seven states served by the REL Midwest (Illinois, Indiana, Iowa, Michigan, Minnesota, Ohio, and Wisconsin).

c. Schools that met the same eligibility criteria but were located in the 50 states and the District of Columbia.

d. All schools in the 50 states and the District of Columbia that had at least 10 students in grade 4 and at least 10 students in grade 5 during 2007/08, were defined as regular schools by the Common Core of Data, and were operational at the time of the Common Core of Data report.

Source: Authors' analysis based on data from the National Center for Education Statistics Common Core of Data 2008/09.

Table A.3. Characteristics of study schools, 2008/09

Characteristic	Mean		Estimated difference	*p*-value
	Grade 4 MAP, grade 5 control schools	Grade 5 MAP, grade 4 control schools		
Number of schools	16	16		
Title I and school composition				
Percentage of Title I schools	89.1	90.6	–1.6	.833
Average percentage of students eligible for free or reduced-price lunch	49.3	59.6	–10.3	.039*
Average percentage of White students	62.3	60.3	2.0	.695
Average percentage of male students	44.8	44.9	–0.2	.867
Enrollment and number of teachers				
Average total school enrollment	438	323	115	.004*
Average number of students in grade 4	73	53	20	.022*
Average number of students in grade 5	75	53	22	.007*
Average number of full-time teachers	30	25	5	.059
School locale (percentage of schools)[a]				
City	56.2	43.8	12.5	.480
Suburb	25.0	37.5	–12.5	.446
Town	0	6.2	–6.2	.310
Rural	18.8	12.5	6.3	.626
Illinois Standards Achievement Test (ISAT) 2009 reading scale score (mean)				
Grade 3 students[b]	202.0	203.2	–1.3	.677
Grade 4 students[b]	214.2	214.7	–0.5	.852
Joint test of difference in school characteristics between MAP and control groups[b] (χ^2 = 10.3, df = 9)				.323

*Difference statistically significantly different from zero at the .05 level.

Note: Means and differences were regression adjusted using ordinary least squares to account for district effects and weighted by the number of schools in each district. *n* represents the actual number of schools used to calculate the average characteristic across schools. *p*-values are from a two-tailed test of the null hypothesis of equality of MAP and control means.

a. Chi-squared test of homogeneity of distributions was not statistically significant (χ2 = 1.85, *p*-value = .604).

b. An overall test of the difference between the MAP and control groups based on the school characteristics in this table was conducted using a chi-square test. The chi-square test is from a logistic regression model with the binary treatment indicator as outcome and the school characteristics as covariates (school locale was included in the model as the combined percentage of city and suburb, because no schools located in towns were included in the grade 4 MAP/grade 5 control sample).

Source: Authors' analysis based on data from the National Center for Education Statistics Common Core of Data 2008/09.

Table A.4. Characteristics of grade 5 teachers, 2008/09 (before Year 2 implementation)

Characteristic	MAP (n = 37)	Control (n = 50)	Estimated difference	p-value
Percentage female	92.8	85.3	7.5	.263
Percentage with graduate degree	63.1	77.9	–14.8	.147
Years teaching experience in English language arts	10.1	10.8	–.7	.708
Percentage with permanent license	80.5	91.3	–10.8	.188
Percentage White	93.8	87.1	6.7	.295
Joint test of difference in student characteristics between MAP and control groups[a] (χ^2= 2.74, df = 5)				.740

Note: Means and differences were regression adjusted to account for district effects and weighted by the number of schools in each district. Where data are missing, *p*-values are from a two-tailed test of the null hypothesis of equality of MAP and control means. None of the estimated differences is statistically significant at the .05 level.

a. An overall test of the difference between the MAP and control groups based on the teacher characteristics in this table was conducted using a chi-square test. The chi-square test is from an ordinary logistic regression model with the binary treatment indicator as outcome and the teacher characteristics in this table as covariates.

Source: Authors' analysis based on Year 2 student baseline data collected from study districts in spring 2009, when students were in grade 3.

Table A.5. Characteristics of grade 5 students, 2008/09 (before Year 2 implementation)

Characteristic	MAP (*n* = 701)	Control (*n* = 1,105)	Estimated difference	*p*-value
Percentage eligible for free or reduced-price lunch	57.9	54.6 (*n* =1,104)	3.30	.525
Percentage White	59.5	57.9	1.6	.803
Percentage with disability	15.8 (*n* = 700)	16.4 (*n* = 1,104)	–0.60	.836
Percentage English proficient	97.2 (*n* = 697)	96.5 (*n* = 1,092)	–0.80	.539
Percentage male	50.9	52.9	–2.00	.442
Illinois Standards Achievement Test (ISAT) 2009 reading scale score	214.2 (*n* = 659)	214.7 (*n* = 1,027)	–0.50	.852
Joint test of difference in student characteristics between MAP and control groups[a] (*F* = 0.07, df = (11, 27))				1.000

Note: Means and differences were regression adjusted to account for district effects and clustering of students within schools and weighted by the number of schools in each district. Where data are missing, *n* is the actual number of students used to calculate the average characteristic in each treatment group. *p*-values are from a two-tailed test of the null hypothesis of equality of MAP and control means. None of the estimated differences was statistically significant at the .05 level.

a. An overall test of the difference between the MAP and control groups based on all student characteristics in this table was conducted using an *F*-test adjusted for the randomization of blocks within districts, and the clustering of students within schools. The *F*-test is from a two-level logistic regression model with the binary treatment indicator as outcome and the student characteristics in this table as covariates.

Source: Authors' analysis based on Year 2 student baseline data collected from study districts in spring 2009, when students were in grade 3.

Table A.6. Grade 5 attrition rates on posttest scores

	Illinois Standards Achievement Test (ISAT) 2010			Measures of Academic Progress (MAP) spring 2010		
Item	**Overall**	**MAP**	**Control**	**Overall**	**MAP**	**Control**
Observed	1,719	671	1,048	1,669	612	1,057
Missing	87	30	57	137	89	48
Total number of students	1,806	701	1,105	1,806	701	1,105
Attrition rate (percent)	4.8	4.3	5.2	7.6	12.7	4.3
Chi-square test of equality of proportions	$\chi^2 = 0.72$, df = 1, p-value = .395			$\chi^2 = 42.68$, df = 1, p-value < .0001*		

* Difference statistically significantly different from zero at the .05 level.

Source: Authors' analysis based on data from the study districts.

Table A.7. Illinois Standards Achievement Test (ISAT) pretest scores of grade 5 students with missing ISAT and Measures of Academic Progress (MAP) scores

	Missing ISAT 2010 scores			Missing MAP spring 2010 scores		
Characteristic	**MAP**	**Control**	**Difference (p-value)**	**MAP**	**Control**	**Difference (p-value)**
Number of students	30	57		89	48	
Mean ISAT 2009 reading scale score[a]	219.4 (n = 21)	208.3 (n = 31)	11.1 (.175)	208.7 (n = 78)	207.6 (n = 36)	1.1 (.840)

Note: n includes only students with nonmissing ISAT 2009 scores. A two-tailed *t*-test for equality of means was used.

a. Scores are pretest means: the grade 4 average score on the 2009 ISAT assessment that was administered in the spring before the Year 2 implementation, when grade 5 students in study Year 2 were in grade 4.

Source: Authors' analysis based on data from the study districts. .

Table A.8. Illinois Standards Achievement Test (ISAT) pretest scores of Year 2 grade 5 "dropouts" and "stayers"

Characteristic	ISAT 2010 scores			MAP spring 2010 scores		
	Dropouts	Stayers	Difference (*p*-value)	Dropouts	Stayers	Difference (*p*-value)
Number of students	87	1,719		137	1,669	
Mean ISAT 2009 reading scale score[a]	213.0 (*n* = 52)	214.5 (*n* = 1,634)	–1.5 (.676)	210.0 (*n* = 114)	214.8 (*n* = 1,572)	–4.8 (.100)

Note: Means are weighted by the number of schools in each district. *n* includes only students with nonmissing ISAT 2009 scores. A two-tailed *t*-test for equality of means was used.

a. Scores are pretest means: the grade 4 average score on the 2009 ISAT assessment that was administered in the spring before the Year 2 implementation, when grade 5 students in study Year 2 were in grade 4.

Source: Authors' analysis based on data from the study districts.

Table A.9. Correlations between pretest scores and Year 2 outcome measures for Year 2 grade 4 students

Outcome measure	Measures of Academic Progress (MAP) 2010 reading score	MAP 2010 language use score	MAP 2010 composite score	ISAT 2010 reading scale score
MAP 2010 language use score	0.86			
MAP 2010 composite score	0.97	0.96		
Illinois Standards Achievement Test (ISAT) 2010 reading scale score	0.81	0.78	0.82	
Illinois Standards Achievement Test (ISAT) 2009 (pretest) reading scale score	0.80	0.80	0.83	0.82

Source: Authors' analysis based on data from the study districts.

Table A.10. Correlations between pretest scores and Year 2 outcome measures for Year 2 grade 5 students

Outcome measure	Measures of Academic Progress (MAP) 2010 reading score	MAP 2010 language use score	Map 2010 composite score	ISAT 2010 reading scale score
MAP 2010 language use score	0.81			
MAP 2010 composite score	0.96	0.95		
Illinois Standards Achievement Test (ISAT) 2010 reading scale score	0.78	0.77	0.81	
Illinois Standards Achievement Test (ISAT) 2009 reading scale (pretest) score	0.77	0.78	0.81	0.81

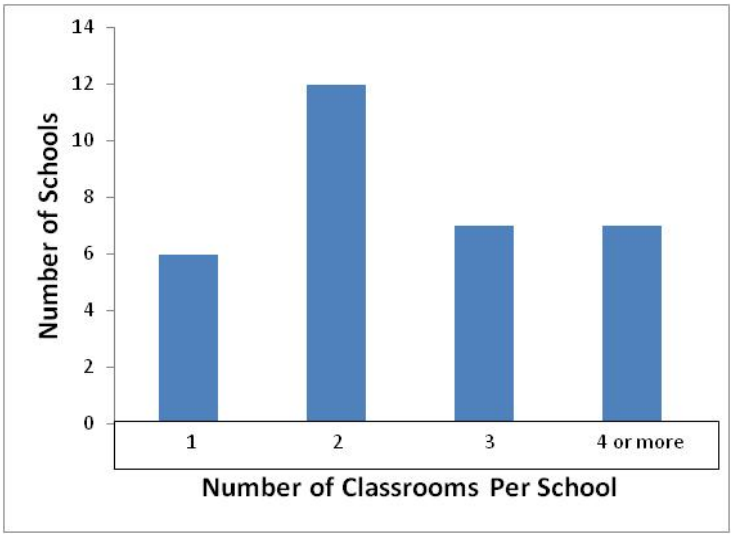
Source: Authors' analysis based on data from the study districts.

Table A.11. Scale score ranges of student performance levels on the 2009 Illinois Standards Achievement Test (ISAT) in reading

Grade	Academic warning (W)	Below standards (B)	Meets standards (M)	Exceeds standards (E)
3	120–155	156–190	191–226	227–329
4	120–157	158–202	203–236	237–341

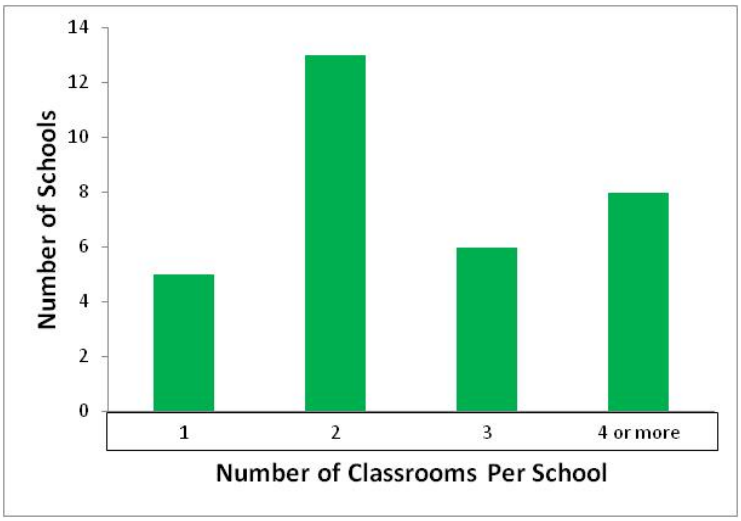
Source: Illinois State Board of Education (http://www.isbe.state.il.us/assessment/htmls/isat_general_info.htm#cut).

Figure A.1. Frequency distribution of number of grade 4 classrooms per school in Year 2 of implementation (2009/10)



Source: Authors' analysis based on data from the study districts.

Figure A.2. Frequency distribution of number of grade 5 classrooms per school in Year 2 of implementation (2009/10)



Source: Authors' analysis based on data from the study districts.

Appendix B. Impact Estimation and Impact Estimates

This appendix presents the model used to estimate the impact of MAP on student achievement in grades 4 and 5. It also provides estimates of the impact.

Model for estimating impact

The intent-to-treat impacts of the MAP intervention on student achievement were estimated (separately for grades 4 and 5) using two-level hierarchical linear models with students nested within schools. The composite model is given by

$$Y_{ij} = \sum_k \gamma_{00k} D_k + \sum_k \gamma_{01k} Trt_j D_k + \gamma_{02} MEANISAT09_j + \gamma_{10} ISATREADSCALE09_{ij} + \mathbf{S_{ij}\Gamma_{20}} + \mathbf{T_{ij}\Gamma_{30}} + r_{0j} + \varepsilon_{ij}, \quad \text{(B.1)}$$

where

Y_{ij} = achievement of student *i* in school *j*, as measured by either the ISAT 2010 reading scale score or the MAP 2010 composite reading and language usage scale score

D_k = 1 if school *j* is in district *k* and 0 otherwise, $k = 1, \cdots, 5$

Trt_j = 1 if school *j* is in the treatment group and 0 otherwise

$MEANISAT09_j$ = ISAT 2009 mean pretest score in reading (for grade 4 or grade 5) for school *j* (centered on its grand mean)

$ISATREADSCALE09_{ij}$ = ISAT 2009 pretest score in reading of student *i* in school *j* (centered on its grand mean)

S_{ij}= vector of student characteristics for student *i* in school *j* (centered on their grand means across the sample)

T_{ij}= vector of teacher characteristics for teacher of student *i* in school *j* (centered on their grand means across the sample)

r_{0j} = school-level residual error, assumed to be independently and identically distributed as $r_{0j} \sim N(0, \tau^2)$

ε_{ij} = student-level residual error, assumed to be independently and identically distributed as $\varepsilon_{ij} \sim N(0, \sigma^2)$.

In this model, γ_{01k} $(k = 1, \cdots, 5)$ represents the average program impact in District k; γ_{00k} is the regression-adjusted mean achievement for students in schools randomly assigned to control in District k; and $\gamma_{00k} + \gamma_{01k}$ is the regression-adjusted mean achievement of students in schools randomly assigned to MAP in the kth district. The mean of the estimates of γ_{01k} $(k = 1, \cdots, 5)$, weighted by the number of schools in each district, is an estimate of the overall average impact,

γ (that is, $\gamma = \sum_{k=1}^{5} w_k \gamma_{01k}$, where the weights $w_k = m_k / \left(\sum_{k=1}^{5} m_k\right)$, and m_k is the number of schools in each district). Similarly, the weighted mean of γ_{00k} represents the overall mean achievement in the control schools, and the weighted mean of ($\gamma_{00k} + \gamma_{01k}$) is the overall mean achievement in the MAP schools.[69] Because District 1 had substantially more schools (20 schools) than the other four districts (2–4 schools each), the estimated impact in this district carried more weight (62.5 percent) than the other four districts combined (37.5 percent). This means that the overall impact estimate was pulled toward the impact estimate in District 1, which is appropriate because it is also estimated with the greatest precision.

The outcomes and covariates included in the model are shown in table B.1. The student characteristics vector, S_{ij}, includes free or reduced-price lunch status, disability status, gender, race/ethnicity (coded as an indicator for White), and limited English proficiency status (coded as an indicator for native English speaker).[70] The teacher characteristics vector, T_{ij}, includes gender, an indicator for whether a teacher has a graduate degree, the number of years teaching experience in reading or English language arts, an indicator for whether a teacher has a permanent (standard) teaching license, and an indicator for White versus racial/ethnic minority. A check for baseline balance on the student characteristics (see chapter 2) indicated that the MAP and control groups did not systematically differ on baseline student, teacher, or school characteristics in grade 4 or grade 5. Although no systematic baseline imbalance was found that could bias the impact estimates, these covariates were included in the impact models, because these subgroups were of interest in the study and could potentially increase the precision of the regression estimates.

[69] A similar weighting procedure was applied in computing the standard error of the estimates. For example, the standard error for the overall impact estimate, $\hat{\gamma}$, is given by $\sqrt{(\sum_{k=1}^{5} w_k^2 Var(\hat{\gamma}_k))}$, where $Var(\hat{\gamma}_k)$ is the variance of the estimated impact in the kth district.

[70] Eligibility for free or reduced-price lunch, race/ethnicity, and limited English proficient status were recoded from the original variables, which had more categories, some of which had very few or no entries. Specifically, eligibility for free or reduced-price lunch originally included the following categories: free lunch, reduced-price lunch, and full pay); race/ethnicity originally included the following categories: White, Black, Hispanic, Native American/Alaskan, Asian, and more than one race). Limited English proficient status originally included the following categories: native, fluent, limited English, non-English speaking, reclassified fluent).

Table B.1. Variables included in the impact model

Outcomes	Covariates
• Illinois Standards Achievement Test [ISAT] 2010 reading scale score (*ISATReadscale10*) • Composite (average) of scale scores on spring 2010 MAP tests in reading (reading Survey Goals 2–5 IL V2) (N*RRITscoreSP10*) and in Language Usage (Language Survey Goals IL V2) (*LRITScoreSP10*)	*Student variables* • ISAT 2009 reading scale score (*ISATReadscale09*) • Race/ethnicity 0 = racial/ethnic minority 1 = White • Free or reduced-price lunch status 0 = not eligible 1 = eligible • Gender 0 = female 1 = male • Limited English proficiency status[a] 0 = not proficient in English 1 = proficient in English • Disability status 0 = disability 1 = no disability
	Teacher variables • Gender 0 = female 1 = male • Graduate degree 0 = has graduate degree 1 = has no graduate degree • Teaching experience (years of teaching experience in reading/English language arts) • Licensure status 0 = has initial license 1 = has permanent license • Teacher race/ethnicity 0 = racial/ethnic minority 1 = White
	School variable • Average ISAT 2009 scores in reading of all grade 4 or grade 5 students in the school (*MEANISAT*09)

a. A student was considered English proficient if he or she was classified as native or fluent.

Source: Authors' compilation based on data from the study districts and the Northwest Evaluation Association.

Impact estimates

The impact estimates presented in this report are intent-to-treat estimates. The analyses of overall impacts presented in chapters 4 and 5 and discussed in more detail below are based on the full randomized sample of eligible schools, teachers, and students regardless of their actual receipt of the MAP intervention.[71] The full sample included 32 schools that were randomized to treatment or control condition in either grade 4 or grade 5; 172 teachers (85 grade 4 teachers and 87 grade 5 teachers); and 3,720 students (1,914 grade 4 students and 1,806 grade 5 students) present in the fall of Year 2. It included teachers and students from less than four schools that opted not to participate in the study shortly after randomization. Students from these schools did not take any of the MAP tests and therefore have missing MAP outcomes. All of them have nonmissing ISAT outcomes, and most have nonmissing baseline characteristics.

For the core analyses, the study team used a multiple imputation method (described in appendix D) to impute missing values on both outcomes and covariates (Puma et al. 2009). To examine the robustness of the findings to the approach chosen to deal with the missing data, the study team also performed listwise deletion of these students and repeated the analyses separately by grade. Appendix C presents the results of these analyses, along with other sensitivity analyses that included only students with complete outcomes. Results of these sensitivity analyses are consistent with the core analysis results that there were no significant overall impacts on the ISAT or MAP composite scores in either grade 4 or grade 5. In all core and sensitivity analyses conducted, a multiple imputation procedure was used to impute missing values on the outcomes and covariates, creating 20 sets of completed datasets. Each completed dataset was analyzed separately; estimates from these analyses were then pooled, as described in appendix D. Throughout this report, the estimated impact results presented are combined estimates from the separate analyses.

Details of the analyses that produced the overall impacts presented in chapters 4 and 5, including the estimated relationships between baseline covariates and outcomes, are discussed below.

The core model (Model 4) is the full model given by equation B.1, the basis of the overall impact results in chapters 4 and 5. To assess the effect of other covariate specifications on the impact estimates, the study team also fitted three other models.

Model 1 (unadjusted model) includes the first two factors in equation B.1:

$$Y_{ij} = \sum_k \gamma_{00k} D_k + \sum_k \gamma_{01k} Trt_j D_k + r_{0j} + \varepsilon_{ij}.$$

Model 2 (pretest model) extends Model 1 by adding the student- and school-level pretests as covariates:

$$Y_{ij} = \sum_k \gamma_{00k} D_k + \sum_k \gamma_{01k} Trt_j D_k + \gamma_{02} MEANISAT09_j + ISATREADSCALE09_{ij} + r_{0j} + \varepsilon_{ij}.$$

[71] Chapter 2 gives the eligibility criteria for students, teachers, and schools.

Model 3 (student + school covariates model) extends Model 2 by adding the student demographic characteristics as covariates:

$$Y_{ij} = \sum_k \gamma_{00k} D_k + \sum_k \gamma_{01k} Trt_j D_k + \gamma_{02} MEANISAT09_j + \gamma_{10} ISATREADSCALE09_{ij} + S_{ij}\Gamma_{20} + r_{0j} + \varepsilon_{ij}.$$

Model 4, the full model given by equation B.1, adds teacher demographic variables to Model 3.

These models are cumulative, in the sense that every model contains more covariates than the model that precedes it. Model 1 is an unadjusted model that includes a dummy variable for the treatment group and dummy indicators for districts but no covariates. This model yields estimates of mean achievement levels of the control group in each district and estimates of the impact of the MAP intervention (that is, differences in achievement levels between MAP and control groups) that are unadjusted for student, teacher, or school characteristics. Model 2 extends Model 1 by adjusting the mean achievement levels and impact estimates for student- and school-level pretests. Model 3 extends Model 2 further by adjusting for the differences between the MAP and control groups using both student- and school-level pretests as well as student demographic characteristics. Model 4, the core model, adjusts impact estimates for both pretests as well as for student and teacher baseline characteristics. Because the study team already controlled for pretest at the student level in Models 2, 3 and 4, the parameter for the school-level pretest in these models captures the contextual effect of school mean ability on student reading achievement or its effect beyond what can be explained by individual achievement (Model 2), by individual achievement and other baseline student characteristics (Model 3), or by individual achievement and student and teacher characteristics.

Tables B.2–B.5 summarize the parameter estimates from the four models considered. The parameters labeled "Intercept (Control)" give estimates of the adjusted mean achievement levels of control schools in each district. The parameters labeled "Impact: MAP—Control" give the estimated impacts (that is, the difference in the adjusted mean achievement levels between the MAP and control groups) in each district.[72] The other parameter estimates capture the effect of each control variable on reading achievement. The last rows in each table show the decomposition of the total variance in student achievement into between-school and within-school components, the ICC (which measures the proportion of the total variance in achievement that lies between schools),[73] and the percentages of the between- and within-school residual variances explained by the covariates.

Impact on grade 4 student achievement

MAP had no statistically significant district-specific impacts on ISAT reading scores, either before controlling for baseline characteristics (Model 1) or after controlling for them (Models 2–4) (tables B.2 and B.3). The school pretest and the individual pretest are the most important predictors of ISAT reading achievement. Together (without the help of other baseline covariates) they explain about 82 percent of the school-to-school variability and 57 percent of the within-school variability in the ISAT scores and about 77 percent of school-to-school variability and 45

[72] These estimates were pooled to obtain the impact estimates shown in table 4.1 in chapter 4.

[73] Intraclass correlation coefficient is defined as the ratio of the between-school variance to the total variance.

percent of the within-school variability in the MAP composite scores (see last two rows under Model 2 in tables B.2 and B.3). Only the student-level pretest was statistically significant in Models 2–4, suggesting that there were no compositional effects attributable to overall school achievement levels. In the models that adjusted for student demographic characteristics (Models 3 and 4), three of the five student demographic variables were consistently significantly related to both ISAT and MAP posttest scores: being eligible for free or reduced-price lunch, being White, and having disability status. The coefficients of these variables indicate that economically more advantaged students outscore those of lower socioeconomic status, White students outperform racial/ethnic minority students, and students with no disability status score better than students with disability status. None of the teacher variables was significantly related to the posttest ISAT scores; two— gender and licensure status—were related to the MAP posttest scores.

Impact on grade 4 ISAT 2010 reading scores. Model 1, the unadjusted model, serves as a baseline model against which the other models can be compared. The ICC for this model shows that 9.0 percent of the variability in student achievement was attributable to school differences (table B.2). This ICC indicates that there is some degree of clustering of reading achievement within schools that warrants the use of hierarchical modeling. The estimate is less than the ICC value of .13 that was assumed for the power calculations conducted before the study.[74]

Model 2 controls for student- and school-level pretests, which together explain about 57 percent of the within-school variability and 82 percent of the between-school variability in reading achievement. Only the estimated coefficient for the student-level pretest was significant, with a value of 0.70, indicating that a one-point increase in a student's prior ISAT score was associated with an average increase of 0.70 scale score points in student achievement level. Controlling for these covariates substantially increased the precision of all impact estimates relative to the unadjusted model (Model 1), as shown by the smaller standard errors. Nevertheless, all the impact estimates remained statistically nonsignificant.

Model 3 extends Model 2 by adding the baseline student characteristics. Inclusion of these covariates resulted in only slight increases in the within-school variance (from 57 percent to 58 percent) and between-school variance (from 82 percent to 84 percent) explained. These results indicate that the school and individual pretests combined explain almost all of the variability in achievement scores within and between schools, a finding that is consistent with prior research (for example, Bloom, Richburg-Hayes, and Black 2007). Three of the student demographic characteristics were statistically significant: students who are economically more advantaged had an achievement level that was 3.3 scale score points higher than less advantaged students, the achievement level of White students was 4.0 points higher than that of racial/ethnic minorities, and students with no disability status scored 5.0 points higher on average than students with disability status. The student-level pretest remained statistically significant, with its estimated effect about the same as in Model 2. The conditional ICC estimate in Model 3 was substantially lower (0.04), indicating that after controlling for both pretests and individual student characteristics, only 4 percent of the total variability in ISAT scores was between schools.

Model 4 adds teacher demographic variables to the covariates in Model 3. Inclusion of these demographic variables did not explain within-school variability over and above that explained by

[74] See chapter 2 for a discussion of the statistical power calculation for the confirmatory analysis of Year 2 outcomes.

Model 3 (which explained about 58 percent). In fact, it caused a 3.8 percent decrease in the proportion of between-school variability explained (from 84.5 to 80.7 percent).[75] Only the individual pretest covariate was statistically significant, suggesting that, among students with similar baseline characteristics, there was no added advantage in being a member of a school with a higher mean achievement level. The same three student demographic variables as in Model 3 were significant, with comparable effects on achievement. The teacher demographic variables had no significant effects on achievement levels. In all models, the directions of the effects of all statistically significant covariates were in the direction expected based on prior research.

Table B.2. Estimates of regression coefficients for the impact of Measures of Academic Progress (MAP) on Illinois Standards Achievement Test (ISAT) 2010 scores of grade 4 students in Year 2

	Model 1 (unadjusted)		Model 2 (pretest)		Model 3 (student and school covariates)		Model 4 (full)	
Parameter	**Estimate (standard error)**	**Effect size**	**Estimate (standard error)**	**Effect size**	**Estimate (standard error)**	**Effect size**	**Estimate (standard error)**	**Effect size**
District								
District 1								
Intercept (control)	209.2	8.11	216.9	8.41	213.4	8.27	213.5	8.28
	(2.64)		(1.68)		(2.05)		(2.53)	
Impact: MAP—control	0.6	0.02	2.1	0.08	1.4	0.05	1.7	0.07
	(3.86)		(1.94)		(1.86)		(2.06)	
District 2								
Intercept (control)	224.6	8.71	218.1	8.45	212.4	8.23	212.6	8.24
	(8.09)		(3.83)		(3.94)		(4.44)	
Impact: MAP—control	3.5	0.14	1.4	0.06	1.6	0.06	1.7	0.07
	(11.59)		(5.41)		(5.15)		(5.57)	
District 3								
Intercept (control)	231.1	8.96	219.5	8.51	213.8	8.29	214.4	8.31
	(6.07)		(3.30)		(3.36)		(3.83)	
Impact: MAP—control	1.5	0.06	4.2	0.16	4.1	0.16	3.7	0.14
	(8.46)		(4.03)		(3.84)		(4.17)	
District 4								
Intercept (control)	228.9	8.88	224.5	8.7	219.7	8.52	219.8	8.52
	(8.11)		(3.79)		(3.86)		(4.43)	
Impact: MAP—control	0.2	0.01	–5.1	–0.2	–4.8	–0.18	–4.5	–0.17

[75] Changes in the "wrong" direction can be attributed to either chance fluctuations or to misspecification of the covariates with fixed coefficients in the expanded model (that is, the model that contains the added covariates) (Snijders and Bosker 1999, p. 104). With large sample sizes, a decrease of 5 percent or more may be an indication of misspecification; small decreases may reflect random fluctuations (Snijders and Bosker 1999, p. 123). Based on these criteria, it is reasonable to assume that the observed decrease (of 3.8 percent) in the proportion of between-school variance explained was simply a chance fluctuation.

Parameter	Model 1 (unadjusted)		Model 2 (pretest)		Model 3 (student and school covariates)		Model 4 (full)	
	Estimate (standard error)	Effect size	Estimate (standard error)	Effect size	Estimate (standard error)	Effect size	Estimate (standard error)	Effect size
	(9.89)		(4.58)		(4.35)		(4.74)	
District 5								
Intercept (control)	226.0	8.76	220.1	8.53	214.7	8.32	215.0	8.33
	(8.32)		(4.12)		(4.21)		(4.50)	
Impact: MAP—control	–4.9	–0.19	0.2	0.01	0.0	0	0.5	0.02
	(10.11)		(4.82)		(4.61)		(4.97)	
Pretest								
MEANISAT09			0.2	0.01	0.1	0	0.1	0.00
			(0.11)		(0.10)		(0.11)	
ISATReadscale09			0.7*	0.03	0.7*	0.03	0.7*	0.03
			(0.02)		(0.03)		(0.03)	
Student demographics								
Free or reduced-price lunch					3.3*	0.13	3.2*	0.13
Eligible—noneligible					(1.07)		(1.08)	
Race/ethnicity					4.0*	0.16	4.1*	0.16
White—racial/ethnic minority					(1.08)		(1.08)	
Disability status					5.0*	0.19	4.9*	0.19
No—yes					(1.35)		(1.36)	
Limited English proficiency status					0.2	0.01	0.3	0.01
Proficient—not proficient					(2.25)		(2.25)	
Gender					0.0	0.00	0.0	0.00
Female—male					(0.81)		(0.81)	

Parameter	Model 1 (unadjusted)		Model 2 (pretest)		Model 3 (student and school covariates)		Model 4 (full)	
	Estimate (standard error)	Effect size	Estimate (standard error)	Effect size	Estimate (standard error)	Effect size	Estimate (standard error)	Effect size
Teacher demographics								
Gender							–0.2	–0.01
Female—male							(1.41)	
Graduate degree							–0.5	–0.02
With—without							(1.25)	
Teaching experience in English language arts							0.0	0.00
Years							(0.08)	
Licensure status							1.6	0.06
Permanent—initial							(1.42)	
Race/ethnicity							–2.2	–0.09
White—racial/ethnic minority							(2.01)	
Variation between and within schools								
Between school	58.2		10.3		9.1		11.2	
Within school	610.6		261.9		255.7		255.4	
Intraclass correlation	0.09		0.04		0.03		0.04	
Percentage of between-school variance explained			82.4		84.5		80.7	
Percentage of within-school variance explained			57.1		58.1		58.2	

* Statistically significantly different from zero at the .05 significance level.

Note: Estimated MAP and control means and the difference between the two were regression adjusted to account for clustering of students within schools, district effects, and baseline student, teacher, and school characteristics; they were then weighted by the number of schools in each district. Effect sizes were computed by dividing the impact by the standard deviation of the outcome for the control group. *p*-values are from a two-tailed test of the null hypothesis of equality of MAP and control means.

Source: Authors' analysis based on Year 2 (2009–10) data from the study districts and the Northwest Evaluation Association.

Impact on grade 4 2010 MAP reading and language use composite scores. Applying Models 1–4 to the MAP composite scores in grade 4 yielded similar results. The district-specific impacts on MAP composite scores were not statistically significant either before adjusting for baseline characteristics (Model 1) or after (Models 2–4). The school and individual (ISAT) pretests were the most important predictors of MAP composite scores. Together (without other baseline covariates), they explain about 45 percent of the within-school and 77 percent of the between-school variance in MAP scores. Although they explain substantial portions of the total variance, these percentages are considerably smaller than the figures for the ISAT posttest scores, probably because the ISAT pretest and the MAP outcome represent different instruments (that is, different

content and scales). As before, only the student-level pretest was statistically significant in the adjusted models (Models 2–4). The same student demographic variables were significantly related to MAP scores (Models 3 and 4). Their estimated coefficients indicate that economically more advantaged students scored about 2.0 points higher than less advantaged students, White students outperformed racial/ethnic minority students by about 1.6 points, and students with no disability status scored 4.1 points higher than those with disability status. None of the teacher variables was significantly related to ISAT scores. In contrast, teacher gender and licensure status were significantly related to MAP scores: on average students taught by men scored 2.0 points higher than those taught by women, and students of teachers with permanent teaching licenses outperformed students of teachers with initial licenses by about 1.9 points.

Table B.3. Estimates of regression coefficients for the impact of Measures of Academic Progress (MAP) on MAP 2010 composite scores in reading and language usage of grade 4 students in Year 2

	Model 1 (unadjusted)		Model 2 (pretest)		Model 3 (student and school covariates)		Model 4 (full)	
Parameter	**Estimate (standard error)**	**Effect size**	**Estimate (standard error)**	**Effect size**	**Estimate (standard error)**	**Effect size**	**Estimate (standard error)**	**Effect size**
District								
District 1								
Intercept (control)	199.2	15.19	203.0	15.48	199.3	15.21	200.9	15.32
	(1.42)		(1.04)		(1.31)		(1.48)	
Impact: MAP—control	0.1	0.01	0.9	0.07	0.5	0.04	1.2	0.09
	(2.03)		(1.20)		(1.19)		(1.19)	
District 2								
Intercept (control)	204.5	15.60	201.3	15.36	196.6	15.00	198.8	15.16
	(4.25)		(2.28)		(2.41)		(2.49)	
Impact: MAP—control	3.5	0.27	2.5	0.19	2.7	0.21	2.3	0.17
	(6.04)		(3.19)		(3.17)		(3.09)	
District 3								
Intercept (control)	208.6	15.91	202.8	15.47	198.1	15.11	200.9	15.33
	(3.14)		(1.92)		(2.02)		(2.14)	
Impact: MAP—control	1.9	0.14	3.2	0.25	3.1	0.23	2.3	0.17
	(4.38)		(2.38)		(2.35)		(2.30)	
District 4								
Intercept (control)	210.8	16.08	208.6	15.91	204.1	15.57	206.1	15.72
	(4.21)		(2.22)		(2.35)		(2.44)	
Impact: MAP—control	1.4	0.10	–1.3	–0.10	–1.0	–0.08	–0.7	–0.06
	(5.13)		(2.69)		(2.67)		(2.60)	
District 5								
Intercept (control)	210.2	16.03	207.3	15.81	202.7	15.46	204.2	15.58
	(4.27)		(2.34)		(2.45)		(2.44)	
Impact: MAP—control	–4.2	–0.32	–1.7	–0.13	–1.9	–0.14	–1.3	–0.10
	(5.19)		(2.81)		(2.78)		(2.70)	

Parameter	Model 1 (unadjusted)		Model 2 (pretest)		Model 3 (student and school covariates)		Model 4 (full)	
	Estimate (standard error)	Effect size	Estimate (standard error)	Effect size	Estimate (standard error)	Effect size	Estimate (standard error)	Effect size
Pretests								
MEANISAT09			0.1	0.01	0.1	0.01	0.1	0.01
			(0.06)		(0.06)		(0.06)	
ISATReadscale09			0.3*	0.02	0.3*	0.02	0.3*	0.02
			(0.01)		(0.01)		(0.01)	
Student demographics								
Free or reduced-price lunch					2.0*	0.15	2.0*	0.15
Eligible—noneligible					(0.62)		(0.62)	
Race/ethnicity					1.6*	0.12	1.7*	0.13
White—racial/ethnic minority					(0.62)		(0.61)	
Disability status					4.1*	0.31	4.0*	0.31
No—yes					(0.76)		(0.76)	
Limited English proficiency status					0.3	0.02	0.5	0.03
Proficient—not proficient					(1.21)		(1.21)	
Gender					0.5	0.04	0.5	0.04
Female—male					(0.47)		(0.47)	
Teacher demographics								
Gender							–2.2*	–0.17
Female—male							(0.76)	
Graduate degree							–1.1	–0.09
With—without							(0.68)	
Teaching experience in English language arts							0	0
Years							(0.05)	
Licensure status							1.9*	0.14
Permanent—initial							(0.79)	
Race/ethnicity							–2.0	–0.15
White—racial/ethnic minority							(1.15)	
Variation between and within schools								
Between school	15.8		3.7		3.7		3.4	
Within school	153.4		84.1		81.5		80.8	
Intraclass correlation	0.09		0.04		0.04		0.04	

Parameter	Model 1 (unadjusted)		Model 2 (pretest)		Model 3 (student and school covariates)		Model 4 (full)	
	Estimate (standard error)	Effect size	Estimate (standard error)	Effect size	Estimate (standard error)	Effect size	Estimate (standard error)	Effect size
Percentage of between-school variance explained	—		76.9		77.2		79.0	
Percentage of within-school variance explained	—		45.2		46.9		47.3	

* Statistically significantly different from zero at the .05 significance level.

Note: Estimated MAP and control means and the difference between the two were regression adjusted to account for clustering of students within schools, district effects, and baseline student, teacher, and school characteristics; they were then weighted by the number of schools in each district. Effect sizes were computed by dividing the impact by the standard deviation of the outcome for the control group. *p*-values are from a two-tailed test of the null hypothesis of equality of MAP and control means.

Source: Authors' analysis based on Year 2 (2009–10) data from the study districts and the Northwest Evaluation Association.

Impact on grade 5 student achievement

There were no statistically significant district-specific impacts on ISAT reading scores, either before controlling for baseline characteristics (Model 1) or after (Models 2–4). As in grade 4, the school pretest and the individual pretest were the most important predictors of ISAT reading achievement. Together (without other baseline covariates) they explain about 86 percent of school-to-school variability on the ISAT scores and 59 percent on the MAP composite scores; they explain about 49 percent of within-school variability on the ISAT and 41 percent on the MAP composite scores (see last two rows under Model 2 in tables B.4 and B.5). Only the student-level pretest was statistically significant in Models 2–4, however, suggesting that there were no compositional effects attributable to overall school achievement levels. In the models that adjusted for student demographic characteristics (Models 3 and 4), four of the five student demographic variables were consistently significantly related to ISAT posttest scores (the indicators for free or reduced-price lunch, disability status, English proficiency, and gender), and three (the indicators for free or reduced-price lunch, race/ethnicity, and disability status) were consistently significantly associated with MAP posttest scores. The coefficients of these variables indicate that females outscored males, students not eligible for free or reduced-price lunch outscored students who were eligible, White students outperformed racial/ethnic minority students, and students with no disability status outscored students with disability status. None of the teacher variables was statistically significantly related to the posttest MAP scores, but teacher experience was significantly related to the ISAT posttest scores. Results for each outcome are discussed below.

Impact on grade 5 ISAT 2010 reading scores. Model 1, the unadjusted model, serves as a baseline model against which the other models can be compared. The ICC for this model shows that 7.0 percent of the variability in student achievement was explained by school differences (table B.4). This correlation indicates that there is some degree of dependence of reading achievement within schools that warrants the use of multilevel modeling. As with the ICC

estimate in grade 4, this estimate is lower than the value of .13 that was assumed for the power calculations conducted before the study was implemented.[76]

Model 2 controls for student- and school-level pretests, which together explain about 49 percent of the within-school variability and 86 percent of the between-school variability in reading achievement. Only the estimated coefficient for the student-level pretest was significant, with a value of 0.70, indicating that a one-point increase in a student's prior ISAT score was associated with an average increase of 0.70 scale-score points in student achievement level. Controlling for these covariates substantially increased the precision of all impact estimates relative to the unadjusted model (Model 1), as shown by the smaller standard errors Nevertheless, all differences between MAP and control students' achievement remained statistically nonsignificant.

Model 3 extends Model 2 by adding the baseline student characteristics. Inclusion of these covariates resulted in only slight increases in the proportion of within-school variance explained (from 49 percent to 52 percent) and between-school variance explained (from 86 percent to 88 percent) over Model 2. There results indicate that the school and individual pretests combined explain almost all of the variability in achievement scores within and between schools, a finding that is consistent with prior research (for example, Bloom, Richburg-Hayes, and Black 2007). Four of the student demographic characteristics were statistically significant, indicating that on average students who were not eligible for free or reduced-price lunch outscored students who were eligible by 4.3 scale score points, students with no disability status fared outscored students with disability status by 6.8 points, English-proficient students outperformed students who are not proficient by 5.4 points, and females outperformed males by 2.3 points. The student-level pretest remained statistically significant; its estimated effect was about the same as in Model 2. The conditional ICC estimate in Model 3 was substantially lower (.02), indicating that after controlling for both pretests and individual student characteristics, only 2 percent of the total variability in ISAT scores was between schools.

Model 4 adds teacher demographic variables to the covariates in Model 3. Inclusion of these demographic variables did not explain student- or school-level variability over and above that explained by Model 3: both models explain about 52 percent of the student-level variance and about 88–89 percent of the school-level variance. Only the individual pretest covariate was statistically significant, suggesting that among students with similar baseline characteristics, there was no added advantage of attending a school with a higher mean achievement level. The same four student demographic variables as in Model 3 were significant; their effects were comparable to those in Model 3. Teacher experience was statistically significantly associated with lower ISAT scores—that is, for every additional year of teacher experience, students scored 0.1 points worse—although the magnitude of this association is probably too small to be of practical significance.

[76] See chapter 2 for a discussion of the statistical power calculation for confirmatory analysis of Year 2 outcomes.

Table B.4. Estimates of regression coefficients for the impact of Measures of Academic Progress (MAP) on Illinois Standards Achievement Test (ISAT) 2010 scores of grade 5 students in Year 2

Parameter	Model 1 (unadjusted)		Model 2 (pretest)		Model 3 (student and school covariates)		Model 4 (full)	
	Estimate (standard error)	Effect size	Estimate (standard error)	Effect size	Estimate (standard error)	Effect size	Estimate (standard error)	Effect size
District								
District 1								
Intercept (control)	224.6	8.87	232.4	9.18	224.7	8.88	223.9	8.85
	(2.58)		(1.61)		(2.03)		(2.46)	
Impact: MAP—control	–3.6	–0.14	–3.3	–0.13	–3.5	–0.14	–3.2	–0.13
	(3.58)		(1.9)		(1.78)		(1.86)	
District 2								
Intercept (control)	241.1	9.52	232.3	9.18	223.0	8.81	222.6	8.79
	(7.35)		(3.51)		(3.66)		(4.18)	
Impact: MAP—control	–5.8	–0.23	–0.8	–0.03	–0.9	–0.04	0.8	0.03
	(10.38)		(4.7)		(4.37)		(4.51)	
District 3								
Intercept (control)	239.2	9.45	233.2	9.21	223.7	8.83	223.6	8.83
	(5.36)		(2.66)		(2.94)		(3.16)	
Impact: MAP—control	2.8	0.11	2.1	0.08	2.1	0.08	1.8	0.07
	(7.73)		(3.7)		(3.45)		(3.46)	
District 4								
Intercept (control)	238.1	9.41	228.8	9.04	221.1	8.73	219.6	8.67
	(5.07)		(2.49)		(2.83)		(3.34)	
Impact: MAP—control	–2.7	–0.1	0.7	0.03	1.1	0.04	0.1	0
	(8.85)		(3.91)		(3.63)		(3.63)	
District 5								
Intercept (control)	232.5	9.18	231.8	9.16	222.9	8.8	221.5	8.75
	(5.2)		(2.36)		(2.74)		(3.33)	
Impact: MAP—control	6.1	0.24	3.9	0.15	3.1	0.12	2.6	0.1
	(8.99)		(4.03)		(3.75)		(3.83)	
Pretests								
MEANISAT09			0.2	0.01	0.1	0	0.1	0
			(0.11)		(0.1)		(0.1)	
ISATReadscale09			0.7*	0.03	0.6*	0.02	0.6*	0.03
			(0.02)		(0.02)		(0.02)	
Student demographics								
Free or reduced-price lunch					4.3*	0.17	4.2*	0.16
Full—free/reduced					(1.18)		(1.18)	
Race					1.9	0.07	1.9	0.07

Parameter	Model 1 (unadjusted)		Model 2 (pretest)		Model 3 (student and school covariates)		Model 4 (full)	
	Estimate (standard error)	Effect size	Estimate (standard error)	Effect size	Estimate (standard error)	Effect size	Estimate (standard error)	Effect size
White—racial/ethnic minority					(1.12)		(1.12)	
Disability status					6.8*	0.27	6.9*	0.27
No—yes					(1.4)		(1.4)	
Limited English proficiency status					5.4*	0.21	5.7*	0.23
Proficient—not proficient					(2.62)		(2.63)	
Student gender					2.3*	0.09	2.3*	0.09
Female—male					(0.84)		(0.84)	
Teacher demographics								
Teacher gender							1.9	0.08
Female—male							(1.64)	
Graduate degree							2.3	0.09
With—without							(1.17)	
Teaching experience in English language arts							–0.1*	–0.01
Years							(0.07)	
Licensure status							0.5	0.02
Permanent—initial							(1.52)	
Teacher race							1.1	0.04
White—racial/ethnic minority							(1.79)	
Variation between and within schools								
Between schools	46.0		6.6		5.3		5.2	
Within schools	569.2		287.8		275.4		274.8	
Intraclass correlation	0.07		0.02		0.02		0.02	
Percentage of between-school variance explained			85.7		88.4		88.7	
Percentage of within-school variance explained			49.4		51.6		51.7	

* Statistically significantly different from zero at the .05 significance level.

Note: Estimated MAP and control means and the difference between the two were regression adjusted to account for clustering of students within schools, district effects, and baseline student, teacher, and school characteristics; they were then weighted by the number of schools in each district. Effect sizes were computed by dividing the impact by the standard deviation of the outcome for the control group. *p*-values are from a two-tailed test of the null hypothesis of equality of MAP and control means.

Source: Authors' analysis based on Year 2 (2009–10) data from the study districts and the Northwest Evaluation Association.

Impact on grade 5 2010 MAP reading and language use composite scores. The results of applying Models 1–4 to the MAP composite scores in grade 5 were similar to those for the ISAT reading scores. The district-specific impacts on MAP composite scores were not statistically significant either before adjusting for baseline characteristics (Model 1) or after (Models 2–4). The school and individual ISAT pretests were the most important predictors of MAP composite scores. Together (without other baseline covariates) they explained about 41 percent of the within-school and 59 percent of the between-school variances in MAP scores. As observed in the grade 4 analysis results, the ISAT pretests explained substantial portions of the total variance in MAP scores but considerably less of the total variance in ISAT posttest scores. The difference probably reflects the fact that the ISAT pretest and the MAP outcome represent different instruments (that is, different content and scales). As before, only the student-level pretest was statistically significant in the adjusted models (Models 2–4). Three student demographic variables were significantly related to MAP scores (Models 3 and 4). Their estimated coefficients indicate that on average, economically more advantaged students scored about 1.5 points higher than less advantaged students, White students outperformed racial/ethnic minority students by about 1.6 points, and students with no disability scored 4.7 points higher than those with a disability. In contrast to the ISAT scores, no teacher variables were significantly related to the MAP scores.

Table B.5. Estimates of regression coefficients for the impact of Measures of Academic Progress (MAP) on MAP 2010 composite scores in reading and language usage scores of grade 5 students in Year 2

	Model 1 (unadjusted)		Model 2 (pretest)		Model 3 (student and school covariates)		Model 4 (full)	
Parameter	**Estimate (Standard error)**	**Effect size**	**Estimate (standard error)**	**Effect size**	**Estimate (standard error)**	**Effect size**	**Estimate (standard error)**	**Effect size**
District								
District 1								
Intercept (control)	205.4	16.4	208.9	16.65	204.8	16.32	204.7	16.31
	(1.34)		(1.16)		(1.35)		(1.55)	
Impact: MAP—control	–0.2	0.0	–0.1	–0.01	–0.2	–0.01	–0.1	–0.01
	(1.89)		(1.35)		(1.31)		(1.37)	
District 2								
Intercept (control)	212.6	16.9	208.7	16.63	204.0	16.25	204.2	16.27
	(3.9)		(2.74)		(2.83)		(3.07)	
Impact: MAP—control	–0.9	–0.1	1.3	0.1	1.1	0.09	1.6	0.13
	(5.52)		(3.73)		(3.68)		(3.82)	
District 3								
Intercept (control)	214.3	17.1	211.6	16.86	206.6	16.46	206.7	16.47
	(2.8)		(1.96)		(2.09)		(2.2)	
Impact: MAP—control	–0.1	0.0	–0.2	–0.01	0.0	0.00	0.1	0.00
	(4.03)		(2.7)		(2.65)		(2.71)	

Parameter	Model 1 (unadjusted)		Model 2 (pretest)		Model 3 (student and school covariates)		Model 4 (full)	
	Estimate (Standard error)	Effect size	Estimate (standard error)	Effect size	Estimate (standard error)	Effect size	Estimate (standard error)	Effect size
District 4								
Intercept (control)	216.4	17.3	212.3	16.92	208.2	16.59	207.9	16.57
	(2.67)		(1.95)		(2.07)		(2.3)	
Impact: MAP—control	–2.5	–0.2	–1.0	–0.08	–1.0	–0.08	–1.6	–0.12
	(4.66)		(3.07)		(3.03)		(3.1)	
District 5								
Intercept (control)	212.2	16.9	211.8	16.88	207.0	16.49	206.6	16.46
	(2.72)		(1.81)		(1.97)		(2.2)	
Impact: MAP—control	4.1	0.3	3.3	0.26	3.1	0.24	2.7	0.21
	(4.71)		(3.13)		(3.08)		(3.19)	
Pretest								
MEANISAT09			0.1	0.01	0.1	0.00	0.1	0.00
			(0.08)		(0.08)		(0.08)	
ISATReadscale09			0.3*	0.02	0.3*	0.02	0.3*	0.02
			(0.01)		(0.01)		(0.01)	
Student demographics								
Free or reduced-price lunch					1.6*	0.13	1.5*	0.12
Full—free/reduced					(0.61)		(0.61)	
Race/ethnicity					1.6*	0.13	1.6*	0.13
White—racial/ethnic minority					(0.57)		(0.57)	
Disability status					4.7*	0.37	4.7*	0.37
No—yes					(0.72)		(0.72)	
Limited English proficiency status					1.5	0.12	1.6	0.13
Proficient—not proficient					(1.11)		(1.11)	
Gender					0.5	0.04	0.5	0.04
Female—male					(0.44)		(0.44)	

Parameter	Model 1 (unadjusted)		Model 2 (pretest)		Model 3 (student and school covariates)		Model 4 (full)	
	Estimate (Standard error)	Effect size	Estimate (standard error)	Effect size	Estimate (standard error)	Effect size	Estimate (standard error)	Effect size
Teacher demographics								
Gender							0.5	0.04
Female—male							(0.82)	
Graduate degree							1.0	0.08
With—without							(0.68)	
Teaching experience in English language arts							–0.1	0.00
Years							(0.04)	
Licensure status							–0.5	–0.04
Permanent—initial							(0.81)	
Race/ethnicity							0.6	0.04
White—racial/ethnic minority							(1.01)	
Variation between and within schools								
Between school	13.1		5.4		5.2		5.5	
Within school	131.1		78.0		74.3		74.3	
Intraclass correlation	0.09		0.06		0.07		0.07	
Percentage of between-school variance explained			59.1		60.4		58.4	
Percentage of within-school variance explained			40.5		43.3		43.3	

* Statistically significantly different from zero at the .05 significance level.

Note: Estimated MAP and control means and the difference between the two were regression adjusted to account for clustering of students within schools, district effects, and baseline student, teacher, and school characteristics; they were then weighted by the number of schools in each district. Effect sizes were computed by dividing the impact by the standard deviation of the outcome for the control group. *p*-values are from a two-tailed test of the null hypothesis of equality of MAP and control means.

Source: Authors' analysis based on Year 2 (2009–10) data from the study districts and the Northwest Evaluation Association.

Appendix C. Results of Sensitivity Analyses

Shortly after randomization, less than four schools withdrew from the study. Because these schools did not participate, their students were not administered the spring 2010 MAP assessments and therefore have missing MAP scores. Data on ISAT pretest and posttest scores were available on most of these students, as were student and teacher demographic data (see table D.2 in appendix D for the rates of missing data). For the primary intent-to-treat analyses in grades 4 and 5, the study team included these schools in the analytic samples and imputed all missing data on outcomes and covariates not just for these entire schools but for the full sample, using multiple imputation procedure as described in appendix D.

To test the robustness of the findings to the approach used to address missing data, the study team conducted three sets of sensitivity analyses separately for grade 4 and grade 5. In each sensitivity analysis, the data were analyzed separately for the ISAT and MAP outcomes (except in sensitivity analysis I, where the study team analyzed only the ISAT data). The same analytic model used for the core analysis (equation B.1) was used. The goal was to assess how the overall impact estimates vary across these approaches. The samples used for the core analyses and each of these sensitivity analyses are shown in table C.1.

Table C.1. Samples used for core and sensitivity analyses

Type of analysis	Sample description	Sample for ISAT outcome	Sample for MAP outcome
Core analysis (all cases, 32 schools)	Full analytic sample, including all randomized schools, teachers, and students eligible to participate in study. (Missing ISAT and MAP outcomes and covariate data imputed.)	Number of teachers: 85 in grade 4, 87 in grade 5 Number of students: 1,914 in grade 4, 1,806 in grade 5	Same as sample for ISAT outcome
Sensitivity analysis I (complete outcome only, 32 schools)	Subsample of students with nonmissing outcomes from full analytic sample. (Missing covariate data imputed.)	Number of teachers: 85 in grade 4, 87 in grade 5 Number of students: 1,815 in grade 4, 1,719 in grade 5	Complete-case analysis for MAP outcome is not necessary, because sample is same as the one in sensitivity analysis III[a]
Sensitivity analysis II (all cases, with less than four no-show schools excluded)	Subsample obtained by listwise deletion of less than four schools from full analytic sample. (Missing ISAT and MAP outcomes and covariate data imputed.)	Number of teachers: 84 in grade 4, 85 in grade 5 Number of students: 1,885 in grade 4, 1,775 in grade 5	Same as sample for ISAT outcome
Sensitivity analysis III (complete outcome only, with less than four no-show schools excluded)	Subsample of students with nonmissing outcomes only from sample in sensitivity analysis II. (Missing covariate data imputed.)	Number of teachers: 84 in grade 4, 85 in grade 5 Number of students: 1,786 in grade 4, 1,688 in grade 5	Number of teachers: 84 in grade 4, 85 in grade 5 Number of students: 1,752 in grade 4, 1,669 in grade 5

a. Because MAP assessments were not administered in the less than four schools that withdrew from the study, a complete-case analysis of MAP outcome on the 32 schools is equivalent to deleting the less than four schools that withdrew and then conducting a complete-case analysis on the remaining schools, which is identical to the analysis of MAP outcome in sensitivity analysis III.

Source: Authors' compilation based on data from the study districts and the Northwest Evaluation Association.

Sensitivity analysis I

Two recommended approaches for handling missing outcome data on whole schools or missing outcome on students within schools are listwise deletion and multiple imputation (Puma et al. 2009).[77] The first sensitivity analysis was a complete-case analysis that included only grade 4 (or grade 5) students with complete outcome data from the full sample. For this set of analyses, complete-case analysis was conducted only of the ISAT outcome, because a complete-case analysis of the MAP outcome is equivalent to a complete-case analysis of the MAP outcome in the sample that excludes the less than four schools that withdrew (see sensitivity analysis III below).

[77] These recommendations are specific to group randomized trials where the interest is in estimating treatment effect (Puma et al. 2009).

Sensitivity analysis II

The second robustness check was based on the sample that excluded the less than four schools that withdrew from the study. After listwise deletion of these schools, all participants from the remaining schools were included in the analysis. Missing data on outcomes and covariates of these participants were filled in using multiple imputation.

Sensitivity analysis III

A complete-case analysis was conducted on the sample used in sensitivity analysis II. For each of the two outcomes, only students with complete outcome data from the sample that excluded the less than four non-participating schools used in sensitivity analysis II were included.

Results

Following the approach used in the core analyses in grades 4 and 5, for each of the sensitivity analysis, the study team estimated district-specific impacts of the MAP intervention on reading achievement and then weighted them by the number of schools in each district and combined them into an overall impact estimate.

Grade 4

There were no statistically significant overall impacts on grade 4 ISAT (table C.2) or MAP composite scores (table C.3). The impacts were generally small for both outcomes (2.0–5.0 percent in standard deviation units for ISAT scores and 6.3–7.3 percent in standard deviation units for MAP scores). All of the estimated impacts were positive. However, because none was statistically significant, it is not possible to rule out that the impact occurred by chance. Overall, the conclusions on grade 4 overall impacts remained the same across all four analyses, indicating that the manner in which missing outcome data were handled did not alter the overall results.

Grade 5

There were no statistically significant overall impacts on grade 5 ISAT (table C.4) or MAP composite scores (table C.5). The overall impacts were generally small for both outcomes (4.0–7.0 percent in standard deviation units for ISAT scores and 1.0–4.0 percent in standard deviation units for MAP scores). In contrast to grade 4, the estimated impacts on grade 5 ISAT and MAP scores were negative, except for the positive impact on MAP scores for the core (all-case) analysis. The impacts were not statistically significant, however, making it impossible to rule out that they occurred by chance. The conclusions on overall grade 5 impacts were consistent across all analyses, suggesting their robustness to the manner in which missing outcome data were handled.

Table C.2. Results of sensitivity analysis of overall impacts of Measures of Academic Progress (MAP) program on grade 4 Illinois Standards Achievement Test (ISAT) 2010 reading scores

	Mean		Estimated impact			
Type of analysis	MAP	Control	Impact	Standard error	95 percent confidence interval	Effect size
All cases, 32 schools	215.6	214.3	1.29	1.570	–1.79 to 4.37	0.050
Complete outcome only, 32 schools	216.4	215.2	1.21	1.46	–1.66 to 4.08	0.047
All cases, with less than four schools excluded	216.6	216.0	0.59	1.46	–2.26 to 3.44	0.023
Complete outcome only, with less than four schools excluded	216.5	215.6	0.9	1.39	–1.83 to 3.63	0.035

Note: Means were regression adjusted to account for clustering of students within schools and baseline student, teacher, and school characteristics. Means and impact estimates are weighted averages of district-specific estimates, where the weight is equal to the number of schools in each district. Effect sizes were obtained by dividing the impact by the standard deviation of the outcome for the control group of grade 4 students. None of the impact estimates was statistically significant at the .05 significance level.

Source: Authors' analysis based on data from the study districts and the Northwest Evaluation Association.

Table C.3. Results of sensitivity analysis of overall impacts of Measures of Academic Progress (MAP) program on grade 4 MAP 2010 composite scores

	Mean		Estimated impact			
Type of analysis	MAP	Control	Impact	Standard error	95 percent confidence interval	Effect size
All cases, 32 schools	202.49	201.53	0.96	0.891	⁻0.79 to 3.01	0.073
All cases, with less than four schools excluded	202.7	201.9	0.83	0.83	⁻0.79 to 2.45	0.063
Complete outcome only, with less than four schools excluded	202.0	201.0	0.95	0.71	⁻0.43 to 2.33	0.070

Note: Means were regression adjusted to account for clustering of students within schools and baseline student, teacher, and school characteristics. Means and impact estimates are weighted averages of district-specific estimates, where the weight is equal to the number of schools in each district. Effect sizes were obtained by dividing the impact by the standard deviation of the outcome for the control group of grade 4 students. None of the impact estimates was statistically significant at the .05 significance level. Because MAP assessments were not administered in the school that withdrew from the study, a complete-case analysis of MAP outcome on the 32 schools is equivalent to deleting the whole school that withdrew and then conducting a complete-case analysis on the remaining schools. It is therefore not included.

Source: Authors' analysis based on data from the study districts and the Northwest Evaluation Association.

Table C.4. Results of sensitivity analysis of overall impacts of Measures of Academic Progress (MAP) program on grade 5 ISAT 2010 reading scores

	Mean		Estimated impact			
Type of analysis	MAP	Control	Impact	Standard error	95 percent confidence interval	Effect size
All cases, 32 schools	221.7	223.2	–1.48	1.366	–4.16 to 1.20	–0.05
Complete outcome only, 32 schools	221.7	222.6	–0.92	1.253	–3.38 to 1.54	–0.04
All cases, with less than four schools excluded	221.6	223.2	–1.66	1.398	–4.40 to 1.08	–0.07
Complete outcome only, with less than four schools excluded	221.7	222.9	–1.16	1.326	–3.76 to 1.44	–0.05

Note: Means were regression adjusted to account for clustering of students within schools and baseline student, teacher, and school characteristics. Means and impact estimates are weighted averages of district-specific estimates, where the weight is equal to the number of schools in each district. Effect sizes were obtained by dividing the impact by the standard deviation of the outcome for the control group of grade 5 students. None of the impact estimates was statistically significant at the .05 significance level.

Source: Authors' analysis based on data from the study districts and the Northwest Evaluation Association.

Table C.5. Results of sensitivity analysis of overall impacts of Measures of Academic Progress (MAP) program on grade 5 MAP 2010 composite scores

	Mean		Estimated impact			
Type of analysis	MAP	Control	Impact	Standard error	95 percent confidence interval	Effect size
All cases, 32 schools	205.6	205.4	0.15	1.037	–1.89 to 2.17	0.01
All cases, with less than four schools excluded	205.5	205.6	–0.14	0.978	–2.06 to 1.78	–0.01
Complete outcome only, with less than four schools excluded	204.3	204.8	–0.49	0.999	–2.45 to 1.47	–0.04

Note: Means were regression adjusted to account for clustering of students within schools and baseline student, teacher, and school characteristics. Means and impact estimates are weighted averages of district-specific estimates, where the weight is equal to the number of schools in each district. Effect sizes were obtained by dividing the impact by the standard deviation of the outcome for the control group of grade 5 students. None of the impact estimates was statistically significant at the .05 significance level. Because MAP assessments were not administered in the school that withdrew from the study, a complete-case analysis of MAP data outcome on the 32 schools is equivalent to deleting the whole school that withdrew and then conducting a complete-case analysis on the remaining schools. It is therefore not included.

Source: Authors' analysis based on data from the study districts and the Northwest Evaluation Association.

Appendix D. Missing Data Imputation Procedures

This appendix describes the procedures used to impute missing data on implementation fidelity and student outcomes.

Imputation of missing data for assessing implementation

No data collected from attendance sheets or computer-based administrative files on variables related to participation in MAP-related activities or use of MAP resources in instructional planning were missing for MAP teachers (table D.1). Less than 10 percent of data were missing from the teacher surveys for both grade 4 MAP teachers and grade 5 MAP teachers. The percentages of missing data were somewhat higher for MAP teachers in both grade 4 and grade 5. In grade 4, 10.0–12.0 percent of data were missing for MAP teachers and less than 10 percent were missing for control teachers. In grade 5, 18.9–21.6 percent of data on MAP teachers and at most 16.0 percent of data on control teachers were missing.

With regard to classroom observations, the proportions of missing data on measures related to differentiated instruction were 16.0 percent for MAP teachers and 11.4 percent for control teachers in grade 4. In grade 5 these data were missing for 21.6 percent of MAP teachers and 12.0 percent of control teachers.

Missing data for the relevant variables on the teacher logs were less than 10 percent for both grade 4 MAP teachers and grade 4 control teachers. The corresponding percentages for grade 5 teachers were 10.8 percent for MAP teachers and less than 10 percent for control teachers.

The SAS multiple imputation procedure PROC MI was used to impute missing data values. Separate imputations were conducted for MAP teachers and control teachers. Five rounds of imputation were performed. In imputing MAP teachers' use of MAP resources in instructional planning, the study team used teachers' use of other types of resources and grade to impute missing values. On the variables that assessed practices related to differentiated instruction that were collected from the teacher survey, teacher logs, and classroom observations, grade and composite indexes[78] on these variables were used to impute missing values. For example, some teachers had missing composite indexes on the teacher surveys but not the teacher logs and classroom observations. In this case, their composite indexes from the logs and observations were used, along with grade, to impute the differentiated instruction composite indexes on the teacher survey. That is, missing values on composite indexes, rather than missing responses to individual items in the teacher survey, teacher logs, or classroom observation protocol, were imputed. The SAS procedure PROC MIANALYZE was then used to analyze the data from these five imputed data series.

[78] See chapter 3 for a discussion on how these composite indexes were constructed.

Table D.1. Rates of missing data for Year 2 implementation analysis

(percent)

Source of data/variable	Grade 4		Grade 5	
	MAP (*n* = 50)	**Control (*n* = 35)**	**MAP (*n* = 37)**	**MAP (*n* = 50)**
NWEA attendance tracking sheets				
Participation in MAP training				
Attendance at MAP training and consultation sessions	0	na	0	na
NWEA computer-based administrative data				
Overall use of MAP web-based resources in Years 1 and 2				
Proportion of teachers using at least one student goals and planning worksheet	0	na	0	na
Proportion of teachers using at least one MAP Lexile book lists	0	na	0	na
Proportion of teachers using at least one Lexile report	0	na	0	na
Teacher survey				
Use of MAP resources in lesson preparation across Years 1 and 2 (MAP teachers only)				
MAP training and interactions with MAP staff	less than 10.0	na	less than 10.0	na
MAP data reports	less than 10.0	na	less than 10.0	na
MAP DesCartes[a] or Lexiles	less than 10.0	na	less than 10.0	na
Other resources on MAP website	less than 10.0	na	less than 10.0	na
Use of differentiated instruction				
Proportion of ability grouping activities	12.0	less than 10.0	21.6	16.0
Proportion of different literacy topics covered three or more days a week	10.0	less than 10.0	18.9	less than 10.0
Proportion of different instructional strategies used three or more days a week	10.0	less than 10.0	18.9	6.0
Overall survey composite	12.0	less than 10.0	21.6	16.0

Source of data/variable	Grade 4		Grade 5	
	MAP (*n* = 50)	Control (*n* = 35)	MAP (*n* = 37)	MAP (*n* = 50)
Classroom observations				
Use of differentiated instruction across fall, winter, and spring observations				
Average proportion of 10-minute segments during which teacher used instructional groupings	16.0	11.4	21.6	12.0
Average proportion of 10-minute segments during which teacher differentiated content, process, or product	16.0	11.4	21.6	12.0
Average proportion of 10-minute segments during which teacher used any type of differentiated instructional strategies	16.0	11.4	21.6	12.0
Overall observation composite	16.0	11.4	21.6	12.0
Teacher logs				
Average proportion of log rounds during which teacher used different instructional groupings	less than 10.0	less than 10.0	10.8	less than 10.0
Average proportion of log rounds during which teacher covered different topics	less than 10.0	less than 10.0	10.8	less than 10.0
Average proportion of log rounds during which teacher used differentiated instructional strategies for comprehension, writing, or word analysis	less than 10.0	less than 10.0	10.8	less than 10.0
Overall log composite	less than 10.0	less than 10.0	10.8	less than 10.0

na is not applicable.

Note: Percentages less than 10.0 have been suppressed to prevent a disclosure risk.

a. DesCartes is a MAP training tool designed to help teachers target instruction for individual students or groups of students based on MAP test results. DesCartes displays learning statements that describe students' demonstrated knowledge and skills when their MAP test results are within specific ranges.

Source: Authors' analysis based on data from the study districts.

Imputation of missing data for analysis of student outcomes

There were no missing data on teacher variables or the school mean pretest score in either grade 4 or grade 5 (table D.2). Only student variables had missing data. In all cases when the study team imputed missing data on these variables, it used a multiple imputation method. Multiple imputation is a technique that generates several, usually five (Graham, Olchowski, and Gilreath 2007) plausible replacements for missing values in order to generate several sets of completed data. The completed datasets are then analyzed separately using standard analytic tools for complete data. The results from these separate analyses are then combined in a way that takes into account the additional uncertainty introduced into the data by imputing the missing values.

Table D.2. Rates (percent) of missing data

(percent)

	Grade 4			Grade 5		
Variable	**Overall**	**MAP**	**Control**	**Overall**	**MAP**	**Control**
Number of students	1,914	1,149	765	1,806	701	1,105
Student achievement outcomes						
ISATReadscale10	5.2	5.3	5.0	4.8	4.3	5.2
MAPCompositeSP10[a]	8.5	4.3	14.8	7.6	12.7	4.3
NRRITScoreSP10	10.8	5.8	18.2	8.4	13.4	5.2
LRITScoreSP10	16.8	5.6	33.6	18.9	29.2	12.3
Student characteristic						
ISATReadscale09	7.8	7.0	8.9	6.6	6.0	7.1
Eligibility for free or reduced-price lunch	0.6	0.5	0.8	less than 0.5	0.0	less than 0.5
Race/ethnicity	0	0	0	0	0	0
Disability status	1.2	1.9	less than 0.5	less than 0.5	less than 0.5	less than 0.5
Limited English proficiency status	0.9	1.2	0.5	0.9	0.6	1.2
Gender	0	0	0	0	0	0
Teacher characteristic						
Number of teachers	85	50	35	87	37	50
Gender	0	0	0	0	0	0
Graduate degree	0	0	0	0	0	0
Years of experience teaching reading/English language arts	0	0	0	0	0	0
Race/ethnicity	0	0	0	0	0	0
Licensure	0	0	0	0	0	0
School characteristic						

Variable	Grade 4			Grade 5		
	Overall	MAP	Control	Overall	MAP	Control
Number of schools	32	16	16	32	16	16
Mean ISAT 2009 reading scale score	0	0	0	0	0	0
Auxiliary variables[b]						
NRRITScoreFL09	3.7	3.7	na	7.1	7.1	na
LRITScoreFL09	9.5	9.5	na	24.5	24.5	na
NRRITScoreWN10	16.1	16.1	na	12.7	12.7	na
LRITScoreWN10	12.4	12.4	na	28.7	28.7	na

na is not applicable.

Note: Figures include missing values on students from the school in District 1 that withdrew from the study.

a. The MAP composite score was considered missing only if both its components (NRRITScoreSP10 and LRITScoreSP10) were missing.

b. These variables were used in the imputation of student variables but used in the analysis, because data on them were collected only for the MAP group. Missing rates are based on the number of grade 4 or grade 5 students in the MAP group

Source: Authors' compilation based on data from the study districts and the Northwest Evaluation Association.

Several methods and software are available to implement this procedure. The method used in this report was sequential regression multiple imputation method (Raghunathan et al. 2001), described at the end of this section. The procedure was implemented using the IMPUTE module of the IVEware software package, an SAS callable routine developed at the University of Michigan's Survey Research Center (http://www.isr.umich.edu/src/smp/ive/). For each grade, the study team generated 20 sets of filled-in data, analyzed each completed dataset separately using a two-level model (student nested within schools), and then pooled the results of the separate analyses by averaging the point estimates and combining their standard errors, as prescribed by Rubin (1987) and discussed further later in this section.[79] The imputed datasets were analyzed separately by grade using SAS Proc Mixed; the estimates were pooled across imputations using SAS Proc MIANALYZE.

[79] About three to five imputations have been recommended (for example, Rubin 1987; Schafer and Olsen 1998) to obtain excellent results for small to moderate rates of missing data. As Rubin (1987, p. 114) shows, the efficiency of an estimate from m imputations relative to an estimate from infinitely many ($m = \infty$) imputations is approximated, in standard deviation units, by $(1 + \lambda/m)^{-1/2}$, where λ is the rate of missing data. Because the maximum rate of missing data for the outcomes and covariates used in the multiple imputation model is about 34 percent for grade 4 (the missing rate for LRITScoreSP10 scores in the grade 4 control group in table D.2), $m = 20$ imputations would yield an estimate with a standard deviation that is only about 0.8 percent larger than the standard deviation of a fully efficient estimator (because $(1 + \lambda/m)^{1/2} = (1 + 0.34/20)^{1/2} = 1.008$), but $m = 5$ imputations would yield an estimate with a standard deviation that is about 3.3 percent larger. (For grade 5 the maximum missing rate is 29 percent, yielding an estimate whose standard deviation is larger than that of a fully efficient estimator by about 0.7 percent for $m = 20$ and 2.9 percent for $m = 5$ imputations.) Based on relative efficiency considerations, researchers chose to run 20 imputations. This choice also has potential benefits in terms of statistical power. As Graham et al. (2007) note, more imputations are desirable to minimize the decrease in statistical power caused by missing data. Based on a simulation study, they show, for example, that with 30 percent missing rate, the power reduction is less than 1 percent for $m = 20$ and $m = 40$ imputations but about 3 percent $m = 5$ and 7 percent for $m = 10$.

Missing rates

No data were missing on teacher or school variables. For student demographic variables, at most 1.9 percent of data were missing for grade 4 (the missing rate on the disability status for the MAP group) and at most 1.2 percent missing values were missing for grade 5 (the missing rate on limited English proficiency status for the control group). Of the student variables used in the impact models, the ones with the highest rates of missing values are the student achievement variables: ISAT 2009 scores (7.0 percent for the MAP group and 8.9 percent for the control group in grade 4, 6.0 percent for the MAP group and 7.1 percent for the control group in grade 5); ISAT 2010 scores (about 5 percent for both groups in grade 4, 4 percent for MAP group and 5 percent for control group in grade 5); and spring 2010 MAP composite scores (about 4 percent for the MAP group and 15 percent for the control group in grade 4, about 13 percent for the MAP group and 4 percent for the control group in grade 5). [80]

Imputation of student variables

Missing data were imputed separately by grade and treatment group in order to protect against possible bias of coefficients in the impact model (Puma et al 2009, p. 23). The imputation model for the treatment group included other information in the form of auxiliary variables, which may be correlated with the missing variables, the missingness of the variables, or both. These variables included the four MAP assessment scores that were available only for the MAP group: the scores on the fall 2009 tests in reading (NRRITscoreFL09) and language usage (LRITScoreFL09) and the scores on the winter 2010 tests in reading (NRRITscoreWN10) and language use (LRITScoreWN10).

Multiple imputation was carried out separately by grade. Missing data on student characteristics were imputed separately by treatment condition, using student-level, teacher-level, and school variables, as well as district indicators, as predictors in the multiple imputation models. Imputation was carried out in two steps.

Step 1: Imputation of ISAT 2009 pretest scores. Because a student's teacher in 2009/10 was likely to have an impact on the student's ISAT 2010, LRITScoreSP10, and NRRITscoreSP10 scores but not on the student's ISAT 2009 score (when the student was in a prior grade), missing data on ISATReadscale09 were first imputed using the following variables as predictors: [81]

- Student demographic variables
    - Free or reduced-price lunch status
    - Race/ethnicity
    - Disability status
    - English proficiency status

[80] The imputation model used the MAP scale scores in reading (NRRITScoreSP10) and language use (LRITScoreSP10); the average of the two scores (MAPCompositeSP10) was used in the impact models.

[81] Sequential regression multiple imputation allows bounds to be set on the imputed values for a variable. In imputing the ISAT 2009 pretest scores, researchers restricted the imputed values to be between the minimum and maximum possible scores on the on the ISAT 2009 reading test. Imputed values were constrained to be between the minimum and maximum possible scores on the ISAT 2010 reading test.

    - Gender
- Student achievement variables
    - ISATReadscale10
    - LRITScoreFL09 (for MAP students only)
    - LRITScoreWN10 (for MAP students only)
    - NRRITscoreFL09 (for MAP students only)
    - NRRITscoreWN10 (for MAP students only)
    - LRITScoreSP10
    - NRRITscoreSP10
- School demographic variables in 2008/09
    - School mean ISAT 2009 scale score
- District indicator variables

Twenty sets of imputed ISAT 2009 scale scores were generated. Each set was used in the corresponding sets of imputations of student data in Step 2.

Step 2: Imputation of all other student variables. Conditional on the ISAT 2009 scores imputed in Step 1, all other student variables in table D.2 with missing values (ISAT 2010, socioeconomic status, disability status, limited English proficiency status, LRITScoreSP10, and NRRITscoreSP10) were imputed using the following teacher variables in addition to the predictor variables used in Step 1:

- Teacher variables
    - Gender
    - Graduate degree
    - Years of experience teaching reading/English language arts
    - Race/ethnicity
    - Licensure

Steps 1 and 2 resulted in 20 sets of completed data for each grade, each of which was analyzed using PROC MIXED. PROC MIANALYZE was then used to pool the results from the 20 separate analyses in each grade following Rubin's (1987) rules of combining estimates, as described below.

Description of sequential regression multiple imputation method

Let X denote the fully observed variables, and let Y_1, Y_2, ..., Y_k denote k variables with missing values, ordered by the amount of missingness, from least to most.[82] The imputation process for Y_1, Y_2, ..., Y_k proceeds in c rounds. In the first round, Y_1 is regressed on X, and the missing values of Y_1 are imputed. Y_2 is then regressed on X and Y_1 (including the imputed values of Y_1), and the missing values of Y_2 are imputed. Y_3 is then regressed on X, Y_1, and Y_2, and the missing values of Y_3 are imputed. This process continues until Y_k is regressed on X, Y_1, Y_2, ...,Y_{k-1}, and the missing values of Y_k are imputed.

In rounds 2–c, the imputation process carried out in round 1 is repeated, except that in each regression, all variables except the variable to be imputed are included as predictors. Thus, Y_1 is regressed on X, Y_2, Y_3, ..., Y_k, and the missing values of Y_1 are reimputed. Y_2 is then regressed on X, Y_1 , Y_3, ..., Y_k, and the missing values of Y_2 are reimputed. After c rounds, the final imputations of the missing values in Y_1,Y_2, ...,Y_k are used.

IVEware allows the following models to be used:

- A normal linear regression model if the Y-variable is continuous
- A logistic regression model if the Y-variable is binary
- A polytomous or generalized logit regression model if the Y-variable is categorical with more than two categories
- A Poisson log linear model if the Y-variable is a count
- A two-stage model if the Y-variable is mixed (that is, semicontinuous), where logistic regression is used to model the zero/nonzero status for Y, and normal linear regression is used to model the value of Y conditional upon its being nonzero.

In addition, IVEware allows restrictions and bounds to be placed on the variables being imputed.

Combining estimates and standard errors from imputed datasets

This section describes Rubin's (1987) rules for pooling together estimates from m imputed datasets. It is adapted from Schafer and Olsen (1998).

The results from data analyses conducted m times, once for each of the m imputed datasets, are pooled into a single set of results following the rules formulated by Rubin (1987).

[82] The description of sequential regression multiple imputation is from Schenker et al. (2008).

Let $\hat{Q}_j$ denote a point estimate of a parameter of interest (for example, a regression coefficient, Q) obtained from the jth imputed dataset ($j = 1, \cdots, m$), and let U_j be its corresponding variance estimate. The overall point estimate is simply the average of the individual point estimates:

$$\bar{Q} = \frac{1}{m}\sum_{j=1}^{m} \hat{Q}_j.$$

The overall variance of the estimate is obtained by combining two components: the within-imputation variance, which reflects the variability within each dataset,

$$\bar{U} = \frac{1}{m}\sum_{j=1}^{m} U_j,$$

and the between-imputation variance, which measures the variability across imputations,

$$B = \frac{1}{m}\sum_{j=1}^{m} \left(\hat{Q}_j - \bar{Q}\right)^2.$$

The total variance is the sum of these two components (with a correction factor, $1 + 1/m$, that accounts for the finite number of imputations),

$$T = \bar{U} + \left(1 + \frac{1}{m}\right) B.$$

The overall standard error is then the square root of T:

$$SE(\bar{Q}) = \sqrt{T}.$$

Confidence intervals are obtained the usual way:

$$\bar{Q} \pm t_{df} SE(\bar{Q}),$$

where t_{df} is the appropriate percentile of the *t*-distribution with degrees of freedom given by

$$df = (m-1)\left(1 + \frac{m\bar{U}}{(m+1)B}\right)^2.$$

A significance test of the null hypothesis $Q = 0$ is performed by comparing the test statistic $t = \bar{Q}/SE(\bar{Q})$ to the same *t*-distribution.

Appendix E. Response Rates on Surveys, Logs, and Classroom Observations

Teachers in the study were asked to participate in three data collection activities:

- The online teacher survey, which asked for information on basic background variables (for example, highest degree, gender, race/ethnicity) and instructional practices (for example, the extent to which teachers grouped students by ability and the practices that they used for high-achieving versus low-achieving students, and characteristics of their classroom and school)
- Observations of their classroom at three separate times (fall, winter, and spring)
- Teacher Logs describing the instruction given on a designated day to a random sample of four high-achieving and four low-achieving students, completed 10 times throughout the school year

Approximately 93 percent of teachers completed the teacher survey, 95 percent of the scheduled 489 classroom observations (3 observations times 163 teachers) were conducted, and 75 percent of all Teacher Logs were completed (table E.1). Inspection of group differences in response rates by grade also indicated high levels of participation.

Table E.1. Teacher response rates on teacher survey, classroom observations, and teacher logs

Instrument	Grade 4		Grade 5		Total
	MAP	Control	MAP	Control	
Teacher survey					
Eligible[a]	48	33	32	50	163
Completed	44	31	30	47	152
Response rate (percent)	91.7	93.9	93.8	94.0	93.3
Classroom observations					
Fall					
Eligible	48	33	32	50	163
Completed	47	33	28	50	158
Response rate (percent)	97.9	100.0	87.5	100.0	96.9
Winter					
Eligible	48	33	32	50	163
Completed	47	31	25	48	151
Response rate (percent)	97.9	93.9	78.1	96.0	92.6
Spring					
Eligible	48	33	32	50	163
Completed	47	32	30	48	157
Response rate (percent)	97.9	97.0	93.8	96.0	96.3
All observations					

Instrument	Grade 4		Grade 5		Total
	MAP	Control	MAP	Control	
Eligible	144	99	96	150	489
Completed	141	96	83	146	466
Response rate (percent)	97.9	97.0	86.5	97.3	95.3
Teacher logs					
Round 1					
Eligible	384	264	256	400	1304
Completed	205	205	217	309	1012
Response rate (percent)	73.2	77.7	84.8	77.3	77.6
Round 2					
Eligible	384	264	256	400	1304
Completed	317	207	200	340	1064
Response rate (percent)	82.6	78.4	78.1	85.0	81.6
Round 3					
Eligible	384	264	256	400	1304
Completed	324	216	197	327	1064
Response rate (percent)	84.4	81.8	77.0	81.8	81.6
Round 4					
Eligible	384	264	256	400	1304
Completed	292	204	176	331	1003
Response rate (percent)	76.0	71.3	68.8	82.8	76.9
Round 5[a]					
Eligible	382	263	256	396	1297
Completed	292	248	194	341	1075
Response rate (percent)	76.4	94.3	75.8	86.1	82.9
Round 6					
Eligible	381	263	253	393	1290
Completed	245	178	160	290	873
Response rate (percent)	64.3	67.7	63.2	73.8	67.7
Round 7					
Eligible	378	263	252	393	1286
Completed	283	215	198	311	1007
Response rate (percent)	74.9	81.7	78.6	79.1	78.3
Round 8					
Eligible	378	263	250	392	1283
Completed	239	151	179	253	822
Response rate (percent)	63.2	57.4	71.6	64.5	64.1
Round 9					

Instrument	Grade 4		Grade 5		Total
	MAP	Control	MAP	Control	
Eligible	378	263	250	391	1282
Completed	238	194	164	314	910
Response rate (percent)	63.0	73.8	65.6	80.3	71.0
Round 10					
Eligible	378	263	250	391	1282
Completed	274	195	148	307	924
Response rate (percent)	72.5	74.1	59.2	78.5	72.1
All Rounds					
Eligible	3811	2634	2535	3956	12936
Completed	2785	2013	1833	3123	9754
Response rate (percent)	73.0	76.4	72.3	78.9	75.4

a. Data pertain only to active teachers in the Year 2 intent-to-treat analysis.

b. The number of logs to be completed (eligible logs) decreased beginning with Round 5, when students who were chosen for Teacher Log completion left the school and were not replaced with another student.

Source: Authors' analysis based on study records.

Grade 4

For grade 4 teachers, at least 90 percent of both MAP teachers and control teachers completed the teacher survey. Across all rounds, 73 percent of the scheduled logs were completed by MAP teachers and 76 percent by control teachers. Individual Teacher Log completions ranged from 63 percent (Round 9) to 84 percent (Round 3) for the MAP group and were at least 57 percent for the control group. Across the fall, winter, and spring classroom observations, at least 90 percent were carried out for both MAP teachers and control teachers. For each observational period, 90 percent or more of all observations were completed for MAP teachers and control teachers.

Grade 5

Among grade 5 teachers, at least 90 percent of both MAP teachers and control teachers completed and returned the teacher survey. More cooperation by control than MAP teachers was also evident in the Teacher Logs. Across the 10 rounds of Teacher Logs, MAP teachers completed 72 percent and control teachers 79 completed percent (table E.1). In the MAP group, log completion rates ranged from a low of 59 percent (Round 10) to a high of 85 percent (Round 1). The range for the control group was 65 percent (Round 8) to 86 percent (Round 5).

The completion rates across all three observation periods were 87 percent for MAP teachers and at least 90 percent for control teachers. Completion rates for the fall and winter rounds of observations were lower for MAP than for control teachers. For MAP teachers, classroom observations were completed for 88 percent of teachers in the fall and 78 percent in the winter; for control teachers, classroom observations were completed for 100 percent of MAP teachers and at least 90 percent for control teachers. Completion rates in the spring, however, were similar: more than 90 percent for both MAP teachers and control teachers.

Appendix F. MAP Observation Protocol

This appendix describes the MAP observation protocol, reports on the reliability of the observations, and presents the MAP observation protocol form.

Description of protocol

The MAP observation protocol was used to observe teachers' reading instruction on three occasions each in Years 1 and 2 of the study.[83] This instrument is a version of the Center for the Improvement of Early Reading Achievement (CIERA) observation system (Taylor et al. 2003) that was modified and augmented to reflect the MAP instructional components. Baseline observations were conducted in early fall 2009, before the MAP training sessions on using MAP data to differentiate instruction. The second and third observations occurred in January and April 2010.

The observation instrument documented reading and English language arts instruction in intervention and control classrooms. Each classroom observation was completed in one continuous time period that typically lasted 60–90 minutes. The observation instrument recorded information in 10-minute intervals, resulting in about six to nine individual records that described the observed lesson.[84] Observation segments were recorded on a computer using an Access database during a reading/literacy block. Each 10-minute segment consisted of 5 minutes of observing and taking notes of targeted practices, followed by 5 minutes of marking the MAP observation protocol form (given later in this appendix) for that particular segment. At the end of the 10-minute segment, the observer saved the notes and codes recorded for that particular segment and started another 10-minute observation interval. The following information was recorded in these 10-minute intervals:

- Who is teaching
- The student grouping (for example, whole class, small group, pairs, individual)
- The focus of the instructional activities (for example, vocabulary, spelling, fluency, comprehension, writing, speaking, or listening), including teacher activities and student responses to these activities
- The materials used (for example, textbooks, video, computers, board, or chart) during English language arts instruction
- Differentiation focus (differentiating by content, process, or product)

[83] As noted in chapter 3, data from classroom observations conducted in Year 1 were used to develop the observation-based index of differentiated instruction.

[84] The actual number of observation segments depended on the instruction occurring that day. Although most reading and writing lessons lasted between 60 and 90 minutes, some were shorter or longer as a result of scheduled or unscheduled events that occurred in the classroom or school that day. For instance, in a few cases, activities such as fire drills and grade-level or school assemblies occurred, which shortened the scheduled observation time. In a few cases, the reading and writing lesson continued for longer than 90 minutes, although such instances were rare.

- Differentiation type (differentiating the content, process, or product based on academic readiness, learning styles, or interests).

As can be seen from the MAP observation protocol form below, in each segment, observers coded the occurrence or nonoccurrence (that is, binary items) of targeted teacher practices, student practices, and differentiated instructional practices. They also coded, using a four-point Likert scale, their agreement to statements on differentiation of content, process, or product to address student readiness, learning style, or interest.

Protocol development and training

REL Midwest reading content experts developed a set of differentiated instruction items to augment the CIERA observation system. These items were developed to align with the MAP program's definition of differentiated instruction, as well as existing research on the topic (Hall 2002; McTighe and Brown 2006; Tomlinson and McTighe 2005; Tomlinson 2001). Items were designed so that observers would be able to identify variations in a teacher's instructional content, processes, or products across topic areas in vocabulary, spelling, fluency, comprehension, writing, and speaking/listening. After items were developed, two external content experts reviewed the new items and provided feedback, which was incorporated into the protocol. The protocol was then pilot-tested with eight teachers in an Illinois elementary school in spring 2008.[85] Additional items were included and final revisions to the protocol made after the pilot study to better capture specific strategies teachers implemented to differentiate instruction. The final protocol used in Year 2 of the study is given at the end of this appendix.

Although a few observers were research associates at Empirical Education (a REL Midwest subcontractor), most were former teachers and school administrators who lived near one of the five districts in the study. All classroom observers (or raters) participated in an initial two-day training seminar during the fall of 2008. Protocol developers from the REL Midwest facilitated the training event. During the training, raters learned (1) common definitions of key terms in the protocol, (2) how to administer a preobservation survey, and (3) how to consistently code classroom instructional activities. A MAP Classroom Observation Protocol Instructional Guide was developed and provided to all raters to support and reinforce what they learned during the training session. At the end of the two-day session, each rater independently rated five 5-minute video segments of an elementary teacher's reading instruction. A rater's codes were then compared to those of a master rater. Raters who coded a minimum of 80 percent of the items in agreement with the master rater across each of the five video segments were invited to conduct classroom observations in the study schools. One-day follow-up training sessions were conducted before each phase of observations during the study (winter 2009, spring 2009; fall, winter, spring 2009/10) to reassess coders' use of the instrument, address questions from the previous set of observations, and review key concepts in English language arts and differentiated instruction. Weekly conference calls were also conducted during each phase of classroom observations to review the MAP instrument and address questions emerging in the field.

[85] The school in which the protocol was piloted did not participate in the study.

Observation reliability

During each observation phase in Year 2, approximately 18–20 percent of all observations were conducted by a pair of observers who independently coded classroom practices given in the MAP observation protocol. As noted above, these items can be grouped into teacher practice, student practice, differentiated instruction, and agreement items. In all paired observations, we assigned a primary observer and a secondary observer. The data from the primary observer were included in the calculation of indexes and ASRI estimates. Data from the secondary observer was used to calculate inter-rater reliability in the sub-sample of observations that occurred with dual observers. These paired observations allowed the study team to monitor the consistency of the observations over time and routinely address any discrepancies across coders. Thirty-four classrooms in fall 2009, 32 classrooms in winter 2010, and 35 classrooms in spring 2010 were observed by a pair of observers. For these paired observations, there were generally between five and ten observation segments, except for a total of five classrooms that had either less than five segments or more than 10 segments. Table F.1 gives the percentage of agreement between pairs of observers during the fall 2009 and winter and spring 2010 observations for each of the above item groups. It also gives the percentage agreement on the group of (binary) items labeled Main, which includes all items on teacher practice, student practice, differentiated instruction, and instructional modality (e.g., whole group, small group). The percentage of agreement was calculated as the percentage of individual (binary or Likert) items in which a pair of observers coded exactly the same rating. Percentage of agreement was first calculated for each observation segment then averaged over all segments for each teacher, and finally, teacher means were averaged across all teachers. The resulting percentages of agreement were considered high, ranging from 87.6 to 99.6 percent across all item groups and observation periods.

Table F.1. Agreement between pairs of coders of classroom observations, 2009/10

(percent, except where otherwise indicated)

Item type	Number of items	Fall 2009 (*n* = 34)		Winter 2010 (*n* = 32)		Spring 2010 (*n* = 35)	
		Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
Main[a]	145	97.9	0.98	98.5	1.06	98.1	1.18
Teacher practice	29	97.2	2.02	97.8	2.23	97.3	2.19
Student practice	29	96.2	2.38	97.3	2.30	96.6	2.09
Differentiated instruction	81	99.5	0.98	99.6	0.69	99.5	0.83
Agreement	9	87.6	10.01	94.4	9.47	91.6	11.93
Number of observation segments per teacher		7.9 (5–12)	2.03	7.3 (5–9)	1.32	7.5 (5–10)	1.58

Note: *n* represents number of teachers observed by a pair of observers. Mean percentage of agreement was obtained by first averaging the percentage of agreements over all segments for each teacher and then averaging the mean percentage of agreements across all teachers.

a. Main items include teacher practice, student practice, differentiated instruction, and six instructional modality items.

Source: Authors' analysis based on study records.

MAP observation protocol form

I. Background Information

Date: _____ District: _____ School: _____

Grade: ____ Teacher: _____ Coteacher: _____ Observer: _____

Start time: _____ End time: _____

II. Demographics

Teacher gender: ☐ M ☐ F

Other adults in room at time of observation (indicate number):

Resource teacher(s) (e.g., ELL, LD) _____ Paraprofessional(s) _____

Librarian: _____ Volunteer(s) _____

Student teacher(s) _____ No one: _____

Total number of students: _____ Student genders (indicate number): ☐ M ☐ F

Student categories (indicate number): ____Special Education ____ESL/ELL

____Other (gifted, etc. explain):

III. Classroom Environment

Desk arrangement: ☐ Rows ☐ Pairs ☐ Groups ☐ Tables

IV. Materials

Briefly describe the materials used:

By the teacher: _____

By the students: _____

Start Time_____ End Time_____ Page # _____ Classroom_____

Type of Material	Used in Lesson	Definition
Literary text		Narrative text (e.g., novel, short stories, trade books, realistic fiction)
Informational text		Informational text, trade book, reference book (e.g., encyclopedia, etc.), newspapers, magazines, weekly readers
Poetry		All forms of poetry (rhyming, verse, etc.)

Material	Used in Lesson	Definition
Board/Chart		Board, chart, or card is being used (e.g. blackboard, pocket chart, hanging chart, flashcards)
Computer		
Listening center		Students are listening to books on audiotape or CD
Newspapers, magazines		
Text book		Include science, social studies or other content areas
Text sets		All materials will be about one topic. Students maybe reading different books at different levels on one topic. One group of students maybe reading about one title while another reads a different book on the same topic.
Other books		
Paper and pen/pencil		
Worksheet		Worksheet, workbook page, sheet of paper, individual white boards for one-word or one-sentence answers.
Other		Something other than the above is being used

V. Observation

Instructional Focus: ☐ Reading ☐ Writing

Instructional Modality: ☐ Whole Group ☐ Small Group ☐ Pairs ☐ Center

☐ Independent ☐ Other: _____

Topic	Teacher Models or Demonstrates	Students Practice	Differentiated Instruction
Vocabulary	Selects words that are central to understanding the text	Understanding word meaning	☐ Content ☐ Varied vocabulary ☐ Other ☐ Process ☐ Word wall ☐ Varied graphic organizers ☐ Other ☐ Product ☐ Tiered assignments ☐ Other
	Introduces words in meaningful contexts	Understanding and extending word meaning	
	Uses vocabulary strategies to gain meaning from text	Using context, suffixes, affixes, and root words to determine meaning of unknown words	
	Connects new vocabulary to prior knowledge	Using prior knowledge to activate and build meaning	
	Uses new vocabulary in written responses	Increasing word knowledge	
Spelling	Says the words	Spelling the words orally or in writing	☐ Content ☐ Personalized lists ☐ Other ☐ Process ☐ Practice options ☐ Multiple opportunities ☐ Other ☐ Product ☐ Alternate assessments ☐ Other
Fluency	Directs students to read aloud a passage that has been read silently	Reading connected text aloud	☐ Content ☐ Varied readabilities ☐ Varied texts ☐ Audio support ☐ Other
	Directs students to read with a partner	Listening and responding to others reading	

Topic	Teacher Models or Demonstrates	Students Practice	Differentiated Instruction
	Times students as they practice fluency	Improving oral reading rate, accuracy, and prosody	☐ Process ☐ Teacher modeling ☐ Prereading ☐ Partner reading ☐ Multiple opportunities ☐ Simultaneous reading ☐ Goal setting ☐ Self-monitoring (taping) ☐ Other ☐ Product ☐ Multiple audiences ☐ Individual progress charts ☐ Other
Comprehension	Conveys goals of the lesson	Understanding the purpose for reading	☐ Content ☐ Varied readabilities ☐ Multiple texts ☐ Varied audio or visual support ☐ Student choice ☐ Other ☐ Process ☐ Teacher modeling/think aloud ☐ Questioning ☐ Varied graphic organizer ☐ Peer support ☐ Learning logs ☐ Tiered activities ☐ Guided reading ☐ Literature circles ☐ Varied time allotments ☐ Marking text ☐ Other ☐ Product
	Uses strategies to access prior knowledge of topic	Making connections between what they already know and topic at hand	
	Shows anticipation of events in narrative text	Making predictions based on their own knowledge and insight	
	Uses metacognitive strategies to monitor and gain meaning from text	Using prior knowledge, questioning, visualizing, summarizing, and inferring to construct meaning beyond literal recall of text	
	Uses fix-up strategies when comprehension breaks down	Rereading, using context, using pictures, and asking for help in order to monitor comprehension	

Topic	Teacher Models or Demonstrates	Students Practice	Differentiated Instruction
	Uses literary devices to develop understanding of texts	Recognizing literary devices (e.g., figurative language, foreshadowing, alliteration) in text in order to improve understanding	☐ Tiered assignments ☐ Project options ☐ Independent study ☐ Other
	Uses text structure to understand content	Using heading, chapters, and text organization to improve comprehension	
	Uses strategies for organizing information from text	Using graphic organizers	
	Uses strategies for active reading	Taking notes, using notation systems, and marking text while reading	
	Directs students to write in response to what was read to them or what they read	Writing in response to what was read	
Writing	Identifies purpose and type of writing	Setting purposes for writing	☐ Content ☐ Multiple prompts ☐ Student choice ☐ Other ☐ Process ☐ Teacher modeling ☐ Varied graphic organizer ☐ Varied prewriting strategies ☐ Peer support ☐ Tiered activities ☐ Multiple drafts ☐ Revision opportunity ☐ Peer feedback ☐ Writing conference ☐ Writing stations
	Uses prewriting tools	Organizing writing using maps, webs, lists, or outlines	
	Shows awareness of audience when developing a draft	Writing a first draft	
	Shows stages of the writing process	Drafting, revising, editing, or publishing	
	Shows explicit skills and strategies that improve writing	Practicing developing effective leads and endings, support and elaboration of ideas, use of figurative language, and transitions to improve writing	

Topic	Teacher Models or Demonstrates	Students Practice	Differentiated Instruction
	Uses tools for editing and proofreading	Editing writing for grammar, mechanics, formatting, spelling, and word choice	☐ Supportive technology (word processing, voice recognition) ☐ Other ☐ Product ☐ Graduated rubric ☐ Project options ☐ Portfolio ☐ Independent study ☐ Other
	Holds informal or scheduled conferences focused on assisting writers	Improving writing through feedback	
	Shares published writing for the purpose of gathering feedback used in revising writing	Sharing writing with others and providing effective feedback	
	Uses rubrics to evaluate and improve writing	Discussing or using rubrics to guide writing	
Speaking or Listening	Shows behaviors of engaged discussion (asking and answering questions, integrating and extending responses of others)	Holding engaging discussions	☐ Content ☐ Multiple prompts ☐ Other ☐ Process ☐ Teacher modeling ☐ Peer support ☐ Multiple opportunities ☐ Scaffolded note-taking strategies ☐ Practice opportunities ☐ Other ☐ Product ☐ Graduated rubric ☐ Project options ☐ Varied modes of presentation ☐ Other

Write specific notes on differentiated instructional practices observed:

Please indicate how much you agree or disagree with each of these statements:	**Not Present**	**Slightly Integrated**	**Partially Integrated**	**Fully Integrated**
The content is differentiated for student readiness.	1	2	3	4
The content is differentiated for student learning style.	1	2	3	4
The content is differentiated for student interest.	1	2	3	4
The process is differentiated for student readiness.	1	2	3	4
The process is differentiated for student learning style.	1	2	3	4
The process is differentiated for student interest.	1	2	3	4
The product is differentiated for student readiness.	1	2	3	4
The product is differentiated for student learning style.	1	2	3	4
The product is differentiated for student interest.	1	2	3	4

Appendix G. MAP Instructional Logs

This appendix presents the MAP instructional log form. Teachers were trained to complete the instructional logs during fall 2008 and fall 2009. In the four smaller districts, a member of the study team visited teachers in each school and provided a 45 minute training session for teachers to complete the logs either before or after school or during a common planning time during the school day. In the largest district, two training sessions (one in the morning and one in the afternoon on two different days) were offered to all teachers. The training was held in an auditorium in one of the district's high schools. In addition, a member of the study team was identified as the key point of contact for teachers if they had questions or needed technical assistance to complete the logs. This team member's contact information was distributed to teachers during the training.

Identification Information

*****Indicates a required response

**1. ***Please enter your first and last name

1a. Please select the date you observed this student this week

Month
Day

Language Arts Log

2. *How much total time did the target student spend on language arts today? Please include all language arts instruction the target student received including routine times such as morning board work, even if the instruction took place in another room or by another teacher. *Enter the number of minutes.*

Minutes

3. *Of the language arts time recorded in Question 2, how much time were you either the teacher or an observer of the teaching? *Enter the number of minutes.*

Minutes

4. *Did you enter 0 minutes in either Question 2 or Question 3?

D Yes
D No

5. Please mark the reason(s) why you recorded 0 minutes in Question 2 or 3. *(Check all the boxes that apply.)*

D Target student was absent
D I was absent
D School was not in session (e.g., vacation period)
D There was a field trip, assembly, visitor, or other special event

- D Target student participated in standardized testing/test preparation
- D Target student received pull out instruction
- D Other

Comprehension

6. *To what extent was comprehension a focus of your work with the target student in reading/language arts today?

- D A major focus
- D A minor focus
- D Touched on briefly
- D Not taught today

7. Was the work in comprehension in … *(Check all the boxes that apply.)*

- D Listening Comprehension
- D Reading Comprehension

8. What areas of comprehension did the target student work on today? (*For each area you choose below, please indicate whether it was a focus of instruction or was touched on briefly. Check all the boxes that apply.)*

Activating prior knowledge or making personal connections to text	D	A focus of instruction
	D	Touched on briefly
Making predictions, previewing, or surveying	D	A focus of instruction
	D	Touched on briefly
Vocabulary-comprehension relationships	D	A focus of instruction
	D	Touched on briefly
Students generating their own questions	D	A focus of instruction
	D	Touched on briefly
Reading for pleasure or information	D	A focus of instruction
	D	Touched on briefly
Self-monitoring for meaning	D	A focus of instruction
	D	Touched on briefly
Using visualization or imagery	D	A focus of instruction
	D	Touched on briefly
Using charts, graphs, figures, tables, or other visual aids in the text	D	A focus of instruction
	D	Touched on briefly
Using concept maps, story maps, or text structure frames	D	A focus of instruction
	D	Touched on briefly
Answering questions that have answers directly stated in the text	D	A focus of instruction
	D	Touched on briefly
Answering questions that require inferences	D	A focus of instruction
	D	Touched on briefly

Explaining how to find answers or information	D	A focus of instruction
	D	Touched on briefly
Sequencing information or events	D	A focus of instruction
	D	Touched on briefly
Identifying story structure	D	A focus of instruction
	D	Touched on briefly
Practicing other skills such as identifying similes or understanding referents	D	A focus of instruction
	D	Touched on briefly
Comparing and/or contrasting information or texts	D	A focus of instruction
	D	Touched on briefly
Summarizing important details	D	A focus of instruction
	D	Touched on briefly
Analyzing and evaluating text	D	A focus of instruction
	D	Touched on briefly
Examining literary techniques or author's style	D	A focus of instruction
	D	Touched on briefly
Written literature extension project	D	A focus of instruction
	D	Touched on briefly
Nonwritten literature extension project (e.g., puppet show, play, shadow box, book talk)	D	A focus of instruction
	D	Touched on briefly

9. Did the materials used by the target student in work on comprehension include any of the following? *(Check all the boxes that apply.)*

- D Informational text
- D Narrative text with controlled vocabulary (sight words and/or words easily sounded out)
- D Narrative text with patterned or predictable language
- D Literature-based or thematic text: short selection
- D Literature-based or thematic text: chapter book

10. In which of the following ways did the target student demonstrate comprehension? *(Check all the boxes that apply.)*

- D Answered brief oral questions
- D Discussed text with peers
- D Did a think-aloud or explained how they applied a skill or strategy
- D Generated questions about text
- D Answered multiple-choice questions
- D Completed sentences filling in the blanks
- D Worked on concept maps, story maps, or text structure frames
- D Wrote brief answers to questions
- D Wrote extensive answers to questions
- D Worked on a literature extension project

11. Did your instruction in comprehension today include any of the following? *(Check all the boxes that apply.)*

- D I demonstrated or explained a skill (e.g., how to determine the main idea, how to make an inference)
- D I demonstrated or explained how to use a reading strategy (e.g., previewing, generating questions about text)
- D I explained why or when to use a reading strategy
- D I helped students to practice a skill or strategy
- D I administered a comprehension test

Writing

12. *To what extent was writing a focus of your work with the target student in reading/language arts today?

- D A major focus
- D A minor focus
- D Touched on briefly
- D Not taught today

13. What areas of writing did the target student work on today? (*For each area you choose below, please indicate whether it was a focus of instruction or was touched on briefly. Check all the boxes that apply.)*

W**riting forms or genres (e.g., letter, drama, editorial, Haiku)**	D	A focus of instruction
	D	Touched on briefly
Writing practice	D	A focus of instruction
	D	Touched on briefly
Revision of writing- elaboration	D	A focus of instruction
	D	Touched on briefly
Revision of writing- refining or reorganizing	D	A focus of instruction
	D	Touched on briefly
Editing capitals, punctuation, or spelling	D	A focus of instruction
	D	Touched on briefly
Editing word use, grammar, or syntax	D	A focus of instruction
	D	Touched on briefly
Sharing writing with others (e.g., author's chair, share-pair, performances)	D	A focus of instruction
	D	Touched on briefly

14. Did the target student's writing consist of … *(Check all the boxes that apply.)*

- D Letter strings or words (with or without illustration)
- D Separate sentence(s) (with or without illustration)
- D Separate paragraph(s)
- D Connected paragraphs

15. Did your instruction in writing include any of the following? *(Check all the boxes that apply.)*

- D I demonstrated or did a think-aloud using my own writing
- D I explained how to write, organize ideas, revise or edit using student writing
- D I explained how to write, organize ideas, revise or edit using a published author's writing
- D I took dictation from the student
- D I led the student and his/her peers in a group composition
- D I commented on what the student wrote (not how)
- D I described what the student did well in his/her writing
- D I commented on how the student could improve his/her writing
- D I provided a writing or proofreading guide

Word Analysis

16. *To what extent was word analysis a focus of your work with the target student in reading/language arts today?

- D A major focus
- D A minor focus
- D Touched on briefly
- D Not taught today

17. What areas of word analysis did the target student work on today? (*For each area you choose below, please indicate whether it was a focus of instruction or was touched on briefly. Check all the boxes that apply.)*

Letter-sound relationships	D	A focus of instruction
	D	Touched on briefly
Sound segmenting: Counting the number of sounds in words	D	A focus of instruction
	D	Touched on briefly
Sound segmenting: Sound spelling/invented spelling/developmental spelling	D	A focus of instruction
	D	Touched on briefly
Sound segmenting: Segmenting a part of the word (for example, “many” without “m” is “any,” or “upstairs” without “stairs” is “up”)	D	A focus of instruction
	D	Touched on briefly
Sound segmenting: other segmenting tasks	D	A focus of instruction
	D	Touched on briefly
Sound blending: Blending initial sound with a rhyming word (onset-rime)	D	A focus of instruction
	D	Touched on briefly
Sound blending: Blending individual phonemes (sounds) into real words	D	A focus of instruction
	D	Touched on briefly
Sound blending: Blending phonemes (sounds) into nonsense words	D	A focus of instruction
	D	Touched on briefly
Sounding blending: Blending syllables	D	A focus of instruction
	D	Touched on briefly
Sound blending: Other blending tasks	D	A focus of instruction

Word recognition, sight words	D	Touched on briefly
	D	A focus of instruction
	D	Touched on briefly
Structural analysis, examining word families, prefixes, suffixes, contractions, etc.	D	A focus of instruction
	D	Touched on briefly
Use of context, picture, and/or sentence meaning and structure into read words	D	A focus of instruction
	D	Touched on briefly
Use of phonics-based or letter-sound relationships to read words in sentences or stories	D	A focus of instruction
	D	Touched on briefly

18. Did the materials used by the target student in work on word analysis contain any of the following? *(Check all the boxes that apply.)*

- D Sounds only
- D Pictures or objects to identify letters, words
- D Isolated words and letters
- D Individual sentences
- D Connected text (for example, stories, articles, poems, etc.) with controlled vocabulary (sight words and/or words easily sounded out)
- D Connected text (for example, stories, articles, poems, etc.) with patterned or predictable language
- D Connected text (for example, stories, articles, poems, etc.) that is literature-based or thematic

19. What did you do when a student got stuck or made errors in word analysis? *(Check all the boxes that apply.)*

- D I corrected the student's errors or modeled the correct answer
- D I told the student to try again
- D I prompted the student to use the context (other words in sentence, pictures, what they already know) to read the word
- D I gave oral cues—sounding out parts of the word for them
- D I ignored the error and waited for the student to self-correct

20. Did your instruction in word analysis include any of the following? *(Check all the boxes that apply.)*

- D I listened to the target student read
- D I took running records or conducted a miscue analysis
- D I administered a word analysis test

21. *To what extent were the following topics a focus of your work with the target student in reading/language arts today?

Concepts of print	D	A major focus
	D	A minor focus
	D	Touched on briefly
	D	Not taught today

Reading fluency	D	A major focus
	D	A minor focus
	D	Touched on briefly
	D	Not taught today
Vocabulary	D	A major focus
	D	A minor focus
	D	Touched on briefly
	D	Not taught today
Grammar	D	A major focus
	D	A minor focus
	D	Touched on briefly
	D	Not taught today
Spelling	D	A major focus
	D	A minor focus
	D	Touched on briefly
	D	Not taught today
Research strategies	D	A major focus
	D	A minor focus
	D	Touched on briefly
	D	Not taught today

22. During today's reading/language arts lesson, did the target student work: *(Check all the boxes that apply.)*

- D Individually
- D In a group with students at similar levels of achievement
- D In a group with students who had similar interests or learning styles
- D In a cooperative learning group
- D In a pair with another student
- D With the whole class
- D Other (describe)_____

23. Did the target student use any of the following reading/language arts materials today? *(Check all the boxes that apply.)*

- D Materials that are part of the curriculum materials adopted by my school
- D Materials that I found in other books or on websites that provide resources for teachers
- D Materials that were given to me by another teacher/colleague
- D Materials that I developed myself for students
- D Materials that the student found on his/her own (e.g., on the Web or in the library)
- D Other (describe) _____

24. Please share any comments or questions.

Appendix H. MAP Teacher Survey for MAP Teachers

Two teacher surveys were developed for this study: one for teachers assigned to the treatment condition and one for teachers assigned to the control condition. The teacher survey shown in this appendix was administered to teachers assigned to the treatment condition. It includes questions that ask specifically about the MAP training and testing program. Teachers assigned to the control condition received the same survey, with the following MAP-specific questions excluded:

Q8a: Items 3, 5, 6, 7,
Q10b: Items 2, 4, 5, 6, 7,
Q12a: Items 2, 4, 5, 6, 7,
Q13b: Items 2, 4, 5, 6, 7,
Q15a: Items 2, 4, 5, 6, 7,
Q36, Q37, Q38, and Q39.

OMB Control No: 1850-0850 Expiration: 01/31/2011

Survey of Teachers on Instructional Practices: The 2008–09 School Year

Conducted by Learning Point Associates and Vanderbilt University for the Institute for Educational Sciences, U.S. Department of Education, April 2009.

Thank you for your participation in this study. We know that your time is valuable, and we greatly appreciate your willingness to complete this questionnaire. We realize that the survey asks a substantial number of questions, but answering the questions is important. Otherwise, we will not have a good sense of the topics that you cover in reading/language arts, the instructional practices that you use, and the environment of your school.

Your responses are voluntary and confidential. If there is a question that you do not wish to answer, simply skip it. We hope, however, that you will answer as many questions as possible.

All information that you provide will be reported only in a form that does not personally identify you or your school.

Public reporting burden for this collection of information is estimated to average 60 minute per response, including the time for reviewing instructions, gathering any needed data, and completing and reviewing the questionnaire. An agency may not conduct or sponsor, and a person is not required to respond to, a collection of information unless it displays a currently valid OMB control number. Send comments regarding this burden estimate or any other aspect of this questionnaire, including suggestions for reducing this burden to: Angela Arrington, IES Clearance Officer, 202-245-6409.

1a. Enter your first and last name here

_____

1b. HiddenID

1c. HiddenName

Remember that you can click "Save" at any time that you want to exit the survey and finish it at a later time. Your responses up to this point will be saved, and when you log onto the survey again, you will immediately go to the screen of the last question that you answered.

Section A. Your Perspective on the School

Teachers, students and the overall school environment often are not the same in different schools. The questions in this section ask for your perceptions about the school in which you are teaching.

1. To what extent do you agree or disagree with each of the following statements about your school?

Agreement or Disagreement	Completely Disagree	Mostly Disagree	Mostly Agree	Completely Agree
Teachers are encouraged to take risks in order to improve their teaching				
Teachers are expected to continually learn and seek out new ideas in this school				
Teachers are encouraged to experiment in their classrooms in this school				

2. To what extent do you agree or disagree with each of the following statements?

Agreement or Disagreement	Completely Disagree	Mostly Disagree	Mostly Agree	Completely Agree
Policies about how I should teach are often contradictory				
I often have difficulty choosing what to do in my classroom out of all the options I hear about				
Out of all the information about teaching I receive, I am often unsure about how to prioritize things				
Overall, the instructional policies I am supposed to follow seem inconsistent				

3. To what extent do you agree or disagree with each of the following statements?

Agreement or Disagreement	Completely Disagree	Mostly Disagree	Mostly Agree	Completely Agree
I have detailed knowledge of the content covered and instructional methods used by other teachers at this school				
When I begin working with a new group of students, I have detailed knowledge of what those students learned previously				
It's easy for other teachers in this school to know what students learned in my class				
I frequently plan and coordinate instruction with my students' other teachers				
In this school, teachers who work with students at the same achievement level use similar methods and cover the same content				

4. Do you teach reading/language arts as part of your assignment?
Yes
No

Characteristics of Your Reading/Language Arts Class

The questions in this section ask about the amount of time that you teach reading/language arts on a typical day and your views about the resources that you have and the school's perspective on teaching reading/language arts.

5. On a typical day, approximately how many minutes do you teach reading/language arts to the students in your class?
Number of minutes _____

6. To what extent do you agree or disagree with each of the following statements?

Agreement or Disagreement	Completely Disagree	Mostly Disagree	Mostly Agree	Completely Agree
I have adequate time during the regular school week to work with my peers on reading/language arts curriculum or instruction				
I have adequate curriculum materials available for teaching reading/language arts				
I have adequate equipment (e.g., computers) for teaching reading/language arts				

Agreement or Disagreement	Completely Disagree	Mostly Disagree	Mostly Agree	Completely Agree
In this school, teachers can develop their own "teaching style" in reading/language arts				
In this school, teachers can pick and choose the reading/language arts curricular content				
I am asked to implement instructional strategies that conflict with my best professional judgment about teaching reading/language arts				

Section B. Topics Covered and Instructional Practices Used in Your Reading/Language Arts Class

Different reading/language arts classes may pay different amounts of attention to topics, depending on district/state standards, student needs and maturity, students interests and learning styles, available resources, and other factors. Instructional practices may differ as well for the same reasons. The questions in this section ask about the topics that you covered in your reading/language arts class as well as types of instructional practices that you may have used (if at all) during the course of this school year.

7a. Think back over all the time that students spent in your reading/language arts class over the course of the school year.
Approximately what **percentage** of time was devoted to teaching … ? *(If no time was spent on a topic, type "0".)*

Comprehension _____
Writing _____
Word analysis (e.g., decoding, word families, context cues, or sight words) _____
Reading fluency (e.g., repeated reading or guided oral reading) _____
Vocabulary _____
Grammar _____
Spelling _____
Other _____

Total instructional time spent in Language Arts: 100%

7b. If you filled in a nonzero response for "other" above, briefly describe.

8a. How much did you rely on each of the following sources of information for preparing lessons for your reading class during the past school year?

Amount of Reliance	Not at All	Only a Little	Some	A Great Deal
Curricular frameworks or standards documents				
Teachers' materials that are part of the curriculum materials adopted by this school				
MAP training sessions or other interactions with MAP trainers				
Suggestions from other instructional leaders at my school (e.g., reading specialists or coordinators of other reading programs)				

Amount of Reliance	Not at All	Only a Little	Some	A Great Deal
MAP data reports (for example, the class breakdown by goal report, individual student progress report, student goal setting reports)				
MAP DesCartes or Lexile information				
Other information such as instructional materials on the MAP website or reports site				
Other websites that provide resources for teachers				
Other sources (describe below)				

8b. If you relied on other sources, briefly describe.

_____

9. When teaching your reading class, how often did you … ?

Frequency	Never	Less Than 1 Time Per Month	About 1 Time Per Month	2–3 Times Per Month	About 1 Day Per Week	More Than 1 Day Per Week but Not Every Day	Every Day
Use whole class grouping (i.e., all students were taught the same thing at the same time)							
Assign students to pairs (e.g., partnering or peer-assisted learning)							
Use individualized instruction (e.g., students worked individually on learning assignments specifically tailored to their own achievement levels, interests, or learning styles)							

10. Did you group students by ability or achievement for reading/language arts instruction?
- Yes
- No

10a. When teaching your reading class, how often did you group students by ability or achievement?
- Less than once a month
- About once per month
- 2–3 times per month
- About 1 day per week
- More than 1 day per week but not every day
- Every day

10b. When you first grouped students by ability or achievement for reading/language arts instruction, how much did you rely on each of the following sources of information for assigning students to specific groups?

Amount of Reliance	Not at All	Only a Little	Some	A Great Deal
Students' performance on state or local achievement tests				
Students' performance on MAP tests				
Students' performance on other types of assessments (e.g., tests administered by the school psychologist or reading specialist)				
MAP training sessions or other interactions with MAP trainers				
MAP data reports (for example, the class breakdown by goal report, individual student progress report, student goal setting reports)				
MAP DesCartes or Lexile information				
Other information such as instructional materials on the MAP website or reports site				
Students' performance on homework or in-class assignments, quizzes, and exams				
Your own observations of or discussions with individual students				
Other (describe below)				

10c. If you relied on other sources, briefly describe.

_____

11. Once you assigned students to ability or achievement groups, approximately how often did you change the composition of these groups?

Rarely or never
A few times a month
Once a month
About once every two months (4–5 times during the school year)
About once every three or four months (2–3 times during the school year)

12a. In making changes to which students were in these groups, how much did you rely on … ?

Amount of Reliance	Not at All	Only a Little	Some	A Great Deal
Students' performance on state or local achievement tests				
Students' performance on MAP tests				
Students' performance on other types of assessments (e.g., tests administered by the school psychologist or reading specialist)				
MAP training sessions or other interactions with MAP trainers				
MAP data reports (for example, the class breakdown by goal report, individual student progress report, student goal setting reports)				
MAP DesCartes or Lexile information				

Amount of Reliance	Not at All	Only a Little	Some	A Great Deal
Other information such as instructional materials on the MAP website or reports site				
Students' performance on homework or in-class assignments, quizzes, and exams				
Your own observations of or discussions with individual students				
Other (describe below)				

12b. If you relied on other sources, briefly describe.

13. Did you group students by their interests or learning styles or assign students to cooperative learning groups for reading/language arts instruction?

Yes
No

13a. When teaching your reading class, how often did you group students by interests or learning styles or established cooperative learning groups?

Less than once a month
About once per month
2–3 times per month
About 1 day per week
More than 1 day per week but not every day
Every day

13b. When you first grouped students by their interests or learning styles or established cooperative learning groups, how much did you rely on each of the following sources of information for assigning students to specific groups?

Amount of Reliance	Not at All	Only a Little	Some	A Great Deal
Students' performance on state or local achievement tests				
Students' performance on MAP tests				
Students' performance on other types of assessments (e.g., tests administered by the school psychologist or reading specialist)				
MAP training sessions or other interactions with MAP trainers				
MAP data reports (for example, the class breakdown by goal report, individual student progress report, student goal setting reports)				
MAP DesCartes or Lexile information				
Other information such as instructional materials on the MAP website or reports site				
Students' performance on homework or in-class assignments, quizzes, and exams				
Your own observations of or discussions with individual students				
Other sources (describe below)				

13c. If you relied on other sources, briefly describe.

14. Once you assigned students to ability or achievement groups, approximately how often did you change the composition of these groups?

Rarely or never
A few times a month
Once a month
About once every two months (4–5 times during the school year)
About once every three or four months (2–3 times during the school year)

15a. In making changes to which students were in these groups, how much did you rely on … ?

Amount of Reliance	**Not at All**	**Only a Little**	**Some**	**A Great Deal**
Students' performance on state or local achievement tests				
Students' performance on MAP tests				
Students' performance on other types of assessments (e.g., tests administered by the school psychologist or reading specialist)				
MAP training sessions or other interactions with MAP trainers				
MAP data reports (for example, the class breakdown by goal report, individual student progress report, student goal setting reports)				
MAP DesCartes or Lexile information				
Other information such as instructional materials on the MAP website or reports site				
Students' performance on homework or in-class assignments, quizzes, and exams				
Your own observations of or discussions with individual students				
Other (describe below)				

15b. If you relied on other sources, briefly describe.

Section C. Topics Covered and Instructional Practices Used in Teaching Reading/Language Arts to High-Achieving Students

16a. Approximately how many students in your reading/language arts class are high-achieving (they "Exceeded Standards," according to the definition used by the ISAT)?

Number of high-achieving students _____

16b. Did you answer “0” above?

Yes
No

17. When you further grouped these high-achieving students for reading/language arts instruction according to their learning styles, interests, or other characteristics, approximately how large were these individual groups?

I did not group these high-achieving students
I typically assigned them to pairs
I typically created groups of 3 students
I typically created groups of 4 students
I typically created groups of 5 or more students

18. How often were the following topics a primary focus of instruction for these high-achieving students?

Frequency	Never	Less Than Once a Month	1–3 Times Per Month	1–2 Times Per Week	3–4 Times Per Week	Every Day
Word analysis (e.g., decoding, word families, context cues, and sight words)						
Reading fluency (e.g., repeated reading and guided oral reading)						
Listening comprehension						
Reading comprehension						
Grammar						
Spelling						
Written composition (e.g., writing sentences, paragraphs, and stories)						

19. Looking back over the school year and thinking about how you taught reading/language arts to these high-achieving students, how often did you … ?

Frequency	Never	Less Than Once a Month	1–3 Times Per Month	1–2 Times Per Week	3–4 Times Per Week	Every Day
Use basic skills worksheets						
Use enrichment worksheets						
Assign reading of more advanced level work						
Assign reports						
Assign projects or other work requiring extended time for students to complete						
Make time available for students to pursue self-selected interests						
Use pretests to determine if students had mastered the material covered in a particular unit or content area						

Frequency	Never	Less Than Once a Month	1–3 Times Per Month	1–2 Times Per Week	3–4 Times Per Week	Every Day
Repeat instruction on the coverage of more difficult concepts for some students						
Encourage students to move around the classroom to work in various locations						
Allow students to leave the classroom to work in another location, such as the school library or media center						
Use learning centers to reinforce basic skills						
Use enrichment centers						
Teach thinking skills such as critical thinking or creative problem-solving						
Use contracts or management plans to help students organize their independent study projects						
Establish interest groups which enable students to pursue individual or small group interests						
Consider students' opinions in allocating time for various subjects within your classroom						
Provide opportunities for students to use programmed or self-instructional materials at their own pace						
Use computers						
Encourage student participation in discussions						

20. About how many minutes, on average, did you expect a high-achieving student to spend on normal homework assignments in reading/language arts outside of class?

I did not assign homework in comprehension
Less than 15 minutes
15–30 minutes
31–60 minutes
61–90 minutes
More than 90 minutes

21. How often did you usually assign these high-achieving students reading/language arts homework to be completed outside class?

Less than once per week
Once or twice per week
3–4 per week
Every day

In teaching reading/language arts, comprehension is one specific topic for high-achieving students. The next series of questions ask about the topics in comprehension that you covered, how you had students demonstrate competence and the types of texts that you used.

22. How often were the following comprehension topics a primary focus of instruction for these high-achieving students this year?

Frequency	Never	Less Than Once a Month	1–3 Times Per Month	1–2 Times Per Week	3–4 Times Per Week	Every Day
Activating prior knowledge or making personal connections to text						
Making predictions, previewing, or surveying text						
Students generating their own questions						
Summarizing important or critical details						
Examining literary techniques						
Identifying the author's purposes						
Using concept maps, story maps, or text structure frames						
Answering questions that have answers directly stated in the text						
Answering questions that require inferences						

23. This year, how often did the high-achieving students in your reading class demonstrate comprehension in the following ways?

Frequency	Never	Less Than Once a Month	1–3 Times Per Month	1–2 Times Per Week	3–4 Times Per Week	Every Day
Wrote brief answers to questions						
Wrote extensive answers to questions						
Did a think-aloud or explained how they applied a skill or strategy						

Worked on a written literature extension project

24. This year, how often did the high-achieving students in your reading class work on the following areas in written comprehension?

Frequency	Never	Less Than Once a Month	1–3 Times Per Month	1–2 Times Per Week	3–4 Times Per Week	Every Day
Editing the capitalization, punctuation, or spelling of their own writing						
Editing the word use, grammar, or syntax of their own writing						
Revising their writing by elaborating and extending what they wrote						
Revising their writing by recognizing or refining what they wrote						

25. This year, how often did the high-achieving students in your reading class work on comprehension using … ?

Frequency	Never	Less Than Once a Month	1–3 Times Per Month	1–2 Times Per Week	3–4 Times Per Week	Every Day
Informational text						
Narrative text with patterned or predictable language						
Narrative text with controlled vocabulary (sight words and/or easily sounded out words)						
Short narrative text without any attempt to control vocabulary (literature-based or thematic)						
Chapter book						

Section D. Topics Covered and Instructional Practices Used in Teaching Reading/Language Arts to Low-Achieving Students

26a. Approximately how many students in your reading/language arts class are low-achieving (they are identified as “Academic Warning” or “below Standards,” according to the ISAT)?
Number of low-achieving students _____

26b. Did you answer “0” above?
Yes
No

27. When you further grouped these low-achieving students for reading/language arts instruction according to their ability, approximately how large were these individual groups?

I did not group these low-achieving students
I typically assigned them to pairs
I typically created groups of 3 students
I typically created groups of 4 students
I typically created groups of 5 or more students

28. How often were the following topics a primary focus of instruction for these low-achieving students?

Frequency	Never	Less Than Once a Month	1–3 Times Per Month	1–2 Times Per Week	3–4 Times Per Week	Every Day
Word analysis (e.g., decoding, word families, context cues, and sight words)						
Reading fluency (e.g., repeated reading and guided oral reading)						
Listening comprehension						
Reading comprehension						
Grammar						
Spelling						
Written composition (e.g., writing sentences, paragraphs, and stories)						

29. Looking back over the school year and thinking about how you taught reading/language arts to these low-achieving students, how often did you … ?

Frequency	Never	Less Than Once a Month	1–3 Times Per Month	1–2 Times Per Week	3–4 Times Per Week	Every Day
Use basic skills worksheets						
Use enrichment worksheets						
Assign reading of more advanced level work						
Assign reports						
Assign projects or other work requiring extended time for students to complete						
Make time available for students to pursue self-selected interest						
Use pretests to determine if students had mastered the material covered in a particular unit or content area						
Repeat instruction on the coverage of more difficult concepts for some students						

Frequency	Never	Less Than Once a Month	1–3 Times Per Month	1–2 Times Per Week	3–4 Times Per Week	Every Day
Encourage students to move around the classroom to work in various locations						
Allow students to leave the classroom to work in another location, such as the school library or media center						
Use learning centers to reinforce basic skills						
Use enrichment centers						
Teach thinking skills such as critical thinking or creative problem-solving						
Use contracts or management plans to help students organize their independent study projects						
Establish interest groups which enable students to pursue individual or small group interests						
Consider students' opinions in allocating time for various subjects within your classroom						
Provide opportunities for students to use programmed or self-instructional materials at their own pace						
Use computers						
Encourage student participation in discussions						

30. About how many minutes, on average, did you expect a low-achieving student to spend on normal homework assignment in reading/language arts outside of class?

I did not assign homework in reading/language arts
Less than 15 minutes
15–30 minutes
31–60 minutes
61–90 minutes
More than 90 minutes

31. How often did you usually assign these low-achieving students reading/language arts homework to be completed outside class?

Less than once per week
Once or twice per week
3–4 per week
Every day

In teaching reading/language arts, comprehension is one specific topic for low-achieving students. The next series of questions ask about the topics in comprehension that you covered, how you had students demonstrate competence and the types of texts that you used.

32. How often were the following comprehension topics a primary focus of instruction for these low-achieving students this year?

Frequency	Never	Less Than Once a Month	1–3 Times Per Month	1–2 Times Per Week	3–4 Times Per Week	Every Day
Activating prior knowledge or making personal connections to text						
Making predictions, previewing, or surveying text						
Students generating their own questions						
Summarizing important or critical details						
Examining literary techniques						
Identifying the author's purposes						
Using concept maps, story maps, or text structure frames						
Answering questions that have answers directly stated in the text						
Answering questions that require inferences						

33. This year, how often did the low-achieving students in your reading class demonstrate comprehension in the following ways?

Frequency	Never	Less Than Once a Month	1–3 Times Per Month	1–2 Times Per Week	3–4 Times Per Week	Every Day
Wrote brief answers to questions						
Wrote extensive answers to questions						
Did a think-aloud or explained how they applied a skill or strategy						
Worked on a written literature extension project						

34. This year, how often did the low-achieving students in your reading class work on the following areas in written comprehension?

Frequency	Never	Less Than Once a Month	1–3 Times Per Month	1–2 Times Per Week	3–4 Times Per Week	Every Day
Editing the capitalization, punctuation, or spelling of their own writing						
Editing the word use, grammar, or syntax of their own writing						
Revising their writing by elaborating and extending what they wrote						
Revising their writing by recognizing or refining what they wrote						

35. This year, how often did the low-achieving students in your reading class work on comprehension using … ?

Frequency	Never	Less Than Once a Month	1–3 Times Per Month	1–2 Times Per Week	3–4 Times Per Week	Every Day
Information text						
Narrative text with patterned or predictable language						
Narrative text with controlled vocabulary (sight words and/or easily sounded out words)						
Short narrative text without any attempt to control vocabulary (literature-based or thematic)						
Chapter book						

The MAP Program and Other Instructional Improvement Programs at My School

The questions in this section ask about formal, organized efforts that your school has taken to improve instruction for your students as well as both your own formal and informal learning experiences (e.g., professional development, staff development, and interactions with colleagues).

36. During this school year, there were four MAP training sessions. How many of these sessions were you able to attend?

- None
- One
- Two
- Three
- Four

37. The MAP trainers also may have scheduled visits to your school during the school year to follow-up on the MAP training and answer specific questions that teachers might have had. To date, how many of these sessions were you able to attend?

No visits were scheduled this year.
None. Although visits were scheduled, I was not able to attend.
One
Two
Three or more

38. To what extent do you agree or disagree with the following statements about the MAP Program at your school?

Agreement or Disagreement	Completely Disagree	Mostly Disagree	Mostly Agree	Completely Agree
I am capable of making the kinds of changes called for by the MAP Program				
The kinds of changes called for by the MAP Program are helping students in my classroom reach higher levels of achievement				
The MAP Program requires me to make major changes in my classroom practices				
I strongly value the kinds of changes called for by the MAP Program				
I am committed to using the resources provided by the MAP Program in my classroom				
The MAP Program has improved my ability to place and/or group my students				
The MAP Program has improved my ability to determine student mastery of skills				
The MAP Program has improved my ability to identify the core deficits of struggling students				

39. During this academic year, did your school participate in any formal school improvement programs other than MAP that involved students in your class?

Yes
No

41. During this school year, how often did the following things occur as part of these other school improvement programs?

Frequency Over the School Year	Never	1–2 Times	3–5 Times	6–10 Times	More Than 10 Times
I watched an instructional leader (e.g., coach, coordinator, or facilitator) model instruction					
An instructional leader observed me teach and gave me feedback about improving my teaching techniques					
An instructional leader observed me teach and gave me feedback about my use of curriculum materials					

Frequency Over the School Year	Never	1–2 Times	3–5 Times	6–10 Times	More Than 10 Times
An instructional leader gave me feedback about ways of assessing my students, how to interpret the results of these assessments, or how to use those assessments to improve my teaching					
An instructional leader studied my students' work and commented on ways I could improve their learning of subject matter					

42. To what extent do you agree or disagree with the following statements about these other school improvement program(s) in your school?

Agreement or Disagreement	Completely Disagree	Mostly Disagree	Mostly Agree	Completely Agree
I am capable of making the kinds of changes called for by these other school improvement programs				
The kinds of changes called for by these other school improvement program are helping my students reach higher levels of achievement				
These other school improvement programs require me to make major changes in my classroom practice				
I strongly value the kinds of changes called for by these other school improvement programs				

43a. Over this past school year, what was the total number of hours of professional development you received?
Number of professional development hours _____

43b. Did you enter "0" above?
Yes
No

44. During the past year, how many professional development sessions did you participate in that focused on … ?

Number of Sessions	None	1–2 Sessions	3–7 Sessions	8 or More Sessions
Student assessment				
Curriculum materials or frameworks				
Content or performance standards				
Teaching methods				
Use of technology in instruction				
Multicultural or diversity issues				
Classroom management and/or student discipline				

Number of Sessions	None	1–2 Sessions	3–7 Sessions	8 or More Sessions
Parent involvement and/or community relations				

45. Considering formal and informal professional development opportunities you had in reading/language arts this school year, how much time and effort did you devote to … ?

Amount of Time and Effort	None	1	2	3	4	5	A Great Deal
Analyzing or studying reading/language arts curriculum material							
Improving my skills at doing miscue analysis							
Improving my skills at designing reading/language arts tasks for my students							
Improving my knowledge of phonetics							
Improving my knowledge of guided reading strategies that help students use context cues							
Improving my knowledge of the writing process							
Extending my knowledge about different ways to help students blend and segment sounds							
Extending my knowledge about different reading comprehension strategies such as KWL or reciprocal teaching							
Using test data to help identify the needs of students and monitor their progress							

46. How would you rate the quality of your formal and informal learning experience this year in terms of … ?

Quality of Learning Experiences	Poor	Fair	Good	Very Good	Excellent
Giving you opportunities to work on aspects of your teaching that you are trying to develop					
Providing you with knowledge or information that you have found useful in the classroom					
Relating coherently to each other					
Providing you with useful feedback on your teaching					
Making you pay closer attention to particular things that you were doing in the classroom					
Leading you to seek our additional information from another teacher, an instructional leader, or some other source					
Leading you to think about an aspect of your teaching in a new way					
Leading you to try new things in the classroom					
Enhancing your understanding of assessment data and how to use data in your teaching					

Your Background Characteristics

Finally, we would like to ask you about some basic demographic characteristics.

47a. Which best describes your MAIN teaching assignment?
Self-contained classroom teacher (i.e., you teach all core subjects: math, reading, language arts, science, social studies, etc.)
Specialist teacher

47b. Please mark your primary subject area assignment *this year.*
English as a Second Language
Language Arts
Reading Specialist
Writing Specialist
Mathematics
Special Education
Other, *specify* _____

48. How many years have you taught language arts or reading prior to this year? If you have been teaching for less than a year, enter "0."

Number of years _____

49. How long have you been assigned to teach at your current school?
Number of years _____

50. What was your undergraduate major field of study?
Do not have an undergraduate degree
Education
English
Social or behavior sciences (economic, history, sociology, or psychology)
Foreign language
Mathematics
Natural/physical sciences
Other, *specify* _____

51. What was your major field of study for your highest graduate degree?
Do not have a graduate degree
Education
English
Social or behavior sciences (economic, history, sociology, or psychology)
Foreign language
Mathematics
Natural/physical sciences
Other, *specify* _____

52. What type of teaching certification do you hold from the state where you teach? *(Choose ALL that apply)*
Permanent or standard certification
Probationary certification
Temporary, provisional, or emergency certification
Alternative certification
Not certified
Other, *specify* _____

53. About how many undergraduate- or graduate-level classes have you taken at a college or university in … ?

Number of Classes	None	1–3 Classes	4–6 Classes	7–9 Classes	11–15 Classes	16 or More Classes
English or a related language arts field						
Methods of teaching reading, English, and/or language arts						

54. Are you … ?
Female
Male

55. Are you … ? *(Choose ONE)*
Hispanic, regardless of race
Black, not of Hispanic origin
White, not of Hispanic origin
Asian or Pacific Islander
American Indian or Alaskan Native
Biracial/Multiracial
Other, *specify* _____

56. Please provide any additional information that would help us understand your response to a question or any comments that you have on this questionnaire.

Thank you for your help! Please click the "Submit Survey" button to submit your responses.

Appendix I. MAP Student Engagement Survey

Identification Information

*****Indicates a required response

**1a. ***Enter your first and last name

1. Indicate how strongly you agree or disagree that each of the following statements is true about this student.

This Student:	**Agreement or Disagreement**
Is eager to learn	◻ Strongly Disagree ◻ Disagree ◻ Agree ◻ Strongly Agree
Usually pays attention in class	◻ Strongly Disagree ◻ Disagree ◻ Agree ◻ Strongly Agree
Frequently argues with others	◻ Strongly Disagree ◻ Disagree ◻ Agree ◻ Strongly Agree
Completes school work in an organized way	◻ Strongly Disagree ◻ Disagree ◻ Agree ◻ Strongly Agree
Often talks back to adult**s**	◻ Strongly Disagree ◻ Disagree ◻ Agree ◻ Strongly Agree
Works well independently	◻ Strongly Disagree ◻ Disagree ◻ Agree ◻ Strongly Agree
Wants to do well in school	◻ Strongly Disagree ◻ Disagree ◻ Agree ◻ Strongly Agree

This Student:	Agreement or Disagreement
Keeps his/her personal belongings organized	◻ Strongly Disagree ◻ Disagree ◻ Agree ◻ Strongly Agree
Often acts impulsively	◻ Strongly Disagree ◻ Disagree ◻ Agree ◻ Strongly Agree
Works hard on school assignments	◻ Strongly Disagree ◻ Disagree ◻ Agree ◻ Strongly Agree
Disrupts the work of others	◻ Strongly Disagree ◻ Disagree ◻ Agree ◻ Strongly Agree
Persists when work is difficult	◻ Strongly Disagree ◻ Disagree ◻ Agree ◻ Strongly Agree
Gets angry easily	◻ Strongly Disagree ◻ Disagree ◻ Agree ◻ Strongly Agree
Usually completes work on time	◻ Strongly Disagree ◻ Disagree ◻ Agree ◻ Strongly Agree
Uses free time in constructive ways	◻ Strongly Disagree ◻ Disagree ◻ Agree ◻ Strongly Agree
Sometimes damages property	◻ Strongly Disagree ◻ Disagree ◻ Agree ◻ Strongly Agree
Works carefully and methodically	◻ Strongly Disagree ◻ Disagree ◻ Agree ◻ Strongly Agree

This Student:	Agreement or Disagreement
Gets into fights with other children	◻ Strongly Disagree ◻ Disagree ◻ Agree ◻ Strongly Agree
Enjoys reading	◻ Strongly Disagree ◻ Disagree ◻ Agree ◻ Strongly Agree
Enjoys writing	◻ Strongly Disagree ◻ Disagree ◻ Agree ◻ Strongly Agree

2. Rate how well this student performs in each of the following areas compared with the other students in your class**.** If you do not have the target student for a particular subject, mark "Don't teach subject to this student."

Comprehension
Writing
Word Analysis

◻ Below Average
◻ About Average
◻ Above Average
◻ Don't teach subject to this student

3. Indicate whether the student participates or is enrolled in each of the following programs or services offered by this school.

Title I reading instruction or tutoring
Title I English/Language Arts instruction or tutoring
Other reading instruction or tutoring program
Other English/Language Arts instruction or tutoring
ESL/bilingual
Special education
Gifted and talented

◻ Yes
◻ No
◻ Don't know

4. Is this student involved in other reading or language arts programs that are not offered by this school? This includes individual tutors, private learning centers, and other services that are paid by the family.

◻ Yes
◻ No
◻ Don't know

5. Have this student's parents (or primary caregiver) participated in parent conferences or other meetings with you during this school year?

- ☐ Yes
- ☐ No
- ☐ Can't remember

6. To what extent have this student's parents (or primary caregiver) demonstrated interest or concern in their child's schoolwork?

- ☐ A great deal
- ☐ Some
- ☐ Only a little
- ☐ Not at all
- ☐ Don't know

7. Provide any additional information that would help us understand your response to a question or any comments that you have on this questionnaire.

Appendix J. MAP School Leader Survey

OMB Control No: 1850-0850 Expiration: 01/31/2011
Survey of School Leaders: The 2009–2010 School Year

Conducted by Learning Point Associates and Vanderbilt University for the Institute for Educational Sciences U.S. Department of Education, April 2010

- **Thank you for your participation in this study. We know that your time is valuable, and we greatly appreciate your willingness to complete this questionnaire. We realize that the survey asks a substantial number of questions, but answering these questions is important. Otherwise, we will not have a good sense of the topics that you cover in reading/language arts, the instructional practices that you use, and the environment of your school.**
- **Your responses are voluntary and confidential. If there is a question that you do not wish to answer, simply skip it. We hope, however, that you will answer as many questions as possible.**
- **All information that you provide will be reported only in a form that does not personally identify you or your school.**

Public reporting burden for this collection of information is estimated to average 30 minutes per response, including the time for reviewing instructions, gathering any needed data, and completing and reviewing the questionnaire. An agency may not conduct or sponsor, and a person is not required to respond to, a collection of information unless it displays a currently valid OMB control number. Send comments regarding this burden estimate or any other aspect of this questionnaire, including suggestions for reducing this burden to: Angela Arrington, IES Clearance Officer, 202-245-6409.

Involvement with Measures of Academic Progress (MAP)

1 During the 2008-2009 and 2009-2010 school years, have you been involved with the Measures of Academic Progress (MAP) program? By "involved", we mean one or more of the following: (1) you attended one or more MAP training sessions or consultative visits; (2) you reviewed MAP test scores of students; or (3) you worked with MAP teachers to identify students for placement in other reading classes or with reading specialists, plan students' instruction, or monitor students' progress.

- ○ Yes
- ○ No → ***Skip to Question 17***

2 How familiar are you with the MAP program and its resources?

- ○ Very familiar
- ○ Somewhat familiar
- ○ Only a little familiar → ***Skip to Question 17***
- ○ Not familiar at all → ***Skip to Question 17***

Your Roles and Activities

1 What is your primary role at this school?

- ○ Assistant Principal
- ○ Literacy Coach
- ○ Master/Mentor Teacher
- ○ Principal
- ○ Reading/Literacy Program Coordinator
- ○ Reading Specialist
- ○ Special Education Coordinator
- ○ Special Education Teacher
- ○ Teacher Consultant
- ○ Other (*specify*) _____-

3 **How many years have you been working in this role at this school?**

_____ years

4 **When working directly with teachers this year, how often did you . . .?**

	Frequency Over the School Year				
	Never ▼	A few times throughout the school year ▼	A few times per month ▼	1–2 days per week ▼	More than 2 days per week ▼
a. Demonstrate instructional practices and/or the use of curricular materials in a classroom	○	○	○	○	○
b. Observe a teacher who was trying new instructional practices or using new curricular materials	○	○	○	○	○
c. Share information or advice about classroom practices with a teacher	○	○	○	○	○
d. Examine and discuss what students were working on during a teacher's lesson	○	○	○	○	○
e. Examine and discuss the standardized norm-referenced or curriculum-referenced test results for students in a teacher's class	○	○	○	○	○
f. Examine and discuss the MAP test results for	○	○	○	○	○

students in a teacher's class					
g. Discuss other reading placements for a student in a MAP teacher's class, based at least partly on the student's MAP test results	○	○	○	○	○
h. Discuss in-class reading support strategies for a student in a MAP teacher's class, based at least partly on the student's MAP test results or other MAP resources	○	○	○	○	○
i. Discuss with a MAP teacher possible strategies for using MAP test results or other MAP re-sources to develop lesson plans for students in the teacher's class	○	○	○	○	○

The School Improvement Process

5 Does your school have a written improvement plan?

- ○ Yes
- ○ No, but we are in the process of developing one → ***Skip to Question 16***
- ○ No, and we are not currently developing one → ***Skip to Question 16***
- ○ Don't know → ***Skip to Question 16***

6 To what extent were each of the following an important priority in your school's improvement plan this year?

	Level of Priority		
	Not in our plan ▼	In the plan, but not a top priority ▼	A top priority ▼
a. Improving the school's facilities	○	○	○
b. Improving school climate (e.g., making school safer or fostering respect for others)	○	○	○
c. Improving parent participation	○	○	○
d. Improving student attendance	○	○	○
e. Improving the health and welfare of students	○	○	○
f. Improving the reading/language arts program	○	○	○
g. Improving the mathematics program	○	○	○
h. Improving the school's library, technology, or media	○	○	○
i. Improving another academic program or programs	○	○	○
j. Improving the use of data in making instructional decisions	○	○	○
k. Improving efforts to identify struggling readers as well as provide them with programs, tutoring, or similar efforts at this school to improve their reading/language arts skills	○	○	○
l. Improving efforts to identify gifted readers as well as provide them with programs, tutoring, or similar efforts at this school to improve their reading/language arts skills	○	○	○
m. Helping teachers to better differentiate instruction in their own classrooms by offering them expanded professional	○	○	○

development opportunities, reducing class size, providing special instructional materials, or other similar efforts

7 During the current school year, was this school formally identified as "in need of improvement" or placed in a formal status requiring school improvement by any of the following agencies? (*Check all that apply*)

- ○ The state education agency
- ○ The federal Title I program
- ○ The school district
- ○ Other agency (*specify*) _____

8 Which of the following programs provide reading/language arts instruction on-site at your school? (*Check all that apply*)

- ○ Gifted education
- ○ Special education
- ○ English as a Second Language
- ○ Title I
- ○ Other (*specify*) _____

9 During this school year, to what extent did you use data for each of the following purposes?

	Extent of Use			
	Data are not used in this way ▼	Used Minimally ▼	Used moderately ▼	Used extensively ▼
a. Identifying individual students who need remedial assistance in one or more subjects	○	○	○	○
b. Setting learning goals for individual students in one or more subjects	○	○	○	○
c. Tailoring instruction to individual students' needs	○	○	○	○
d. Developing recommendations for tutoring or other educational services for students	○	○	○	○
e. Assigning or reassigning students to classes or groups	○	○	○	○
f. Identifying and correcting gaps in the curriculum for all students	○	○	○	○
g. Encouraging parent involvement in student learning	○	○	○	○
h. Identifying areas where teachers need to strengthen their content knowledge or teaching skills	○	○	○	○

	Extent of Use			
	Data are not used in this way ▼	Used Minimally ▼	Used moderately ▼	Used extensively ▼
i. Determining topics for professional development	○	○	○	○
j. Setting school improvement goals	○	○	○	○
k. Celebrating the achievement of school goals	○	○	○	○

10 **We would now like to ask you about the MAP program in particular. To what extent do you agree or disagree with the following statements about the MAP Program at your school?**

► *Select one for each*	Agreement or Disagreement			
	Completely disagree ▼	Mostly disagree ▼	Mostly agree ▼	Completely agree ▼
I am capable of making the kinds of changes called for by the MAP Program.	○	○	○	○
The kinds of changes called for by the MAP Program are helping students in the participating classrooms reach higher levels of achievement	○	○	○	○
The MAP Program requires me to make major changes in my instructional or administrative practices	○	○	○	○
I strongly value the kinds of changes called for by the MAP program	○	○	○	○
Teachers who are participating in the MAP program are committed to using it in their classrooms.	○	○	○	○
District staff and officials who make decisions about reading/language arts instruction are committed to schools' use of MAP testing and resources	○	○	○	○
The MAP program has improved participating teachers' ability to place and/or group their students	○	○	○	○
The MAP program has improved participating teachers' ability to determine student mastery of skills	○	○	○	○
The MAP program has improved participating teachers' ability to identify the core deficits of struggling students	○	○	○	○

Q11 **Over the course of the current school year, to what extent have you . . . ?**

	Frequency Over the School Year			
	Never	A few times throughout the school year	A few times per month	More than a few times per month
	▼	▼	▼	▼
Accessed MAP test data on individual students	○	○	○	○
Generated MAP reports for students in a classroom	○	○	○	○
Used the information in the reports to group students	○	○	○	○
Used the MAP test data to identify those students who should have different reading placements (e.g., in a gifted education teacher's classroom or with a reading specialist, literacy coach, or tutor) for all or some of their reading instruction	○	○	○	○
Accessed and used Lexiles and Lexile book lists	○	○	○	○
Accessed and used Descartes	○	○	○	○
Used MAP resources to obtain information about instructional strategies for struggling readers	○	○	○	○
Used MAP resources to obtain information about instructional strategies for students who are high-achievers in reading	○	○	○	○

The Reading and Language Arts Program

12 To what extent do you agree or disagree with the following statements about the reading/language arts program at this school?

	Agreement or Disagreement			
	Completely disagree ▼	Disagree ▼	Agree ▼	Completely agree ▼
a. The reading/language arts program at this school needs major improvement	○	○	○	○
b. The reading/language arts instruction provided to students is much better than it was last year	○	○	○	○
c. The reading comprehension skills of most students in this skill are at or above grade level	○	○	○	○
d. The ability of students in this school to write for a variety of purposes and audiences is at or above grade level	○	○	○	○
e. Using strategies that accelerate reading/language arts instruction for above-grade-level students is a major goal at this school.	○	○	○	○
f. Using strategies that remediate instruction for students who are achieving below grade level in reading/language arts is a major goal at this school.	○	○	○	○
g. Teachers are expected to use appropriate grouping procedures in the reading/language arts program.	○	○	○	○

13 Because the MAP program is targeted at either 4th- or 5th- grade students in your school, we are especially interested in the your school's overall approach to reading/language arts instruction for these grades. Which of the following statements best describe the reading/language arts instruction for <u>struggling</u> <u>readers</u> in these grades?

- ○ Struggling readers receive the majority of their formal reading/language arts instruction from instructional staff other than their classroom teacher (e.g., pullout or plug-in programs).
- ○ Struggling readers receive equal amounts of their formal reading/language arts instruction from both instructional staff in other programs and their primary classroom teacher.
- ○ Struggling readers receive the majority of their formal reading/language arts instruction from their primary classroom teacher.
- ○ Other (*specify*) _____

14 Which of the following statements best describe the reading/language arts instruction for <u>gifted readers</u> in 4th- and 5th-grade?

- ○ Gifted readers receive the majority of their formal reading/language arts instruction from instructional staff other than their classroom teacher (e.g., pullout or plug-in programs).
- ○ Gifted readers receive equal amounts of their formal reading/language arts instruction from both instructional staff in other programs and their primary classroom teacher.
- ○ Gifted readers receive the majority of their formal reading/language arts instruction from their primary classroom teacher.
- ○ Other (*specify*) _____

Professional Development

15 Over this past school year, what was the total number of hours of professional development you received?

_____ hours → ***If "0", skip to Question 17***

16 **Considering <u>both</u> formal and informal professional development opportunities you had this year, how would you rate the quality of your formal and informal learning experiences this year in terms of . . . ?**

► *Select <u>one</u> for each*	Amount of Time and Effort None ▼	1 ▼	2 ▼	3 ▼	4 ▼	5 ▼	A great deal ▼
Assessment of student skills in reading/language arts	○	○	○	○	○	○	○
Identification of students with exceptionally low or exceptionally high skills in reading/language arts	○	○	○	○	○	○	○
Curriculum materials or frameworks for reading/language arts	○	○	○	○	○	○	○
Content or performance standards in reading/language arts	○	○	○	○	○	○	○
Teaching methods in reading/language arts	○	○	○	○	○	○	○
Use of technology in instruction in reading/language arts	○	○	○	○	○	○	○
Teaching strategies targeted at helping struggling readers	○	○	○	○	○	○	○
Teaching strategies targeted at exceptionally skilled readers	○	○	○	○	○	○	○
The use of ability grouping in reading/language arts instruction	○	○	○	○	○	○	○

The use of assessment data to identify students for placement with other reading professionals at my school	○	○	○	○	○	○	○
The use of other types of grouping strategies in reading/language arts	○	○	○	○	○	○	○
Multicultural or diversity issues	○	○	○	○	○	○	○
Classroom management and/or student discipline	○	○	○	○	○	○	○
School governance (e.g., local school council, site-based management, decision making)	○	○	○	○	○	○	○
School improvement planning or goal setting	○	○	○	○	○	○	○
Social services for students	○	○	○	○	○	○	○
Safety or school climate issues	○	○	○	○	○	○	○
Improving parent involvement and/or community relations	○	○	○	○	○	○	○

Your Background

17 Are you . . . ?

- ○ Female
- ○ Male

18 Are you . . .? *(Mark ONE)*

- ○ Hispanic, regardless of race

- ○ Black, not of Hispanic origin
- ○ Tab, not of Hispanic origin
- ○ Asian or Pacific Islander
- ○ American Indian or Alaskan Native
- ○ Biracial/Multiethnic
- ○ Other (*specify*) _____

19 Which best describes your employment status in this school system?

- ○ Full-time administrative appointment
- ○ Part-time administrative appointment
- ○ Regular full-time teaching appointment
- ○ Regular part-time teaching appointment
- ○ Permanent substitute teaching appointment
- ○ Other (*specify*) _____

20 How many years have you worked as an administrator? _____ years

21 How many years have you worked as a teacher? _____ years

22 What was your undergraduate major field of study?

- ○ Do not have an undergraduate degree
- ○ Education
- ○ English
- ○ Social or behavioral sciences (economics, history, sociology, or psychology)
- ○ Foreign language
- ○ Mathematics
- ○ Natural/physical sciences
- ○ Other (*specify*) _____

23 What was your major field of study for your highest graduate degree?

- ○ Do not have a graduate degree
- ○ Education
- ○ English

- ○ Social or behavioral sciences (economics, history, sociology, or psychology)
- ○ Foreign language
- ○ Mathematics
- ○ Natural/physical sciences
- ○ Other (*specify*) _____

24 About how many undergraduate or graduate level classes have you taken at a college or university in . . . ?

	Number of Classes					
	None ▼	1–3 classes ▼	4–6 classes ▼	7–9 classes ▼	11–15 classes ▼	16 or more classes ▼
English or a related language arts field	○	○	○	○	○	○
Methods of teaching reading, English, and/or language arts	○	○	○	○	○	○

Appendix K. MAP Recruitment Process

This appendix describes the process of recruiting districts and schools to participate in the MAP study.

Identification of targeted sites (spring 2008)

To identify eligible schools for recruitment, REL Midwest used the National Center on Education Statistics (NCES) Common Core of Data to create a file with contact information and demographics on all elementary schools in the Midwest Region that contained both grade 4 and grade 5 classrooms. This file was merged with information from NWEA that classified districts into one of four categories according to their enrollment status:

- Enrollment Status 1: District expressed initial interest in MAP program
- Enrollment Status 2: District has demonstrated ongoing interest in implementing the MAP program
- Enrollment Status 3: District is planning to implement MAP program
- Enrollment Status 4: District has implemented MAP program and is a district partner.

Using this information, the study team eliminated districts that had already adopted the program and prioritized the remaining districts for recruitment according to their enrollment status. The study team believed Status 2 and Status 3 districts were most likely to demonstrate interest in the study. They therefore planned to contact them first, followed by Status 1 districts and eventually all remaining districts in the file.[86]

Because of the difficulties associated with equating performance across different state accountability measures, the study team narrowed this initial list of school candidates to those in a single state. Illinois was selected because it had the largest number of potentially eligible and interested districts among the seven Midwest states. In addition, unlike some Midwest states, which administer their state accountability tests in the fall, Illinois administers its test in the spring of each school year. This provided nearly two years between initial MAP implementation (fall 2008) and final testing (spring 2009), the minimum time the study team deemed necessary for teachers to fully implement the MAP training and testing program.

[86] Remaining districts were districts that were included in the NCES file but were not listed in the NWEA file. These districts had not expressed a prior interest in MAP.

Initial contact with districts (spring 2008)

The study team identified 88 districts in Illinois representing 553 elementary and intermediate schools serving grades 4 and 5 to contact about the study.[87] The recruitment process was implemented in three stages (table K.1). In the first stage, the study team sent an initial letter by regular mail and e-mail to superintendents and other key district staff (for example, directors of curriculum, directors of assessment) to each potentially eligible district. The letter introduced the study and briefly explained the benefits and requirements of participation. Empirical Education, a REL Midwest subcontractor, followed up with phone calls to district staff to determine interest and, if applicable, schedule a one-hour web conference session with the district. During the initial web conference, members of the study team provided details about the MAP program; discussed the benefits of participation for districts and school staff and students; and specified the roles, responsibilities, and eligibility criteria for school participation. After this web conference, districts expressing continued interest were asked to complete a fact sheet that asked for each school's student demographic characteristics, grade configurations, assessments administered, professional development, and other information needed to confirm school eligibility. Districts sent a completed fact sheet on their elementary schools to a member of the study team, who reviewed the information to verify whether these districts had at least one eligible school before proceeding to the next recruitment stage.

Table K.1. Recruitment stages and sample sizes

Stage	Number of districts	Number of schools
Initial district contact (spring 2008)		
Sent introduction letter and made follow-up call	88	553
Conducted initial web-based conference call	14	93
Approved school eligibility/verified interest	7	54
District site visits (spring/summer 2008)		
Presented MAP program and study	7	54
Collected memorandum of understanding	5	32
Conducted grade-level random assignment	5	32
District and school follow-up site visits (fall 2008)		
Presented study, confirmed teacher eligibility	5	32
Conducted random assignment	5	32

Source: Authors.

[87] Letters were sent first to Status 3 schools, then to Status 2 and Status 1 schools. Because of delayed and lower than anticipated responses from study districts, letters were eventually sent to all elementary schools in Illinois.

District site visits (spring/summer 2008)

During the second stage of recruitment, an NWEA representative and members of the study team met with district staff and school administrators to present in-depth information about the MAP program and eligibility requirements and the responsibilities and benefits of study participation. After the presentation, the study team used a protocol to gather information from administrators on the number of eligible grade 4 and grade 5 classrooms in each school; classroom structures (for example, number of split classrooms, team teaching); assessments and assessment-related programs currently used in the school; planned professional development; and other relevant information to confirm school eligibility.

The study team followed up with these districts shortly after the site visits to confirm which schools were eligible for participation. If at least one school in the district was confirmed as eligible and agreed to participate in the study, a memorandum of understanding was drafted and sent to the district for review and approval. In order for a school to participate in the study, the district superintendent, key district point of contact, and the school's principal had to sign the memorandum of understanding. Once the study team received the memorandum of understanding, eligible schools were placed in the random assignment pool.

District and school follow-up site visits (fall 2008)

In the third stage of recruitment, consent was gathered from principals, and all regular education teachers in grades 4 and 5 after schools were randomly assigned to one of two conditions.[88] Shortly after school began in fall 2008, a member of the study team visited the study schools and presented an overview of the study to school administrators and regular education classroom teachers in grades 4 and 5.[89] Teachers reviewed a packet of information that included details about the study and a teacher consent form. To compensate participants for their time, principals

[88] The REL Midwest and its subcontractors rendered research services to each participating district in a manner consistent with the provisions of the Family Educational Rights and Privacy Act (FERPA). All data provided to REL Midwest was the property of the participating district, even while stored in REL Midwest's databases. REL Midwest, in collaboration with our subcontractors, took precautions to ensure the data were accessible only to company personnel and consultants assigned to this project or to authorities legally authorized for access. Before analysis, personally identifying information was removed from all analytic data files so that students and teachers could not be identified. Teacher and student demographic and achievement data was transferred by a district employee directly to a password protected, secure website. Trained observers conducted classroom observations using a computer based observation protocol. Observers entered data into a database and uploaded the data to a secure website at the end of the each observation session. Empirical Education, a REL Midwest subcontractor, sent unique, password protected URL's to each teacher, via email, to complete the annual spring survey and each set of instructional logs. Teachers clicked on the link and then entered a unique username and password to access the survey or instructional log. When a teacher completed the survey and instructional log, they clicked "submit," and data was uploaded to a secure database. The IRB approved protocol did not require that parent permission be obtained through the use of a signed permission form. Procedures for acquiring parental consent were followed in accordance with local district policies and requirements. Parents were notified by each of the participating schools that the study was occurring and were given the option to remove their child from the study.

[89] The third stage of recruitment was handled differently in the largest district because of the large number of participating schools. During the initial site visit, in spring 2008, researchers presented the study to administrators and teachers from all of the study schools at one time in the auditorium of the local high school. They conducted a follow-up site visit shortly before the 2008/09 school year began to gather administrators' consent, distribute information to teachers, and describe the process for working with administrators to gather teachers' consent.

and teachers were offered a $250 stipend each school year for completing data collection activities. All eligible staff were given time to review the information carefully and sign and submit the consent form shortly after the presentation. Teachers who were still uncertain about participating in the study after this meeting were invited to email or fax their signed consent at a later time. Staff from nearly all of the 32 randomly assigned schools participated in the study.

Appendix L. Assessment of Control Group Contamination and of Integrity of Year 2 Intervention–Control Contrast

Randomly assigning schools to treatment-control conditions has been advocated as a means of minimizing contamination between conditions. In designing the MAP study we considered simply assigning whole schools to conditions (MAP or control). We also considered random assignment of teachers within schools. Neither of these options was acceptable to the principals in schools who initially expressed willingness to participate in the study. As such, we devised a design where grades (4 or 5) would be randomly assigned to either MAP or control condition within each school (see table L.1)[90]. Specifically, for each school, grade 4 or grade 5 was randomly assigned *as a grade-level group* to either receive the MAP program or to conduct business as usual without exposure to the MAP tests, training program, or other resources provided by the program (e.g., Descartes and Lexile reports). If grade 5 teachers in School A were assigned to the treatment condition, then grade 4 teachers in this same school were assigned to the control condition. Similarly, if grade 5 teachers in School B were assigned to the control condition, then this school's grade 4 teachers were assigned to treatment condition. In this way, the control group for grade 4 consists of fourth-grade classes from those schools where MAP was randomly assigned to grade 5; the control group for grade 5 consists of fifth-grade classes from those schools where MAP was randomly assigned to grade 4.

Table L.1. Study design for the MAP RCT study

Study year	**Results of randomization**			
	Fourth grade assigned to MAP		**Fifth grade assigned to MAP**	
	Grade 4 = MAP	**Grade 5 = control**	**Grade 4 = control**	**Grade 5= MAP**
1 (2008-09)	T_1^{4th} $S_{Cohort1}$	C_1^{5th} $S_{Cohort1}$	C_1^{4th} $S_{Cohort1}$	T_1^{5th} $S_{Cohort1}$
2 (2009-10)	T_1^{4th} $S_{Cohort2}$	C_1^{5th} $S_{Cohort2}$	C_1^{4th} $S_{Cohort2}$	T_1^{5th} $S_{Cohort2}$

Because the above design gives rise to the possibility of between-group contact or contamination within a school, this appendix examines if elements of the MAP intervention "leaked" into the control classes, thereby weakening the treatment-control contrast or the achieved relative strength (defined in chapter 3).

[90] To encourage participation, principals were told that the control group could receive MAP training and testing after the study was completed. To maintain MAP's organizational presence, third-grade teachers were, at the principal's request, provided MAP training and testing in Year 2. For simplicity, these inducements are not shown in table L.1. The teachers and students associated with these inducements are not part of the samples that were included in the intent-to-treat analyses.

Overview of the issues

As shown in table L.1, the fourth-grade students assigned to the MAP condition in Year 1 would be fifth-grade control students in Year 2. To the extent that MAP was fully implemented in Year 1, these students would have been exposed to elements of the MAP program in Year 1, and this contamination would have weakened the Year 2 MAP versus control contrast for grade 5. Furthermore, the first year implementation results showed equivalent levels of differentiated instruction in MAP and control conditions. The latter could be due to several reasons. Of particular concern is that the control teachers could have received assistance from MAP teachers, MAP personnel, or school personnel. Because of the above possibilities, this study treats the Year 2 grade 5 cohort results as exploratory.

In this appendix, we first identify possible sources of contamination in the Year 2 grade 5 cohort. Second, we examine the extent to which school leaders encouraged MAP-like professional development for their control teachers. Third, we examine the implementation timeline (especially spring 2009), which revealed that the extended time required to implement the MAP program left little opportunity for teachers to use tactics for differentiated instruction or to disseminate these tactics to other teachers in the control condition. Fourth, we identify the potential origins of the equivalent levels of differentiated instruction seen in control and treatment classrooms. And finally, we revisit the logic model and training implementation timeline. Findings from these examinations suggest that fourth-grade students of MAP teachers (in Year 1) were likely only minimally exposed, if at all, to the treatment during Year 1.

Issue 1: Possibilities for between-group contamination

This study used a two-year, cluster-randomized design to obtain unbiased estimates of the impact of the MAP program on student reading achievement.[91] As described above, within each school, grades 4 or 5 were randomly assigned *as a grade-level group* to treatment or control conditions. Although control teachers in either grade did not have access to MAP training, MAP resources, and MAP testing of students, it is conceivable that MAP-related principles and best practices could be shared between fourth- and fifth-grade teachers and school leaders, especially those who were enthusiastic about MAP. For example, school leaders could encourage control teachers to seek out additional professional development, communicate directly with the control teachers about MAP ideas, or encourage sharing of MAP-related information across conditions.

Issue 2: The potential for some augmentation of professional development for teachers in fourth-grade control condition classes

To examine the extent to which MAP-related training or information may have been shared between fourth and fifth grade teachers and school leaders, we compared survey results on teachers' professional development experiences during Year 1. Specifically, we compared the amount of professional development that both grade 4 and grade 5 teachers reported receiving in assessment or data use, as well as the total amount (in hours) of professional development they received during Year 1. Among grade 4 teachers, there were no differences between MAP and

[91] The intent-to-treat (ITT) model used to estimate the effects of the MAP program on student reading achievement is described in appendix B.

control teachers in the total hours of professional development or in the percentages of teachers who reported receiving professional development in assessment or in data use. However, among fifth grade teachers, the MAP program had impacts in the expected direction on each of these three measures. That is, grade 5 treatment teachers reported that they spent significantly more time than control teachers in professional development activities during Year 1. In addition, treatment teachers reported spending significantly more time in professional development activities focused on assessment and data use than control teachers. These results were significant at the p=.05 level.

Based on results of our Year 1 analyses pertaining to different levels of school leadership support for the MAP program across schools, as well as data on self-reported receipt of professional development, the possibility exists that the level of professional development for fourth-grade control group teachers was enhanced by those principals who were more supportive or enthusiastic about the MAP program. Specifically, the following possibilities exist:

- If the principal's enthusiasm was a source of motivation for school staff in the treatment group to implement new practices related to differentiation of instruction, this same enthusiasm may have also affected control group teachers.
- The fact that the percentages of professional development in data use were very similar for fifth-grade treatment teachers and fourth-grade control teachers in the same school (83 percent vs. 76 percent, respectively), and the fact that the percentages for fourth-grade treatment teachers and fifth-grade control teachers (in the same school) were also similar (58 percent vs. 58 percent, respectively) might suggest that "school" is a stronger predictor of professional development in data use than is treatment status. However, one needs to ask whether this reflects no impact or contamination.
- If the level of professional development in the control condition was increased in an effort to make receipt of training equitable between the treatment and control teachers, that could be conceived as contamination because control teachers were brought up to the higher level of professional development for the treatment teachers, which would not have occurred in the absence of the intervention. However, if the level of professional development on this topic was completely unaffected by the intervention, the proper conclusion would be that the level was and would have been high in both conditions without the intervention. If the second explanation is not plausible, this could indicate contamination.
- About half the schools reflect a relatively high level support for MAP. If the overlap between "high support" schools and schools with high levels of professional development in grade 5 treatment and grade 4 control conditions is considerable, this might be evidence of contamination.

Our Year 1 analysis does show that in about half of the schools, school leaders engaged in more MAP training and consultation sessions. The average number of training and consultation sessions was 11.80 in schools where the principal and/or other leaders participated in MAP activities versus 5.92 sessions where only the teachers attended these activities. Although these school-based differences are potentially important because those schools with greater exposure to MAP training and consultation might have higher levels of implementation at the teacher level, the greater exposure (in terms of number of knowledgeable school personnel) also could result in contamination of the control condition, especially if school leaders in high MAP support schools promoted MAP-like professional development in the control group.

Table L.2 reproduces the overall average level of professional development reported by teachers in grades 4 and 5 by MAP and control conditions in Year 1. As can be seen, the average amount of professional development for fourth-grade control teachers is 30.8 hours. On the other hand, the average number of hours of professional development for their fifth-grade control-group counterparts is 23.3 hours.[92] Our analysis of Year 1 data showed that the difference between MAP and control in grade 4 was not statistically significant, but the difference was significant in grade 5. In connection with higher rates of data use for fifth-grade MAP teachers (36.0 hours) and fourth-grade control teachers (30.8 hours) (in the same school), it can be argued that school effects (e.g., principal motivations) may have been responsible for these results (especially the no-difference findings in professional development for grade 4). Below, we show that the average level of professional development for control group teachers was not enhanced in schools where principals were more supportive of the MAP program.

Table L.2. Year 1 average hours of professional development (PD) in schools with high and low engagement

Grade	MAP			Control		
	Average PD (Overall)	Leadership participation		Average PD (Overall)	Leadership participation	
		Low	High		Low	High
4	34.6	30.0	36.8	30.8	34.4	27.9
5	36.0	34.6	37.0	23.3	25.7	22.2

The speculation about the relationship between school-level leadership participation and the principal's enthusiasm for MAP is supported by data. The correlation between school-level leadership participation ($n = 17$ and $n = 14$ for high- and low-participation groups, binary coded),[93] and our index of the principal's attitude toward the MAP program (based on the spring 2009 School Leader Survey) is 0.40 ($p < 0.05$). That is, schools with higher MAP participation by school leaders have principals who are more positively disposed toward the MAP program. On the other hand, this predisposition did not appear to translate into enhanced professional development opportunities for control teachers, especially those in grade 4. Table L.2 provides average levels of professional development for MAP and control group teachers in schools designated as having low or high leadership participation. The subgroup means show that leadership participation level is not related to the average level of professional development. In particular, the fourth-grade teachers in the control group did not receive enhanced hours of professional development when leadership participation was classified as high.

Issue 3: The Role of Delayed Implementation

After examining the implementation timeline and processes for MAP training and consultation and upon the advice of our Technical Advisory Group, we decided to dedicate Year 1 to the implementation of MAP resources (e.g., testing, Lexiles, and DesCartes), formal training, and

[92] The difference between grades 4 and 5 average levels of professional development was not statistically significant. The pattern of data that triggered the concerns of contamination could be the result of sampling error or chance.
[93] The low-participation group includes two schools with no principal or staff participation.

consultation services. We originally estimated that there would be only a 1.5-month window for in-class use of MAP testing and differentiated instruction before state testing was to begin. In turn, we explicitly hypothesized that there would be no detectable relative effects on student reading achievement, mainly because the active ingredients of MAP—use of formative assessment and differentiated instructional practices based on formative assessment—would not be implemented by teachers with enough intensity to be seen as changes in classroom behavior and to induce changes in learning before statewide testing began.

Our original expectation was that formal training (Sessions 1–3) would be completed by the end of January. Our report showed that NWEA completed these sessions by December 2008, but consultation played a major role in spring 2009. Table L.3 provides a more detailed analysis of training in the spring semester (2009) and shows that training continued (through consultation) well into the end of the spring term. Here, we index a teacher's timing of the completion of training as the last month in which the teacher engaged in consultation activity beyond the first three formal training sessions.[94]

Table L.3. Teachers' completion of MAP training in Year 1

Grade	Teacher completion of MAP training and consultation	Timeline				Total
		Dec 2008	Jan to Mar 2009[a]	Apr 2009	May 2009	
4	Number of grade 4 teachers	5	13	24	11	53
	Cumulative percentage	9	34	79	100	
5	Number of grade 5 teachers	5	8	9	12	34
	Cumulative percentage	15	38	65	100	

a. The counts for January, February, and March 2009 were combined to prevent a disclosure risk.

Table L.3 reveals that relatively few teachers (9 percent and 15 percent for fourth and fifth grade, respectively) completed training in December 2008. Instead, the majority of teachers continued training through consultations during the spring semester. When state testing began in March 2009, only 34 percent and 38 percent of fourth- and fifth-grade teachers had completed training with NWEA staff. Many teachers, in both grades completed their training in April and May 2009 (65 percent and 62 percent in grades 4 and 5, respectively).

Issue 4: Preexisting individual differences in teacher practices

In the Year 1 analyses, we found that the percentage of teachers who report using grouping indicates the practice is very high in both treatment and control classes and the sources of information used to form groups are nearly the same. These findings could reflect no-differences

[94] MAP training involves four sessions, the last being delivered at the end of the academic year (May–June). The last session entails principles and practices for planning sustained subsequent growth. For the purposes of examining implementation of practices associated with data use and differentiated instruction, we count attendance at Sessions 1–3 and any follow-up consultation service that is used by individual teachers.

or contamination. In this section, we provide additional data to distinguish between these two hypotheses. Here, we argue that in addition to delayed/incomplete implementation, teachers in both conditions have preexisting classroom behaviors that are the result of prior training, other professional development, access to new instructional strategies on the Web, and so on. As a result of randomization, these preexisting individual differences in teacher practices are likely to be seen in both treatment and control conditions, independent of assignment to conditions.

Individual difference measures and indicators

Three methods of data collection (teacher surveys, teacher logs, and classroom observations) were used to assess whether MAP teachers were more able to adapt their instructional practices to reflect an emphasis on differentiated instruction. The classroom observations were conducted in fall, winter, and spring. Here, the fall observation was completed in September and October prior to the completion of the training on differentiated instruction (Session 3 or "Climbing the Data Ladder"). As a result, the fall observations provide the best reflection of preexisting teacher behaviors. We rely on fall and spring observational data to assess the influence of preexisting teacher behaviors and changes over time.

Observation-based indicators

A chief focus of these observations was to record instances of differentiated instruction. Differentiation could be exhibited in terms of the content of instruction, the processes of instruction, or the evaluation of products. We examine below the extent to which any differentiation occurred within multiple 10-minute observation segments that composed the overall classroom observation session. In our Year 1 analyses, data for up to three observations per teacher were combined into a single score for each teacher. The score reflects the proportion of observation segments in which *any* differentiated instruction occurred. The additional analyses reported here are based on separate indexes for fall and spring. Repeated-measures analysis of variance was used to assess changes over time, group effect, and the group-by-time interaction.

For the additional analyses reported in tables L.4 and L.5, we focus on data about the use of different instructional modalities (e.g., grouping), the use of any type of differentiation, and the extent to which the instructional practices of teachers reflect an integration of differentiated instructional practices.

Instructional modality. Each observation segment is coded for the type of instructional modality used by the teacher. The proportions of segments using whole class, small groups, and pairs are examined across groups for fall and spring.[95]

Differentiated instruction. Based on the work of Tomlinson (2001), the MAP classroom observation protocol collected information as to whether a teacher differentiated instruction in terms of content, process, and product for each of six areas of instructional focus: vocabulary, spelling, fluency, comprehension, writing, and speaking or listening. For each classroom observation, a class session differentiation score was constructed to reflect the proportion of ways that instruction was differentiated (content, process, and/or product) across the topics addressed in that particular class session. The index used in our Year 1 analyses represented the

[95] Because multiple modalities could occur in a 10-minute segment, these proportions do not sum to 100.0 percent.

proportion of observation segments in which any type of differentiation occurred over all topics and all ways. The index ranges from 0.00 to 1.00.

Integrated differentiation. The observation protocol also included separate ratings on the extent to which content, process, and product were differentiated for student readiness, learning style, and interest (see Tomlinson, 2001). According to Tomlinson, differentiated instruction is integrated into the instructional practices when the teachers consider types of differentiation (e.g., content) across potential forms of differentiation (e.g., readiness, interest). A 4-point Likert scale was used for each of these nine ratings (0 = not present, 1 = slightly integrated, 2 = partially integrated, and 3 = fully integrated). For each classroom observation, the nine ratings were recoded as 0 or 1 (not present or slightly integrated or more), summed, and this total was divided by 9 (the maximum sum that could be received for these nine ratings). Again, possible values ranged from 0.00 to 1.00.

Results for pre-post teacher behaviors

The proportions (P) of observation segments in fall and in spring during which teachers used specific modalities of instruction (whole class, small groups, and student pairing) are presented in tables L.4 and L.5 (grades 4 and 5, respectively). Also presented are the proportions of observation segments in fall and spring during which teachers engaged in any type of differentiated instruction (by type) and the extent to which this instruction reflects an integration of differentiated instruction across types, topics, and student attributes (e.g., interests, readiness). To highlight the main results of multiple statistical tests, the results for the repeated-measures analysis of variation (ANOVA) are simply summarized in terms of whether (i.e., yes) the statistical test reached conventional ($p < 0.05$) statistical significance.

Table L.4. Fall 2008 and spring 2009 observation results for instructional modality, differentiated instruction, and integration of differentiated instruction—grade 4

Construct	Measure	Group	Observation period				Significance of effects		
			Fall 2008 average P (SD)	ARSI	Spring 2009 average P (SD)	ARSI	Group	Time	Group x time
Instructional modality	Whole class	Control	0.61 (0.24)	–0.21	0.56 (0.29)	0.00			
		MAP	0.55 (0.33)		0.56 (0.31)				
	Groups	Control	0.13 (0.19)	0.22	0.36 (0.34)	–0.09		Yes	
		MAP	0.18 (0.25)		0.33 (0.32)				
	Pairs	Control	0.13 (0.21)	0.09	0.14 (0.28)	–0.09			
		MAP	0.15 (025)		0.12 (0.20)				
Differentiated instruction	Proportion of segments	Control	0.19 (0.27)	–0.16	0.25 (0.34)	–0.10			
		MAP	0.15 (0.23)		0.22 (0.29)				
Integrated differentiated instruction	Proportion agree	Control	0.03 (0.04)	0.15	0.05 (0.08)	–0.34			
		MAP	0.04 (0.08)		0.03 (0.04)				

Note: The numbers of teachers with both fall and spring observations were 44 and 28 in the MAP and control groups, respectively.

General statistical results. Across all tests of effects (group, time, and group by time), only two effects were statistically significant. Both these effects suggested that teachers changed the frequency with which they used any one instructional modality. Specifically, in the fourth grade (see table L.4), the proportion of segments in which small groups were used as the instructional modality increased for both MAP and control groups (from less than 0.20 to more than 0.30). In the fifth-grade (see table L.5), teachers in both groups increased their use of the whole-class type of instructional modality (from about 0.50 to more than 0.60).[96]

[96] In addition, regarding the concerns about contamination at the fourth-grade control/fifth-grade treatment schools, there does not appear to be any discernable pattern of change across the two groups of schools (as opposed to the two conditions shown in tables M.4 and M.5). Visual inspection of these tables suggests that both groups of schools had similar patterns and magnitude of change.

Table L.5. Fall 2008 and spring 2009 observation results for instructional modality, differentiated instruction, and integration of differentiated instruction—grade 5

Construct	Measure	Group	Observation period				Significance of effects		
			Fall 2008 average P (SD)	ARSI	Spring 2009 average P (SD)	ARSI	Group	Time	Group x time
Instructional modality	Whole class	Control	0.51 (0.26)	–0.07	0.66 (0.30)	–0.14		Yes	
		MAP	0.49 (0.30)		0.62 (0.26)				
	Groups	Control	0.30 (0.34)	–0.25	0.27 (0.31)	0.17			
		MAP	0.22 (0.29)		0.32 (0.29)				
	Pairs	Control	0.16 (0.26)	–0.41	0.11 (0.17)	0.00			
		MAP	0.07 (0.14)		0.11 (0.17)				
Differentiated instruction	Proportion of segments	Control	0.19 (0.26)	0.04	0.21 (0.32)	0.03			
		MAP	0.20 (0.29)		0.22 (0.29)				
Integrated differentiated instruction	Proportion agree	Control	0.03 (0.05)	–0.23	0.04 (0.06)	0.15			
		MAP	0.01 (0.12)		0.05 (0.08)				

Note: The numbers of teachers with both fall and spring observations were 27 and 43 in the MAP and control groups, respectively.

Preexisting teacher practices. Although the results in tables L.4 and L.5 are largely nonsignificant, the between-group differences for the fall and the spring are very small, in numerical values and in terms of ARSI values[97]. The ARSI values reported in tables L.4 and L.5 are consistently well below our a priori threshold. For the fall results, these differences are not statistically significant, as would be expected by virtue of the randomization process. The average fall ARSI (pretest) values for grades 4 and 5 are 0.02 and –0.18, respectively. What is interesting about the fall results is that teachers in both the MAP and the control condition exhibit nonzero levels of behavior that could not have been affected by the MAP training and contamination of MAP training across conditions—the fall observations were obtained prior to training in grouping and differentiated instruction. Moreover, with the exception of changes in the use of small-group instruction in grade 4 and the increased use of whole-class instruction in grade 5, there are no time or group-by-time effects for differentiated instruction and the integration of differentiated instruction. The average spring ARSI values for grades 4 and 5 are 0.12 and 0.04, respectively.

The lack of between-group differences for the spring observations (tables L.4 and L.5) on types of instructional modalities used by MAP and control teachers, the use of any differentiated

[97] Although there are no formal guidelines for interpreting the magnitude of ARSI values, as part of the proposal for this study we designated a minimum ARS threshold of 1.00 as the expected difference between conditions if groups differ as planned.

instruction, and the extent of integration of differentiated instruction are consistent with our expectation that much of the first year of implementation would be devoted to training and consultation of MAP teachers, and that teachers would not have sufficient time to enact what they had learned from the training and consultation sessions. As a result, the achieved relative strength indexes hover around 0.00, on average. For example, the average proportions of differentiated instruction in the spring are 0.25 and 0.22 for fourth-grade control and MAP teachers, respectively; comparable values are seen in table L.5 for grade 5 (i.e., 0.21 and 0.22 for control and MAP teachers). Rather than being the result of contamination of the control conditions, the findings on key variables of the MAP program (grouping and differentiated instruction) are likely to be due to delayed implementation and enactment of differentiated instruction in the classes. The nonzero levels of differentiated instruction and grouping in fourth- and fifth-grade MAP and control classes are likely to be due to preexisting teacher propensities. The observers' ratings (average proportion of agreement) of the integration of the differentiated instruction that was observed in classes are, on average, near zero for both grade 4 and grade 5.

Issue 5: Integrity of the grade 5 treatment-control contrast

Our analysis of how MAP training unfolds for teachers and the timing of state tests strongly suggests that there would be insufficient opportunity for teachers to change their practices and affect their students. In fact, this is the rationale for conducting a two-year study. Because of the length of initial MAP training, we believed that the program could not be implemented at a sufficient dose during the first year. For that reason, we expected to see no effects on state and MAP tests and no material difference between conditions in teacher practices. The data in table L.3 show that implementation of training and consultation lasted well into the spring term.

Grade 5 Year 2 control group contamination (based on its grade 4 MAP status) depends on whether the grade 4 MAP teachers actually implemented the core components of the MAP program in their classes. The mechanism for contamination requires that the students be exposed to different instructional practices than they would have had they not been assigned to the MAP group. The data presented in prior sections of this appendix suggest that the grade 4 MAP students received the same exposure to grouping practices and differentiated instruction as their control counterparts. Specifically, using the spring observation data, the average proportion of segments for small-group instruction was 0.36 and 0.33 for control and MAP, respectively; the use of student pairing was similar across groups (0.14 and 0.12). Exposure to differentiated instruction also was comparable across control and MAP conditions (0.25 and 0.22 on average for control and MAP).[98]

[98] Our Year 1 analyses revealed no evidence of different levels of differentiation (in fourth or fifth grade) based on self-reported behavior of teachers, logs, and observations. We also examined the degree of differentiated instruction using logs from the fall and spring. As argued above, the fall logs do not represent a "pure" pretest because the timing of the logs overlapped the time of the three main training sessions. Even though the fall-spring comparisons on the logs are not ideal, we did analyze them using a repeated-measures ANOVA. Across measures of comprehension, word analysis, and writing, we found no differences between control and MAP conditions in grades 4 and 5.

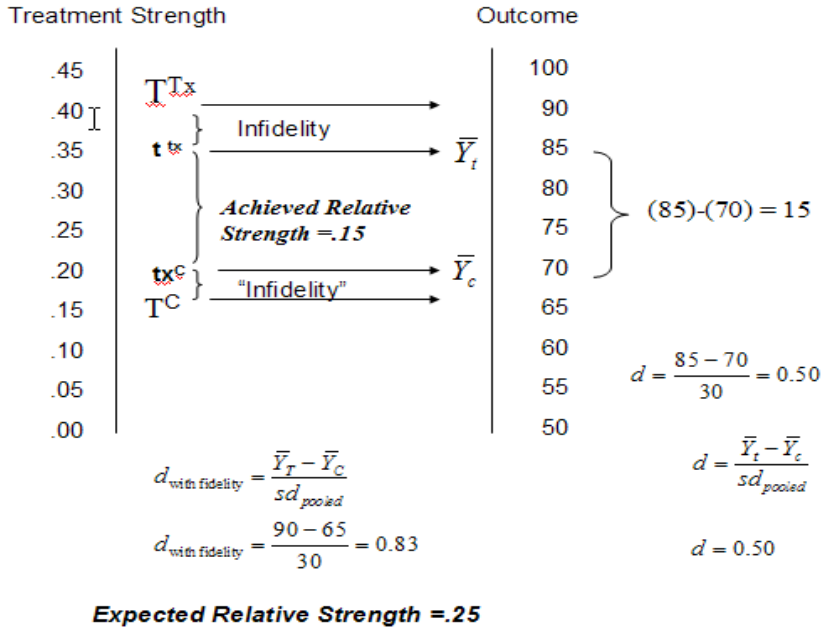
Appendix M. Implementation Fidelity and Achieved Relative Strength

The notion of intervention fidelity has been captured under a broad array of labels, such as treatment integrity, adherence, compliance, dose, exposure, quality of delivery, and treatment differentiation (see Dane and Schneider 1998; Dusenbury et al. 2003; Mowbray et al. 2003; O'Donnell 2008). Treatment integrity, compliance, and adherence refer to the extent to which participants (for example, teachers) deliver the intended innovation and other participants (for example, students) accept, receive, or are responsive to the intended services, at the intended level of treatment strength (see Boruch and Gomez 1977; Cordray and Pion 2006; Yeaton and Sechrest 1981). In practice, these constructs are often operationalized by indexes of dose, exposure, and quality.

Assessments of intervention fidelity involve the specification of a gold standard, or basis for comparison—a theory, model, or conception of the educational intervention to which intervention is faithful. Fidelity assessments must thus begin with a full characterization of the intervention in theory. These theories can be grand or small. What they have in common is a well-stated set of expectations about how the intervention is supposed to work; the logic underlying the intervention; and rationales for how and why these actions will produce the desired enhancements in student learning, motivation, and achievement. Fidelity assessments indicate how closely the intervention met these specifications. Reliable and valid measures of achieved intervention fidelity index the degree of discrepancy between what should have been implemented and what actually was implemented.

Figure M.1 defines intervention fidelity within the context of a randomized controlled trial. Cordray and Pion (2006) state that the outcome (Y_i) for a participant is determined by the achieved fidelity of the treatment, as implemented and received by that individual (t_i^{Tx}). When it is possible to stipulate the intended or theoretical strength of an intervention (T^{Tx}), true indexes of compliance, adherence, or treatment integrity can be derived. In such cases, the achieved intervention fidelity (or treatment integrity) can be represented as the difference between treatment as theorized (T^{Tx}) and the treatment as realized for individuals or groups of individuals (t^{Tx}). In figure M.1, across all participants, the degree of treatment infidelity, $T^{Tx} - t^{Tx}$, is 0.05 strength units ($0.40 - 0.35 = 0.05$).

Figure M.1. Example of representation of fidelity and relative strength in experiments



Source: Adapted from Cordray and Pion (2006), p. 116.

Cordray and Pion (2006) also incorporate treatment differentiation (see Waltz et al. 1993) into their definition of intervention fidelity. Treatment differentiation suggests that T^{Tx} has to be stronger than or different from the counterfactual condition. Counterfactual conditions are rarely unprogrammatic or unorganized collections of activities. Rather, as is often the case in education research, control conditions frequently consist of business as usual in terms of curriculum activities. Holland (1986) stipulates that theories or models of causality are embedded in control conditions. These theories/models can be designated as T^C. In such cases, the causal effect on the outcome $E(Y^{Tx})$ of the target treatment (T^{Tx}) has to be considered relative to the causal components in T^C associated with the production of the outcome $E(Y^C)$ in the counterfactual condition. As above, infidelity can occur when the actual comparison condition (t^C, as opposed to its theoretical counterpart, T^C) becomes more like the T^{Tx}. This migration can be caused by contamination or leakage of the t^C with elements of the T^{Tx} (Orwin et al. 1998; Shadish, Cook, and Campbell 2002), such as when core components of the treatment are provided to participants in the control condition. A parallel fidelity assessment of programmatic components in comparison conditions is required to determine whether this happens. Cordray and Jacob (2005) link fidelity assessment to contemporary statistical models of causal inference and refer to the difference between intervention and control conditions as the *achieved relative strength* of the contrast. The achieved relative strength is the difference between the treatment, as implemented, and the control, as implemented ($t^{Tx} - t^C$). Estimates of effects on the outcome are the result of the achieved relative strength of the contrast ($t^{Tx} - t^C$), not the theoretically expected difference ($T^{Tx} - T^C$). Because of these sources of infidelity, the observed effects can be less than originally expected. For the hypothetical example in figure M.1, the expected effect of a perfectly implemented intervention, based on $T^{Tx} - T^C$, is 0.83 standard deviation. Because of infidelity and contamination in both conditions, the achieved effect ($t^{Tx} - t^C$) on the outcome is only 0.50 standard deviation.

Appendix N. The Achieved Relative Strength Index

This appendix reports the Achieved Relative Strength Index (ARSI) (ARSI); a standardized difference between conditions (see Hulleman and Cordray 2009) is used. This index is modeled after Hedges's *g*. It is adjusted for small sample bias and clustering, as follows:

$$\hat{g} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1-1)SD_1^2+(n_2-1)SD_2^2}{(N_{\text{total}}-2)}}} \times \left(1 - \frac{3}{4(n_1+n_2)-9}\right).$$

Hedges's *g* for means corrected for clustering is calculated as follows:

$$g = (\frac{\bar{X}_1 - \bar{X}_2}{S_T}) \times (1 - \frac{3}{4(n_1+n_2)-9}) \times \sqrt{1 - \frac{2(n-1)p}{N-2}},$$

where

$\bar{X}_1$ = mean for group 1

$\bar{X}_2$ = mean for group 2

S_T = pooled within-group standard deviation

n = average cluster size

ρ = intraclass correlation

N = total sample size.

The variance is calculated as follows:

$$V_g = (\frac{N_T + N_C}{N_T \times N_C})(1 + (n-1)p) + g(\frac{(N-2)(1-p)^2 + n(N-2n)p^2 + 2(N-2n)p(1-p)}{2(N-2)[(N-2)-2(n-1)p]}).$$

Hedges's *g* for proportions (the Absolute Fidelity Index and the Binary Complier Index) are calculated as follows:

$$g = 2 * \arcsin\theta(\sqrt{p_{Tx}}) - 2 * \arcsin\theta(\sqrt{p_C}) \text{ X } (1 - \frac{3}{4(n_1+n_2)-9}),$$

where p_{Tx} = proportion of participants in the treatment group and p_C = proportion of participants in the control group. The variance is calculated as:

$$V_g = \frac{n_C + n_{Tx}}{n_C \times n_{Tx}} + \frac{g^2}{2 \times (n_C + n_{Tx})}$$

where n_C = sample size in the control group, n_{Tx} = sample size in Tx group, and *g* = Hedges's *g*.

Hedges's *g* for proportions corrected for clustering is calculated as:

$$g = 2 * \arcsin\theta(\sqrt{p_{Tx}}) - 2 * \arcsin\theta(\sqrt{p_C}) \text{ X } (1 - \frac{3}{4(n_1 + n_2) - 9}) \text{X } \sqrt{1 - \frac{2(n-1)p}{N-2}},$$

where p_{Tx} = proportion of participants in the treatment group and p_C = proportion of participants in the control group with variance

$$V_g = (\frac{N_T + N_C}{N_T \times N_C})(1 + (n-1)p) + g(\frac{(N-2)(1-p)^2 + n(N-2n)p^2 + 2(N-2n)p(1-p)}{2(N-2)[(N-2) - 2(n-1)p]})$$

where *g* = Hedges's *g* and ρ = intraclass correlation.

The lower and upper bounds for the Hedges's *g* effect sizes are calculated using the formula for the 95 percent confidence interval (g ± 1.96 * Standard Error of *g*) (see Hedges 2007 for further detail).

References

Ash, K. (2008). Adjusting to test takers. *Education Week*, *28*(13), 1–4.

Baenen, N., Ives, S., Lynn, A., Warren, T., Gilewicz, E., & Yaman, K. (2006). *Effective practices for at-risk elementary and middle school students* (E&R No. 06.03). Raleigh, NC: Wake County Public School System.

Baker, E. L., & Linn, R. L. (2003). Validity issues for accountability systems. In S. H. Fuhrman & R. F. Elmore (Eds.), *Redesigning accountability systems for education* (pp. 47–72). New York: Teachers College Press.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B, 57*(1), 289–300.

Bennett, R. E. (2002). *Using electronic assessment to measure student performance*. Princeton, NJ: National Governors Association.

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan International, 80*(2). Retrieved December 14, 2010, from http://blog.discoveryeducation.com/assessment/files/2009/02/blackbox_article.pdf

Blank, R. K., Porter, A. C., & Smithson, J. (2001). *New tools for analyzing teaching, curriculum and standards in mathematics and science: Results from the survey of enacted curriculum project. Final report.* Madison, WI: University of Wisconsin, Wisconsin Center for Education Research.

Bloom, H. S. (2006). *The core analytics of randomized experiments for social research* (MDRC Working Papers on Research Methodology). Retrieved December 14, 2010, from http://www.mdrc.org/publications/437/full.pdf

Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2007). Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. *Educational Evaluation and Policy Analysis, 29*(1), 30–59.

Borman, G., Carlson, D., & Robinson, M. (2010). *The district-level achievement impacts of benchmark assessments: Year 1 outcomes of CDDRE*. Paper presented at the Institute of Education Sciences Research Conference, National Harbor, MD. Retrieved December 14, 2010, from http://ies.ed.gov/director/conferences/10ies_conference/slides.asp?ppt=borman

Borman, G. D., Slavin, R. E., Cheung, A. C. K., Chamberlain, A. M., Madden, N. A., & Chambers, B. (2007). Final reading outcomes of the national randomized field trial of Success for *All. American Educational Research Journal, 44*(3), 701–731.

Boruch, R. F., & Gomez, H. (1977). Sensitivity, bias, and theory in impact evaluations. *Professional Psychology, 8*(4), 411–434.

Brown, R. S., & Coughlin, E. (2007). *The predictive validity of selected benchmark assessments used in the Mid-Atlantic Region* (Issues and Answers Report, REL 2007–017). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic. Retrieved December 14, 2010, from http://ies.ed.gov/ncee/edlabs/regions/midatlantic/pdf/REL_2007017_sum.pdf

Clarke, B. (2006). Breaking through to reluctant readers. *Educational Leadership, 63*(5), 66–69.

Cohen, J. (1988). *Statistical power for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing-data procedures. *Psychological Methods, 6*(4), 330–351.

Cordray, D. S. (2010, June 1). Analyses of potential control group contamination in the REL Midwest's Measures of Academic Progress (MAP) program study. Unpublished manuscript. Vanderbilt University, Nashville, TN.

Cordray, D. S., & Jacobs, N. (2005, June). *Treatment fidelity and core components in school-based ATOD prevention programs.* Paper presented at the Annual Meeting of the Society for Prevention Research, Washington, DC.

Cordray, D. S., & Pion, G. M. (2006). Treatment strength and integrity: Models and methods. In R. Bootzin & P. McKnight (Eds.), *Strengthening research methodology: Psychological measurement and evaluation*, 103–124. Washington, DC: American Psychological Association.

Council of Chief State School Officers. (2007). *Formative assessment and CCSSO: A special initiative—A special opportunity.* Washington, DC: Author.

Cronin, J., Kingsbury, G. G., Dahlin, M., Adkins, D., & Bowe, B. (2007, April). *Alternate methodologies for estimating state standards on a widely-used computer adaptive test.* Paper presented at the Annual Conference of the American Educational Research Association, Chicago.

Dane, A. V., & Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: Are implementation effects out of control? *Clinical Psychology Review, 18*(1), 23–45.

Decker, G. (2003). Using data to drive student achievement in the classroom and on high-stakes tests. *T.H.E. Journal, 30*(6). Retrieved December 14, 2010, from http://www.thejournal.com/articles/16259

Dunn, K. E., & Mulvenon, S. W. (2009). A critical review of research on formative assessment: The limited scientific evidence of the impact of formative assessment in education. *Practical Assessment, Research and Evaluation, 14*(7), 1–11.

Dusenbury, L., Brannigan, B., Falco, M., & Hansen, W. B. (2003). A review of research on fidelity of implementation: Implications for drug abuse prevention in school settings. *Health Education Research, 18*(2), 237–256.

Enders, C. K., & Tofighi, D. (2007). Center predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods, 12*(2), 121–128.

Fuchs, L. S., & Fuchs, D. (1986). Effects of systematic formative evaluation: A meta-analysis. *Exceptional Children, 53*(3), 199–208.

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science, 8*(3), 206–213.

Hall, T. (2002). *Differentiated Instruction* [Monograph]. Washington, DC, National Center on Accessing the General Curriculum. Retrieved August 30, 2006 from http://www.cast.org/system/galleries/download/ncac/DifInstruc.pdf

Hansen, W. B., Collins, L. M., Malotte, C. K., Johnson, C. A., & Fielding, J. E. (1985). Attrition in prevention research. *Journal of Behavioral Medicine, 8*(3), 261–275.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112.

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

Hedges, L. V., & Hedberg, E. (2007). Intraclass correlation values for planning group randomized trials in education. *Educational Evaluation and Policy Analysis, 29*, 60–87.

Henderson, S., Petrosino, A. Guckenburg, S., & Hamilton, S. (2007a). *Measuring how benchmark assessments affect student achievement* (Issues and Answers Report, REL 2007–039). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northeast and Islands. Retrieved December 14, 2010, from http://ies.ed.gov/ncee/edlabs/regions/northeast/pdf/REL_2007039_sum.pdf

Henderson, S., Petrosino, A., Guckenburg, S., & Hamilton, S. (2007b). *A second follow-up year for "Measuring how benchmark assessments affect student achievement"* (REL Technical Brief, REL Northeast and Islands 2007–002). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northeast and Islands. Retrieved

December 14, 2010, from http://ies.ed.gov/ncee/edlabs/regions/northeast/pdf/techbrief/tr_00208.pdf

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*(396), 945–960.

Hulleman, C., & Cordray, D. S. (2009). Moving from the lab to the field: The role of fidelity and achieved relative intervention strength. *Journal of Educational Effectiveness, 2*(1), 88–110.

Hunt, E., & Pellegrino, J. W. (2002). Issues, examples, and challenges in formative assessment. *New Directions for Teaching and Learning, 89*, 73–85.

Kingston, N., & Nash, B. (2009, April). *The efficacy of formative assessment: a meta-analysis*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Lipsey, M. W. (1990). *Design sensitivity: Statistical power for experimental research.* Newbury Park, CA: Sage.

McTighe, J. & Brown, J. L. (2005) Differentiated instruction and educational standards: Is detente possible? *Theory Into Practice* 44(3), 234-244.

Meisels, S. J, Atkins-Burnett, X., Xue, Y., Bickel, D., Son, S., & Nicholson, J. (2003). Creating a system of accountability: The impact of instructional assessment on elementary children's achievement test scores. *Education Policy Analysis Archives, 11*(9). Retrieved December 14, 2010, from http://epaa.asu.edu/epaa/v11n9/

Mowbray, C. T., Holter, M. C., Teague, G. B., & Bybee, D. (2003). Fidelity criteria: Development, measurement, and validation. *American Journal of Evaluation, 24*(3), 315–340.

Northwest Evaluation Association. (2005). *RIT scale norms for use with achievement level tests and measures of academic progress.* Lake Oswego, OR: Author.

Northwest Evaluation Association. (2008). *Teacher hand-book: Measures of academic progress (MAP).* Lake Oswego, OR: Author. Retrieved December 14, 2010, from http://www.nwea.org/sites/www.nwea.org/files/resources/Teacher%20Handbook_0.pdf

Northwest Evaluation Association. (2009). *Technical manual for Measures of Academic Progress™ and Measures of Academic Progress for primary grades™.* Lake Oswego, OR: Author.

Nyquist, J. (2003). *Reconceptualizing feedback as formative assessment: A meta-analysis.* Unpublished master's thesis, Vanderbilt University, Nashville, TN.

O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K–12 curriculum intervention research. *Review of Educational Research, 78*(1), 33–84.

Olson, A. (2007). Growth measures for systemic change. *School Administrator, 64*(1), 10.

Orwin, R. G., Sonnefeld, L. J., Cordray, D. S., Pion, G. M., & Perl, H. I. (1998). Constructing quantitative implementation scales from categorical service data: Examples from a multisite evaluation. *Evaluation Review, 22*(2), 245–288.

Perie, M., Marion, S., & Gong, B. (2007). *The role of interim assessments in a comprehensive assessment system*. Washington, DC: Aspen Institute.

Puma, M. J., Olsen, R. B., Bell, S. H., & Price, C. (2009). *What to do when data are missing in group randomized controlled trials* (NCEE 09-0049). Washington, DC: U.S. Department of Education, Institute of Education Science, National Center for Education Evaluation and Regional Assistance.

Raghunathan, T., Lepkowski, J., van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology, 27*(1), 85–95.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park, CA: Sage.

Regents of the University of Michigan. (2001). *Study of instructional improvement: Teacher Questionnaire 2000–2001.* Ann Arbor, MI: Author. Retrieved December 1, 2011, from http://www.sii.soe.umich.edu/documents/surveys/Teacher%20Questionnaire%202000 - 2001.pdf

Regional Educational Laboratory Midwest at Learning Point Associates. (2008). *REL Midwest task 1.1: Regional education needs analysis* (Year 2 report). Unpublished manuscript. Naperville, IL: Author.

Rowan, B., Camburn, E., & Correnti, R. (2004). *Using teacher logs to measure the enacted curriculum: A study of literacy teaching in 3rd grade classrooms.* Unpublished manuscript, University of Michigan.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York: Wiley.

Russo, A. (2002). Mixing technology and testing. *School Administrator, 59*(4), 6–12.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147–177.

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research, 33*(4), 545–571.

Schenker, N., Raghunathan, T. E., Chiu, P., Makuc, D. M., Zhang, G., & Cohen, A. J. (2008). *Multiple imputation of family income and personal earnings in the National Health Interview Survey: Methods and examples.* Retrieved December 14, 2010, from http://www.cdc.gov/nchs/data/nhis/tecdoc.pdf

Schochet, P. Z. (2008). *Technical methods report: Guidelines for multiple testing in impact evaluations* (NCEE 08-4018). Washington, DC: U.S. Department of Education, Institute of Education Science, National Center for Education Evaluation and Regional Assistance.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* New York: Houghton Mifflin.

Snijders, T. A. B., & Bosker, R. J. (1999) *Multilevel analysis. An introduction to basic and advanced multilevel modelling.* London: Sage.

Steering Committee of the Delaware Statewide Academic Growth Assessment Pilot. (2007). *Toward a more powerful student assessment system: The evaluation and recommendations of the Delaware Statewide Academic Growth Assessment Pilot.* Dover, DE: Author.

Taylor, B. M., Pearson, P. D., Petersen, D. S., & Rodriguez, M. C. (2003). Reading growth in high poverty classrooms: The influence of teacher practices that encourage cognitive engagement in literacy learning. *Elementary School Journal, 104*(1), 3–28.

Tomlinson, C. A. (2001). *How to differentiate instruction in mixed ability classrooms*. Alexandria, VA: Association for Supervision and Curriculum Development.

Tomlinson, C. A. and McTighe, J. (2006) *Integrating differentiated instruction + understanding by design*. Alexandria: Association for Supervision and Curriculum Development.

Torgesen, J. K., & Miller, D. H. (2009). *Assessments to guide adolescent literacy instruction.* Portsmouth, NH: RMC Research, Center on Instruction.

Waltz, J., Addis, M. E., Koerner, K., & Jacobson, N. S. (1993). Testing the integrity of a psychotherapy protocol: Assessment of adherence and competence. *Journal of Consulting and Clinical Psychology, 61*(4), 620–630.

What Works Clearinghouse. (2010). *WWC evidence review protocol for K–12 students with learning disabilities, Version 2.0.* Princeton, NJ: Author. Retrieved September 21, 2011, from http://ies.ed.gov/ncee/wwc/PDF/sld_protocol_v2.pdf

Woodfield, K. (2003). Getting on board with online testing. *T.H.E. Journal, 30*(6), pp. 32, 34–37.

Yeaton, W. H., & Sechrest, L. (1981). Critical dimensions in the choice and maintenance of successful treatments: Strength, integrity, and effectiveness. *Journal of Consulting and Clinical Psychology, 49*(2), 156–167.


DEPARTMENT OF EDUCATION
UNITED STATES OF AMERICA