NCEE 2012-4014

U. S. DEPARTMENT OF EDUCATION

A Study of the Effectiveness of a School Improvement Intervention (Success in Sight)

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

A Study of the Effectiveness of a School Improvement Intervention (Success in Sight)

Final Report

February 2012

Authors:

Stephanie B. Wilkerson, Ph.D., Magnolia Consulting

Lisa C. Shannon, Ph.D., Magnolia Consulting

Mary K. Styers, Ph.D., Magnolia Consulting

Billie-Jo Grant, Ph.D., Magnolia Consulting

Project Officer:

Sandra Garcia
Institute of Education Sciences

NCEE 2012-4014
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
Rebecca A. Maynard
Commissioner

February 2012

This report was prepared for the National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences (IES), under contract ED-06C0-0023 with Regional Educational Laboratory Central administered by Mid-continent Research for Education and Learning.

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

This report is available on the Institute of Education Sciences website at http://ncee.ed.gov and the Regional Educational Laboratory Program website at http://edlabs.ed.gov.

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

Acknowledgments

A study of this size and nature requires the collaboration and contributions of multiple people including educators, students, researchers, technical advisers, reviewers, and editors. The study would not be possible without them, and the authors extend much gratitude to the following individuals for their voluntary participation and efforts.

The authors thank the school principals, leadership team members, and teachers who went beyond their daily call of duty to children by participating in a research study and who willingly and graciously gave their time to complete the data collection. Their contributions were invaluable to the study. The authors greatly appreciate the support provided by school districts and the Minnesota and Missouri departments of education. Their commitment to the study and their time in providing student assessment data reflect their ongoing dedication to supporting high-quality research to promote school improvement.

The authors gratefully acknowledge the contributions of Dr. Elizabeth Palmer, president of ASPEN Associates, Inc., and Suzanne Joyce, program coordinator for ASPEN Associates, Inc., who developed the research design and instrumentation, recruited participants, conducted data collection, and provided project management through the first year and a half of the study.

The authors thank Helen Apthorp, who served as Mid-continent Research for Education and Learning's (McREL) research liaison and manager of the research subcontract, and Ceri Dean and Danette Parsley of McREL for their contributions to defining the theoretical foundation underpinning Success in Sight. The authors are also grateful for the report review provided by Kirsten Miller, Bruce Randel, and Jean Williams of McREL as well as Louis Cicchinelli, who also guided initial project design and management.[1] The authors thank Success in Sight facilitators for providing implementation fidelity data throughout the study period.

The authors express gratitude to members of the study's Technical Working Group, who provided technical consultation and advice throughout the study period: William Clune, Norman Webb, Geoffrey Borman, Robin Morris, Andrew Porter, and Bob St. Pierre. Reviewers at the Analytical and Technical Support contract administered by Mathematica Policy Research Inc. also provided valuable technical consultation and feedback for which the authors are grateful.

The authors extend special thanks to research team members and consultants who collaborated on this project and whose dedication and contributions made this a successfully implemented study: Sara Bullard, Lauren Funk, Julius Najab, deKoven Pelton, Candace Rowland, Monica Savoy, Becky Siceloff, and Arianne Welker.

[1] See Appendix A for firewall procedures related to report review.

Disclosure of potential conflicts of interest

Mid-continent Research for Education and Learning (McREL) explicitly addressed the concern that there could be a perception of a potential conflict of interest regarding this REL Central-supported randomized controlled trial in our Conflict of Interest Plan submitted to the U.S. Department of Education on April 23, 2008. We disclosed this area of potential conflict regarding this randomized controlled trial because McREL is both the developer of *Success in Sight* and the recipient of the REL Central contract. In order to mitigate against this perception, we proposed to subcontract the conduct of the actual evaluation to Magnolia Consulting (located in Charlottesville, Virginia) and install a firewall by establishing formal procedures for communication among McREL, Magnolia, and the U.S. Department of Education regarding study design, analysis, and report issues. This firewall plan is detailed in appendix A of this final report. IES approved both the subcontracting relationship and the firewall as appropriate strategies for addressing the perception of a potential conflict of interest. In addition, we should note that none of the report authors or study staff from Magnolia Consulting has a financial interest in the reported outcomes of this study.

Contents

Figures

Tables

Executive summary

Since the passage of the No Child Left Behind (NCLB) Act of 2001 and its adequate yearly progress (AYP) requirements, the nation's education systems have increased their focus on school improvement interventions that build school and teacher capacity to increase student achievement in reading and mathematics. Despite the intensified focus on school improvement, only 70 percent of schools made AYP in reading and mathematics in 2008 (U.S. Department of Education 2008a). Failing to make AYP in reading or mathematics has important implications for schools, such as risk of closure or restructuring. The challenges preventing low-performing schools from making AYP are rarely singular or simple and call for proven systemic and sustainable interventions (Kutash, Nico, Gorin, Rahmatullah, and Tallant 2010).

Systemic interventions aim to improve school and teacher capacity and increase student achievement by focusing on various parts of an education system, such as professional development, student assessment, curriculum and instruction, and school leadership and support (Clune 1998; Supovitz and Taylor 2005). Because these parts of an education system are interrelated, creating and sustaining change in one part of the system often catalyze or require changes throughout the rest of the system. When implemented effectively, systemic change can lead to positive gains in student reading achievement (Wolf 2007) and mathematics achievement (Clune 1998; Kim and Crasco 2003; Wolf 2007). As systemic interventions build schools' and teachers' capacities to increase student achievement, the likelihood of schools improving their performance to make AYP also increases (Hallinger and Heck 2010).

Sensing a need for systemic improvement interventions, state departments of education in the Central Region (Colorado, Kansas, Missouri, Nebraska, North Dakota, South Dakota, and Wyoming) began to request research-based information and technical assistance on systemic change to address the increasing number of schools failing to make AYP. In addition, Mid-continent Research for Education and Learning (McREL), which provides research and technical assistance to the Central Region, identified the need for a systemic approach to address schools' varying needs by strengthening teacher quality, using research-based classroom practices, preparing students adequately for the workforce or postsecondary education, using technology to enhance instruction, and recruiting and retaining teachers. McREL responded to the needs of both state departments of education and schools by developing Success in Sight, a systemic school improvement intervention. Since 2000, McREL has implemented Success in Sight in schools across the country.

To provide rigorous evidence of the effectiveness of Success in Sight, McREL contracted with independent researchers under its regional educational laboratory (REL) contract with the Institute of Education Sciences (IES) to conduct the first cluster randomized trial of the intervention. The study took place during the 2008/09 and 2009/10 school years in 52 schools in two states.

Success in Sight overview

Success in Sight focuses on the interrelated parts of an education system. This systemic school improvement intervention is designed to address schools' specific needs while building their capacities to plan, implement, and evaluate school improvement practices. It is intended to help

schools, leadership teams, and teachers systematically and systemically engage in continuous school improvement practices to advance the learning of all students (Cicchinelli et al. 2006). Success in Sight facilitators work directly with school leadership teams, which comprise five to seven members, including the principal, teachers, and other staff. As leadership teams increase their capacities for implementing school improvement practices, they expand their efforts to include more teachers. As teachers collaborate with leadership team members in planning and implementing Success in Sight school improvement practices, it is expected that they will also increase their capacities for carrying out school improvement practices, thus increasing schoolwide capacity.

The program is based on school improvement research (Marzano 2000; Marzano, Waters, and McNulty 2005) and targets five main school capacity-building areas:

- Data-based decisionmaking—collecting, analyzing, interpreting, and using data to inform decisions and to establish and monitor goals for improvement at the individual student and school levels.
- Purposeful community—forming and sustaining a community that identifies with and works collectively toward important outcomes, uses all available resources effectively, operates from a set of agreed-upon processes that guide actions and decisions in the school, and shares a collective belief that the community can accomplish its goals (collective efficacy).
- Shared leadership—participating in a process of mutual influence, responsibility, and accountability for achieving collective, organizational goals for school improvement.
- Research-based practices—adopting practices that directly address factors shown to be associated with improved student achievement and that are based on scientific evidence of effectiveness.
- Continuous improvement process—employing a five-stage process to improve student performance by taking stock of the current situation, focusing on the right solution, taking collective action, monitoring progress and adjusting efforts, and maintaining momentum for improvement efforts.

McREL facilitators deliver Success in Sight capacity-building content to school leadership teams through four components: six large-group professional development sessions with consortia of schools, 10 onsite mentoring sessions with leadership teams, distance learning and support, and fractal improvement experiences (projects that build team capacity while addressing specific school needs).

The Success in Sight large-group professional development component is delivered by McREL facilitators over two days, three times a year. During each of the six sessions, which occur with a consortium of leadership teams in the same geographic area, McREL facilitators intend to increase the knowledge and skills of school leadership teams in the five capacity-building areas described above.

The Success in Sight onsite mentoring and support component occurs between the large-group professional development sessions. Specifically, McREL facilitators conduct 10 onsite visits to support leadership teams as leadership team members apply lessons from the professional development sessions. Each onsite meeting is tailored to each school's needs and priorities.

The Success in Sight distance support component occurs with leadership teams between large-group development sessions. McREL facilitators provide leadership teams with ongoing support through phone conferences and email exchanges as the teams implement the continuous improvement process.

The final component of Success in Sight is fractal improvement experiences. Fractal improvement experiences are change initiatives related to student achievement that are identified, planned, and implemented by leadership team members using what they learn during the large-group professional development sessions. Fractal improvement experiences can address a variety of focus areas based on a school's specific needs, such as school culture, parent involvement, and student engagement, but most often they focus on reading and mathematics content areas. Initial fractal improvement experiences are small, intended to result in quick successes in order to build leadership team members' sense of collective efficacy (that is, a belief that by working together they can make a difference in student achievement). The onsite mentoring and distance support provided by facilitators are intended to expand leadership team members' capacities to increase the scope of the fractal improvement experiences and involve increasing numbers of teacher participants. As teachers become involved in fractal improvement experiences, it is expected that they develop their capacity for data-based decisionmaking, purposeful community, shared leadership, research-based strategies, and the continuous improvement process. The fractal improvement experiences, in turn, are intended to result in an increased schoolwide capacity to enact school improvement initiatives using the five Success in Sight areas. Ultimately, the intended result is higher student achievement schoolwide.

Schools have used Success in Sight over the past 11 years to facilitate school improvement efforts. However, there have been no cluster randomized trials to provide causal evidence regarding its effectiveness in improving student and teacher outcomes. Therefore, the main purpose of this study was to provide unbiased estimates of the impact of Success in Sight on student academic achievement in reading or mathematics. The achievement outcome areas of reading and mathematics were chosen for this study based on the NCLB mandate that all students should be proficient in reading and mathematics by 2014. Additionally, all states assess reading and mathematics achievement in grades 3–5, which are the focus of this study. The study also sought to provide an unbiased estimate of the effects of Success in Sight on teacher capacity for school improvement practices related to data-based decisionmaking, purposeful community, and shared leadership.

Research questions

The primary research questions focus separately on reading and mathematics student achievement outcomes:

1. Does implementation of Success in Sight have a significant impact on student achievement in reading?

2. Does implementation of Success in Sight have a significant impact on student achievement in mathematics?

Answers to these primary research questions will be the basis for conclusions about the effectiveness of Success in Sight and are based on study findings related to student achievement outcomes in reading or mathematics.

The secondary research questions focus separately on teacher capacity related to data-based decisionmaking, purposeful community, and shared leadership:

1. Does implementation of Success in Sight have a significant impact on teacher capacity for data-based decisionmaking?

2. Does implementation of Success in Sight have a significant impact on teacher capacity for purposeful community practices?

3. Does implementation of Success in Sight have a significant impact on teacher capacity for shared leadership?

Finally, the study included exploratory research questions to examine the empirical relationship between teacher capacity and student achievement outcomes. The exploratory research questions focus separately on reading and mathematics student achievement outcomes as they each relate to data-based decisionmaking, purposeful community, and shared leadership:

1. What is the relationship between teacher capacity for data-based decisionmaking and student achievement in reading?
2. What is the relationship between teacher capacity for data-based decisionmaking and student achievement in mathematics?
3. What is the relationship between teacher capacity for purposeful community practices and student achievement in reading?
4. What is the relationship between teacher capacity for purposeful community practices and student achievement in mathematics?
5. What is the relationship between teacher capacity for shared leadership and student achievement in reading?
6. What is the relationship between teacher capacity for shared leadership and student achievement in mathematics?

Study timeline

The activities for this study occurred from September 2007 to June 2010. School recruitment occurred from September 2007 until July 2008. Implementation of Success in Sight occurred during the 2008/09 and 2009/10 school years. Baseline data collection occurred from March 2008 through August 2008, and posttest data collection occurred from March 2010 through June 2010.

Study sample

This study's target population was low- to moderate-performing elementary schools located in states served by McREL under its Regional Educational Laboratory (REL) contract from the U.S. Department of Education's Institute of Education Sciences (IES) and Comprehensive

Center grant from the U.S. Department of Education's Office of Elementary and Secondary Education. Low- to moderate-performing schools were defined as schools that did not make AYP for any of the three school years prior to the 2008/09 school year or were at risk of not making AYP as reported by school personnel. Among the states served by McREL's regional programs, Minnesota, Colorado, Missouri, and Kansas had the most schools that did not make AYP in 2004/05. From this set of four states, recruitment efforts for this study focused on Minnesota and Missouri.

School recruitment efforts yielded 52 participating schools (26 treatment schools and 26 control schools) in eight districts. Researchers assigned participating schools to matched pairs based on their 2006 mean school reading scores and the percentage of students qualifying for free or reduced-price lunch. Within each matched pair, one school was randomly assigned to participate in the Success in Sight intervention (as a treatment school), and the other school was assigned to conduct business as usual (as a control school). Within participating schools at baseline, there were 8,467 students with reading achievement scores, 8,331 students with mathematics achievement scores, and 1,374 teacher participants. At posttest, there were 8,182 students with reading achievement scores, 8,213 students with mathematics achievement scores, and 1,516 teacher participants. These sample sizes yielded enough statistical power (that is, greater than 0.80) to detect an effect size of 0.20 for the benchmark impact estimates regarding the primary student achievement outcomes and an effect size of 0.30 for the benchmark impact estimates regarding secondary outcomes related to teacher capacity for school improvement practices.

Researchers conducted preliminary analyses to examine the baseline equivalence of treatment and control groups on reading and mathematics tests, student demographic characteristics, mean baseline teacher capacity for school improvement practice scores, teacher demographic characteristics, and general school characteristics (such as school size). These analyses revealed no statistically significant differences between treatment and control groups.

For this study, student participants included students in grades 3–5 with available baseline or posttest achievement data on reading and/or mathematics state assessments. Including students in these grades enabled the use of existing data from state-administered reading and mathematics assessments, which reduced the data collection burdens for participating schools. Student baseline reading and mathematics scores were used to compute mean school-level baseline achievement covariates, and student posttest reading and mathematics scores served as outcome data. The student sample for the benchmark impact estimate of primary outcomes included students who were in grades 3–5 at posttest with available outcome data. The teacher survey participants included leadership team members, classroom teachers, and specialists with appointments of 0.50 full-time equivalent or greater at their schools. These teachers were included because they were in a position to participate in and implement school improvement practices directly with students. Available teacher baseline school improvement practice scores were used to compute mean school-level baseline capacity for school improvement covariates, and available teacher posttest school improvement practice scores served as outcome data.

Implementation

Eight criteria were developed to gauge fidelity of Success in Sight delivery and participation across 26 treatment schools for the 2008/09 and 2009/10 school years. Four criteria focus on

McREL facilitators' fidelity to delivering Success in Sight as intended by conducting six large-group professional development sessions, implementing a minimum of 80 percent of a content module at each session, conducting 10 onsite mentoring and distance support sessions with leadership teams, and providing 10 onsite mentoring sessions with principals. Four criteria focus on school participation requirements: forming leadership teams with a minimum of five members representing different student support and instructional areas, attending the six large-group professional development sessions, attending 10 onsite mentoring sessions, and completing at least two fractal experiences involving participants not on leadership teams. Success in Sight facilitators' program records and electronic logs provided the data used to assess adequate program delivery and participation.

Success in Sight facilitators and all 26 treatment schools met the eight implementation fidelity indicators for this study. All treatment schools formed leadership teams with at least five members, including the principal and staff representing two or more grades and services for student subgroups. Of the required 130 leadership team members (five per team), 97.69 percent of them attended all six large-group professional development sessions at which Success in Sight facilitators delivered a minimum of 80 percent of each program module (one module per session, six modules total). Success in Sight facilitators provided 10 of 10 onsite mentoring sessions to the 26 schools in which 100 percent of leadership team members and 96 percent of principals attended. All principals in each treatment school received at least 9 of 10 one-on-one mentoring sessions with a Success in Sight facilitator during these site visits. All treatment schools completed a minimum of two fractal improvement experiences that involved participants not on leadership teams.

As part of the Success in Sight fractal improvement experiences, leadership team members and school staff applied lessons from the large-group professional development sessions regarding data-based decisionmaking, purposeful community, shared leadership, research-based practices, and the continuous improvement process. Twenty-six treatment schools completed three to eight fractal improvement experiences across schools (mean = 5.46, standard deviation = 1.48) focusing on salient local issues that included a range of 7–115 participants across schools (mean = 29, standard deviation = 15.26). Of the 142 fractal improvement experiences completed across 26 schools, 39 focused specifically on reading (27.46 percent), and 26 focused specifically on mathematics (18.31 percent). The other 77 (54.23 percent) focused on broader areas of student achievement, such as teacher professional development, school culture, data-based decisionmaking, student behavior and engagement, and parent involvement. Of the 26 treatment schools, 10 focused 50 percent or more of their fractal improvement experiences on reading exclusively or mathematics exclusively with the majority focused on reading, 10 focused 50 percent or more of their fractal improvement experiences on both reading and mathematics, and 6 focused 50 percent or more of their fractal improvement experiences on multiple areas not directly targeting reading or mathematics, such as student behavior, school culture, parent involvement, and teacher professional development.

Treatment and control schools had leadership teams prior to the study and participated in other education initiatives as part of their school improvement process during the two-year study period. In control schools, this was considered "business as usual," as their participation in the study did not require that they conduct specific or formal school improvement initiatives, but rather continue with current and planned efforts. In treatment schools, Success in Sight is meant

to supplement rather than supplant other school improvement initiatives. Through fractal improvement experiences, leadership teams can focus on implementing, evaluating, and improving other initiatives, such as those involving curriculum and assessment.

Based on interview feedback from 155 school representatives and published time estimates from curriculum developers, professional development opportunities over the two study years involved comparable amounts of time whether schools participated in Success in Sight (26 treatment schools; 166 hours), professional learning communities in Missouri (7 treatment and 8 control schools; 192 hours), or leadership academies in Minnesota (three or fewer treatment and 6 control schools; 168 hours). Of the 28 Missouri schools participating in the study, 8 treatment schools and 3 control schools received professional development services from the Regional Professional Development Centers. All treatment and control schools in Missouri implemented Reading First and response to intervention during the study period. In Minnesota, all 24 treatment and control schools implemented the Mondo literacy program and the Phonological Awareness Literacy Screening assessment. Despite the similarity in amount of professional development time, no professional development programs at control schools consisted of systemic school improvement interventions similar to Success in Sight during the study period.

Measures and data collection

This study's impact analyses of primary outcomes examined the effect of Success in Sight on student achievement in grades 3–5, as measured by reading and mathematics state assessments, the Minnesota Comprehensive Assessment II and the Missouri Assessment Program, in 2008 and 2010. The study's impact analyses of secondary outcomes examined the effects on teacher capacity for school improvement practices, as measured by a teacher survey administered in 2008 and 2010. The teacher survey used in this study was derived from two existing surveys: the Teacher Survey of Policies and Practices (Mid-continent Research for Education and Learning 2005) and the 12-item Collective Efficacy Scale (Goddard 2002). The intended school improvement practice outcomes in this study were data-based decisionmaking, purposeful community, and shared leadership. Two of the four Teacher Survey of Policies and Practices scales (professional community and leadership), one of its subscales (assessment and monitoring), and the Collective Efficacy Scale were used to measure the three intended capacities for school improvement practices outcomes. Throughout the study, researchers also collected program records and implementation logs from professional development facilitators to document Success in Sight delivery and participation in treatment schools. These records and logs included information about participant membership, attendance, delivery of professional development content, and fractal improvement experiences and focus areas. In addition, researchers collected interview and focus group data to provide information about the local contexts of the treatment and control schools.

Data collection occurred from March 2008 through August 2010. Baseline student achievement data were collected from March 2008 through May 2008, and baseline teacher survey data were collected from June 2008 through October 2008. The extended survey administration period provided time to identify site coordinators and administer the survey when school was in session rather than during the summer. Baseline principal interviews and school focus groups were conducted from September 2008 through October 2008. Posttest student achievement data were collected from March 2010 through May 2010, posttest teacher survey data were collected from

March 2010 through April 2010, and posttest phone interviews were conducted from April 2010 through June 2010.

Analyses and results

This study's impact analyses examined the effect of Success in Sight on student achievement in reading or mathematics after two years, which was the length of the Success in Sight intervention. Researchers ran separate multilevel models for each student achievement outcome. The achievement test scores were transformed into *z*-scores to make the data from the two different state assessments more comparable. Separate transformations were conducted for each grade, state, and assessment content area. For each student in the study sample, researchers subtracted the appropriate grade-level state mean from each student's corresponding reading or mathematics scale score and divided it by the relevant standard deviation.

The rate of missing data on the outcome measures was less than 5 percent. Results indicated that Success in Sight did not have a statistically significant impact on student achievement in either reading (adjusted posttest mean difference = –0.01, standard error = 0.03, $p = .75$) or mathematics (adjusted posttest mean difference = –0.06, standard error = 0.04, $p = .10$).

Researchers conducted sensitivity analyses to test the robustness of the benchmark estimates to the use of a baseline achievement covariate, to the way the student sample was defined, and to the impact analysis methods combining data across states. Omitting the baseline cluster-level covariate and estimating impacts separately by state and, subsequently, combining the state-level results meta-analytically yielded results consistent with the benchmark impact estimates. The sensitivity analysis that included only student stayers (that is, students enrolled in study schools in grade 3 at 2008 baseline data collection and grade 5 at 2010 posttest data collection who did not change schools over the course of the study) also were consistent with the benchmark estimate of impacts of Success in Sight on student reading achievement, but generated estimates of statistically significant, negative impacts on posttest mathematics scores. Specifically, the sensitivity analysis of mathematics achievement data indicated that Success in Sight had a statistically significant negative impact on mathematics achievement (adjusted posttest mean difference = –0.11, standard error = 0.04, $p = .02$.), with student stayers in treatment schools demonstrating average posttest mathematics achievement lower than that of student stayers in control schools. Although a sensitivity analysis with a student sample comprised of stayers and within-study in-movers (that is, students who were enrolled in grades 1 and 2 at baseline and remained in the same school throughout the study) would have been useful, researchers did not have access to baseline enrollment rosters of grade 1 and 2 students, which made it impossible to identify within-study in-movers.

The study also included impact analyses of secondary outcomes to examine the effect of Success in Sight on teacher capacity for school improvement practices (that is, data-based decisionmaking, purposeful community, and shared leadership) after two years. Researchers ran separate multilevel models for each secondary outcome. The outcome variables were mean posttest scores for teacher capacity for data-based decisionmaking, purposeful community, and shared leadership. The teacher sample included leadership team members, classroom teachers, and specialists with appointments of 0.50 full-time equivalent or greater at that school who had

available data. Wave nonresponse led to missing data for less than 5 percent of teachers for the impact analysis sample, and cases with missing outcome measures were excluded from analyses.

Results indicated that Success in Sight did not have a statistically significant impact on teacher capacity for data-based decisionmaking (adjusted posttest mean difference = 0.03, standard error = 0.02, $p = .13$), purposeful community (adjusted posttest mean difference = 0.03, standard error = 0.04, $p = .49$), or shared leadership (adjusted posttest mean difference = 0.16, standard error = 0.07, $p = .02$, which is not significant after applying the Benjamini-Hochberg correction for multiple comparisons). The sensitivity analyses with no baseline covariates supported these findings.

Finally, the study's analyses included exploratory analyses to examine the hypothesized relationship between the study's intermediate outcomes—teacher capacity for school improvement practices in data-based decisionmaking, purposeful community, and shared leadership—and student achievement in reading and mathematics. These results revealed a statistically significant negative association between teachers' posttest ratings of their capacity for shared leadership and posttest student reading achievement ($p = .03$). Neither teacher capacity for data-based decisionmaking nor purposeful community was statistically significantly associated with posttest student reading achievement ($p = .60$ and $p = .77$, respectively). For mathematics achievement, there was a statistically significant negative association between teachers' posttest ratings of their capacity for data-based decisionmaking and shared leadership and posttest student mathematics achievement ($p = .04$ and $p < .01$, respectively), indicating that higher ratings of teacher capacity in data-based decisionmaking was statistically significantly associated with lower student mathematics scores, and higher ratings of teacher capacity in shared leadership was statistically significantly associated with lower student mathematics scores. Findings also revealed a statistically significant positive association between teachers' posttest ratings of their capacity for purposeful community and posttest student mathematics achievement ($p < .01$), indicating that higher ratings of teacher capacity in purposeful community was statistically significantly associated with higher student mathematics scores. It was not within the scope of these exploratory analyses to generate explanations of the associations between teachers' self-reported ratings of their capacity for data-based decisionmaking, purposeful community, or shared leadership and students' reading and mathematics achievement.

Conclusions

This study was the first cluster randomized trial to examine the effectiveness of Success in Sight on primary outcomes—student achievement in reading and mathematics—and intermediate teacher outcomes—capacity for school improvement practices in data-based decisionmaking, purposeful community, and shared leadership.

The results of the benchmark analyses revealed that Success in Sight did not have a statistically significant impact on student achievement in reading or mathematics or on teacher capacity for school improvement practices in data-based decisionmaking, purposeful community, or shared leadership.

Although this study used rigorous methodology, readers should consider findings in the context of its limitations. One limitation is that the study used a volunteer sample of low- to moderate-

performing schools in Minnesota and Missouri. Therefore, the results do not generalize to schools that differ systematically from this specific sample of schools. In addition, because the study assessed only reading and mathematics at grades 3–5 using state assessments, the study's findings are not generalizable to other content areas, grades, or assessments. Furthermore, the study findings do not generalize to schools that implement Success in Sight for more than two years. The study also had limitations related to how teacher capacity outcomes were measured. Data from the teacher practice impact analyses were based entirely on teacher self-report collected through an online survey.

Chapter 1. Introduction and study overview

Since the passage of the No Child Left Behind (NCLB) Act of 2001 and its adequate yearly progress (AYP) requirements, the nation's education systems have been increasingly focused on school improvement interventions that build school and teacher capacity to increase student achievement in reading and mathematics. Despite the intensified focus on school improvement, only 70 percent of schools made AYP in reading and mathematics in 2008 (U.S. Department of Education 2008a). Failing to make AYP in reading or mathematics has important implications for schools, such as risk of closure or restructuring. The challenges preventing low-performing schools from making AYP are rarely singular or simple and call for proven systemic and sustainable interventions (Kutash, Nico, Gorin, Rahmatullah, and Tallant 2010).

Systemic interventions aim to impact school and teacher capacity and increase student achievement by focusing on various parts of an education system, such as professional development, student assessment, curriculum and instruction, and school leadership and support (Clune 1998; Supovitz and Taylor 2005). Because these parts of an education system are interrelated, creating and sustaining change in one part of the system often catalyzes or requires changes throughout the rest of the system. When implemented effectively, systemic change can lead to positive gains in student reading achievement (Wolf 2007) and mathematics achievement (Clune 1998; Kim and Crasco 2003; Wolf 2007). As systemic interventions build schools' and teachers' capacities to increase student achievement, the likelihood of schools improving their performance to make AYP also increases (Hallinger and Heck 2010).

Mid-continent Research for Education and Learning (McREL) responded to the complex challenges confronting low-performing schools by developing Success in Sight, a systemic school improvement intervention. Success in Sight is designed to address interrelated parts of an education system with the purpose of building schools' and teachers' capacities to increase student achievement. Since 2000, McREL has implemented the Success in Sight systemic school improvement intervention in schools across the country.

In 2008, McREL contracted with independent researchers under its regional educational laboratory contract with the U.S. Department of Education's Institute of Education Sciences (IES) to conduct the first cluster randomized trial to assess the effectiveness of Success in Sight (see appendix A for firewall procedures used to ensure that objective research practices were followed). The study took place during the 2008/09 and 2009/10 school years in 52 schools in two states.

This chapter discusses the study rationale, provides an overview of Success in Sight and its theory of change, and presents a study overview.

Study rationale

In 2005, 21 percent of schools across the seven states served by McREL's regional educational laboratory program (Colorado, Kansas, Missouri, Nebraska, North Dakota, South Dakota, and Wyoming) did not make AYP in student achievement as required by the NCLB Act of 2001 (American Institutes for Research 2005). Despite some improvements in the following years, the number of schools not making AYP continued to grow (Ehlert et al. 2009; Missouri Department

of Elementary and Secondary Education 2009a). These schools face complex challenges in improving student achievement and bringing all students to proficiency in reading and mathematics by 2014. Given the stakes associated with student performance and the impending 2014 deadline, schools do not have the luxury of a trial-and-error approach to school improvement. To meet those challenges, schools need to improve systemically and demonstrate sustained academic progress (Mourshed, Chijioke, and Barber, 2010).

Regional educator needs

McREL identified priority needs in the region by reviewing advisory committee reports, by interviewing the chief state school education officers and key state education agency staff, and by analyzing demographic and education system data. This study responds to the expressed priority needs for research on strengthening teacher quality and on classroom practices.

In addition, based on a need for systemic improvement for the increasing number of Central Region schools failing to make AYP, state departments of education in the region began to request research-based information and technical assistance around systemic change for low-performing districts and schools. This study was developed to help meet Central Region information needs about the effectiveness of a systemic approach to school improvement.

Systemic school improvement

Systemic school improvement interventions focus on building school and teacher capacity to increase student achievement by addressing various interrelated and interdependent components of an education system (Hargreaves, Halász, and Pont 2007). Among other components, these may include a school's curriculum, professional development opportunities, instructional practices, and assessment procedures (Clune 1998; Supovitz and Taylor 2005). Efforts to improve one of the system's components will often instigate changes in other components, as well as changes in the system as a whole. This, in turn, can contribute to greater school and teacher capacities and improvements in student achievement (Hallinger and Heck 2010).

A systemic approach to school improvement considers the local context of education systems and acknowledges that the specific needs, focus areas, and capacities for improvement vary from school to school. Therefore, rather than concentrating on a particular project or narrowly defined prescriptive intervention, effective systemic school improvement interventions have differential emphases on school structures, processes, and capacities depending on particular schools' needs (Herman et al. 2008). This alignment with individual school needs is critical to facilitating change that will lead to sustained student academic growth (Fullan 1999; Hall and Hord 1987). Within a systemic approach to school improvement, districts and schools operate uniquely to organize and facilitate decisionmaking about creating, implementing, and sustaining fundamental school improvement efforts most relevant to their specific needs (Adelman and Taylor 2007).

Implementing systemic change is rarely easy and requires multiple levels of support, as decades of research have shown (Fullan and Steigelbaurer 1991; Sashkin and Egermeier 1993; Massell, Kirst, and Hoppe 1997; Ellsworth 2000). Many school administrators do not have the skills, experience, or time to accomplish the daunting task of school reform. Facilitating the change process involves many individuals at different levels within a school system including district

administrators, principals, and teachers (Goertz, Floden, and O'Day 1996; Datnow, Lasky, and Stringfield 2005). Research suggests that internal or external change agents, or a combination of both, can be effective in assisting schools in building capacity for change and navigating the road to improvement (Hall and Hord 1987; Havelock and Zlotolow 1995; Sun, Creemers, and de Jong 2007; Herman et al. 2008). External pressure and high expectations for student performance from community, state, or national representatives can help catalyze the improvement process. Internal motivators such as empowered school leadership and success with short-term goals can help educators sustain improvement efforts (Fullan 1999).

Program selection

McREL proposed to study the Success in Sight intervention because the program is in widespread use but had not been systematically tested in the field. The Success in Sight intervention is a systemic approach focusing on building capacities across multiple, interconnected areas of school improvement, de-emphasizing intervention ownership and emphasizing collaborative work toward desired outcomes, and providing multiple levels of support to participant teams. By the time it was chosen for study, Success in Sight had been implemented in 60 schools across four states. In development as early as 1995 and fully operational since 2000, the program had reached both urban and rural settings and across all grade levels, for a total of 28 elementary schools, 11 middle schools, 19 high schools, 1 school serving grades K–8, and 1 school serving grades K–12. The program had been implemented in two ways: with McREL acting as the external change facilitator and with McREL training qualified staff at participating schools to act as the change facilitator.

During the 2002/03 school year, McREL field-tested the Success in Sight external change agent model in 12 schools with a one-group pre-post design including rural and urban schools. The percentage of schools making AYP in their focus area (reading or mathematics) was 25.00 percent in 2001/02, 41.66 percent in 2002/03, and 83.33 percent in 2003/04 (Mid-continent Research for Education and Learning n.d.). Although that study was not designed to establish a causal relationship between the Success in Sight intervention and student achievement, its findings did suggest that further investigation was warranted. To this end, McREL contracted with independent researchers to conduct a large-scale, cluster randomized trial to study Success in Sight's impact on student achievement and staff capacity for school improvement practices.

McREL established firewall policies, structures, and procedures to ensure against bias and maintain a separation between the researchers and McREL developers and facilitators. McREL designated a research liaison as the sole point of contact between the researchers and Success in Sight developers and facilitators. This firewall procedure limited communication and prohibited researchers from sharing outcome data with Success in Sight developers and facilitators. The liaison provided researchers with program documentation and records. The firewall was approved by the Institute of Education Sciences and is described in appendix A.

Success in Sight overview and theoretical foundations

Success in Sight, developed by McREL, uses a capacity-building approach to help schools, leadership teams, and teachers systematically and systemically engage in continuous school improvement practices to advance the learning of all students (Cicchinelli et al. 2006). The

intervention focuses on building a culture of shared leadership among school staff to promote collective responsibility for implementing school improvement practices targeting student achievement. Success in Sight facilitators work directly with school leadership teams comprised of five to seven members including the principal, teachers, and other staff. As leadership teams increase their capacities for implementing school improvement practices, they expand their efforts to include more teachers. It is expected that as teachers collaborate with leadership team members in planning and implementing Success in Sight school improvement practices, they also will increase their capacities for carrying out school improvement practices, which are intended to increase schoolwide capacity. The increased capacities at the individual and school levels are expected to mutually support each other and contribute to improved student outcomes.

The program is based on years of school improvement research (Marzano 2000; Marzano, Waters, and McNulty 2005) and aims to build the capacity of schools, leadership teams, and teachers to increase student achievement by targeting five main school capacity-building areas:

- *Data-based decisionmaking*—collecting, analyzing, interpreting, and using data to inform decisions and to establish and monitor goals for improvement at the individual student and school levels.
- *Purposeful community*—forming and sustaining a community that identifies with and works collectively toward important outcomes that matter to all, uses all available resources effectively, operates from a set of agreed-upon processes that guide actions and decisions in the school, and shares a collective belief that the community can accomplish its goals (collective efficacy).
- *Shared leadership*—participating in a process of mutual influence, responsibility, and accountability for achieving collective, organizational goals for school improvement.
- *Research-based practices*—adopting practices that directly address factors shown to be associated with improved student achievement and that are based on scientific evidence of effectiveness.
- *Continuous improvement process*—employing a five-stage process to improve student performance by taking stock of the current situation, focusing on the right solution, taking collective action, monitoring progress and adjusting efforts, and maintaining momentum for improvement efforts.

McREL facilitators deliver Success in Sight capacity-building content to school leadership teams consisting of principals, teachers, and other staff through four components: six large-group professional development sessions with consortia of schools, 10 onsite mentoring sessions with leadership teams, distance learning and support, and fractal improvement experiences (manageable projects that build team capacity while addressing specific school needs). This section describes each capacity building area and delivery component, along with supporting research, then discusses the theory of change involved.

Success in Sight capacity-building areas

Success in Sight aims to build the capacities of schools, leadership teams, and teachers for school improvement practices in five areas. Each area encompasses knowledge and skills deemed essential for focusing on the right problems and solutions and sustaining continuous

improvement, taking into account a school's context, needs, and existing strengths. Each of these five areas is described below.

Data-based decisionmaking—collecting, analyzing, interpreting, and using data to inform decisions and to establish and monitor goals for improvement at the individual student and school levels.

Research has shown that in effective schools, educators collect, analyze, interpret, and use data to identify learning problems and guide improvement efforts at all levels including school, classroom, and individual student levels (Creemers 1994; Teddlie and Reynolds 2000). According to Bernhardt (2003), practitioners can collect and use four categories of data related to student achievement: demographics, programs, teacher perceptions, and student perceptions. Success in Sight facilitators introduce leadership teams to these four data types to build their data-based decisionmaking capacities.

Researchers argue that data-rich information can help not only improve practice, but in some instances also improve student performance (Bernhardt 2003; McIntire 2005; Protheroe 2001; Wayman, Stringfield, and Yakimowski 2004). In a cluster randomized trial, Carlson, Borman, and Robinson (2010) examined the effectiveness of a districtwide data-driven reform initiative that helped district and school leaders implement student benchmark assessments and interpret and use student results to guide education reform efforts. The researchers found that the initiative had no statistically significant effect on reading achievement ($d = .14$), but did have a statistically significant positive effect on mathematics achievement ($d = .21$) after one year of implementation (Carlson, Borman, and Robinson 2010). Based on objective observations examining how 45 elementary school teachers used assessments to inform their mathematics instructional practices, Goertz, Olah, and Riggan (2009) found that teachers accessed and analyzed data for reteaching purposes but did not make fundamental changes in the way they taught mathematics. The researchers recommended that teachers receive more professional development on interpreting student assessment data and linking its use to specific instructional approaches and strategies.

Success in Sight facilitators involve school leadership teams in four steps of data-based decisionmaking that could potentially be applied at any level of school systems (individual, classroom, program, school, or district).

- *Collect and organize data*—define specific questions to investigate, determine the types and sources of data needed, and develop a data collection plan. This step could involve collecting new data or accessing extant data related to student achievement, demographics, programs, and teacher and student perceptions (Bernhardt 2003).
- *Analyze data*—examine data to uncover patterns and relationships, summarize data with charts and graphs, and record factual observations.
- *Interpret data*—summarize observations, generate possible explanations for data patterns, and identifying root causes for those patterns.
- *Plan to take action*—develop measurable and realistic improvement goals, define specific research-based activities intended to accomplish those goals, and devise a plan for monitoring implementation and progress toward intended outcomes.

Successes in Sight leadership teams are introduced to these data-based decisionmaking steps in the second large-group professional development session during the first year of program implementation. With the ongoing mentoring support of Success in Sight facilitators, team members practice and apply these steps in their schools, focusing on their specific identified areas of need. In the fifth large-group professional development session that occurs during the second year of implementation, leadership teams review the four-step process and discuss how to monitor and adjust improvement efforts. During this session, facilitators present a framework for monitoring implementation quality, fidelity, intensity, and consistency to improve practice. Facilitators offer participants information about key data and assessment terms, ways that leadership teams can support monitoring at the school level, components for structuring collaborative time to pursue monitoring and improvement strategies, and ways to use formative data to determine effectiveness of strategies and make adjustments as needed.

Purposeful community—forming and sustaining a community that identifies with and works collectively toward important outcomes that matter to all, uses all available resources effectively, operates from a set of agreed-upon processes that guide actions and decisions in the school, and shares a collective belief that the community can accomplish its goals.

The concept of "purposeful community" is similar to the widely-used "professional learning community," which refers to a community with shared values and a focus on student learning that engages in collaboration, deprivatized practice, and reflective dialogue (DuFour 2004; Louis and Marks 1998). Researchers have argued that professional learning communities in schools—as measured by public classroom practice, reflective dialogue, peer collaboration, proactive new teacher socialization, collective responsibility for school improvement, and a specific focus on student learning—is essential for schoolwide improvement in student achievement (Bryk et al. 2010). Empirical qualitative studies have found that teacher participation in professional learning communities positively influenced student achievement (Berry, Johnson, and Montgomery 2005; Hollins et al. 2004; Phillips 2003; Strahan 2003; Supovitz 2002; Supovitz and Christman 2003). Success in Sight adapts many characteristics of professional learning communities into its systemic school improvement model, but its developers distinguish purposeful community from professional learning community because of the former's emphasis on building collective efficacy.

Collective efficacy is defined as a group's shared perception that it can organize and execute a course of action that makes a difference (Goddard 2002). "The strength of families, communities, organizations, social institutions, and even nations lies partly in people's sense of collective efficacy that they can solve problems they face and improve their lives through united effort" (Bandura 1997, p. 80). In their research on the impact of collective efficacy on schools, Hoy, Smith, and Sweetland (2002) found that schools with high levels of collective efficacy are more likely to accept challenging goals, demonstrate stronger efforts, and persist in efforts to overcome difficulties and succeed. Collective efficacy is task specific in the sense that teachers might experience a high level of collective efficacy in one area and a low level of collective efficacy in another area. During the first large-group professional development session, Success in Sight facilitators discuss the research-based importance of collective efficacy and purposeful community to school improvement, and opportunities are provided for participating school teams to reflect on strengthening these elements through planning, implementing, and evaluating the

effects of change. As leadership teams progress in their implementation of Success in Sight, they involve more teachers in their efforts in order to build increased schoolwide collective efficacy.

Shared leadership—participating in a process of mutual influence, responsibility, and accountability for achieving collective, organizational goals for school improvement.

Leithwood et al. (2004) concluded in their literature review on leadership that there is an association between increased student learning and leaders who develop and rely on leadership contributions from a diverse constituent base within their organizations. Success in Sight promotes shared leadership through an emphasis on collaboration and capacity building at the teacher, school, and district levels. Addressing these different levels within a school system helps ensure sustainability and system coherence in support of school improvement efforts (Lippitt and Lippitt 1986).

Success in Sight focuses on helping schools develop a culture of shared leadership in which principals, teachers, and other staff accept responsibility for helping the school achieve its improvement goals. Facilitators work with leadership teams made up of principals, teachers, and other staff. Through participation on collaborative teams, team members are expected to build their individual capacity for leading change and improving instruction as well as increasing the school's capacity as a whole. These increased school and individual capacities are mutually reinforcing and are believed to lead to the ultimate goal of improved student outcomes (Hallinger and Heck 2010). According to Printy and Marks (2006, p. 130), "Best results occur in schools where principals are strong leaders who also facilitate leadership by teachers; that is, principals are active in instructional matters in concert with teachers whom they regard as professionals and full partners. Where schools have the benefit of shared instructional leadership, faculty members offer students their best efforts and students respond in kind."

Hulpia, Devos, and Rosseel (2009) identify a coherent leadership team as an important characteristic of a shared leadership model, describing it as a team that works together on explicit, agreed-upon objectives for the school with a shared understanding of the tasks expected of them and a willingness to implement tasks. They contend that the function of a leadership team consists of supportive leadership, a concept that includes helping or complimenting teachers, questioning and debating school vision, considering the personal welfare of teachers, and encouraging teachers to seek out practices based on teacher interests, for example. Rhoton (2001, p. 20) refers to supportive leadership as using "a variety of behaviors to show acceptance of and concern for subordinates' needs and feelings" and notes that "supportive leadership increases the satisfaction and productivity of the people involved."

The Success in Sight shared leadership component incorporates both the coherent leadership team characteristic and the supportive leadership function. During large-group professional development sessions and the onsite mentoring sessions, Success in Sight facilitators work with leadership teams to clarify their role, responsibilities, and decisionmaking processes and methods for supporting school improvement efforts related to student achievement. As part of the shared leadership component, Success in Sight facilitators aim to increase leadership teams' capacities for supportive leadership by helping them identify the level of trust in the school, address mistrust, improve communication, and involve other teachers in sharing and participating in improvement efforts.

Research-based practices—adopting practices that directly address factors shown to be associated with improved student achievement and are based on scientific evidence of effectiveness.

Success in Sight emphasizes scientific inquiry as a primary source of guidance for school improvement. The program provides leadership teams with resources and strategies for accessing and understanding research literature, and facilitates the use of the research to identify solutions to problems associated with improving student achievement in their particular context. For example, the first and second large-group professional development sessions include activities that introduce participants to and involve participants in considering applications of meta-analytic research on factors that influence student success, including student-level factors, leadership-level factors, and teacher- and school-level factors (see, for example, Marzano 2003). During their onsite visits, Success in Sight facilitators continue to provide mentoring support to help schools identify appropriate research-based practices that align with their school improvement efforts.

Continuous improvement process—employing a five-stage process to improve student performance by taking stock of the current situation, focusing on the right solution, taking collective action, monitoring progress and adjusting efforts, and maintaining momentum for improvement efforts.

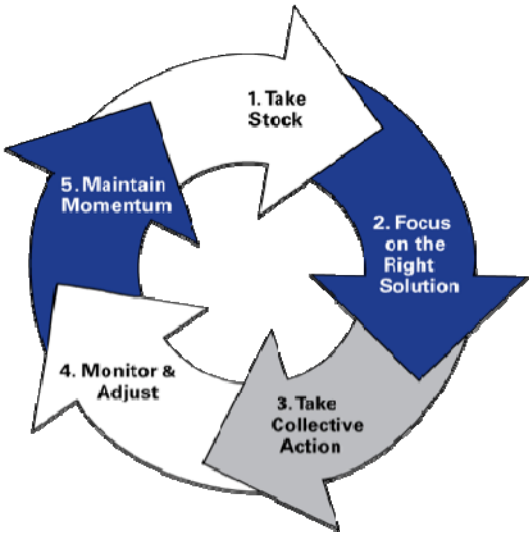
The continuous improvement process is a program of action that integrates the four capacity areas described above. Success in Sight developers theorize that with repeated application of a five-stage continuous improvement process with manageable projects, school leadership teams reinforce their knowledge and skills in the other capacity areas, build their collective efficacy, and attempt to take on larger and more complex change initiatives with confidence. Team members learn about the continuous improvement process through large-group professional development sessions, then apply the process by planning and implementing small, manageable improvement efforts in their schools with mentoring from Success in Sight facilitators. The five stages of the process are (figure 1.1):

1. *Taking stock*—examining the structures, processes, and attitudes in place to support improvement, and identifying problem areas to address. Team members identify structures (such as information and data management systems, collaborative work groups, or meeting schedules) that could support school improvement. They identify processes for making data-based decisions, communicating information, identifying research-based strategies, and defining school improvement strategies. They take stock of staff attitudes regarding shared responsibility and accountability, perceptions of student potentials, and willingness to take risks and work collaboratively. Leadership team members also are introduced to data-based decisionmaking and how to use data to assess student strengths, prioritize needs, and establish goals for improvement.
2. *Focusing on the right solution*—developing appropriate improvement plans for specific problems. Success in Sight facilitators work with each school's leadership team to identify and adopt research-based practices most likely to address problems while ensuring alignment with district priorities and goals.
3. *Taking collective action*—developing and maintaining purposeful communities where everyone works collaboratively and effectively to improve student learning. Leadership

teams explore the use of professional learning communities and collaborative team meetings, take actions to help staff manage school change, and attend to elements of school culture that might influence school improvement (such as trust, communication, participation, productive mindsets, high expectations for students and staff, and optimism).

4. *Monitoring and adjusting*—developing systems to formally and informally collect data and monitor progress in improvement strategies. Schools identify what is working and what is not working, and make necessary adjustments.
5. *Maintaining momentum*—establishing structures and processes to build on successes. To inform ongoing initiatives, schools reflect on and document what led to success with their improvement efforts and what decreased the effectiveness of their efforts.

Figure 1.1 Success in Sight's five stages of the continuous improvement process



Source: McREL 2008.

Success in Sight program delivery

Success in Sight program delivery typically takes place over two years, during which facilitators conduct six large-group professional development sessions with consortia of multiple school leadership teams, 10 onsite mentoring sessions for school leadership teams, distance support for school leadership teams between site visits, and fractal improvement experiences of increasing magnitude. (Each activity is detailed below.) The program is designed to increase the capacity of leadership teams and teachers to implement school improvement practices, which in turn, are expected to increase school capacity as a whole and improve student achievement.

The Success in Sight delivery model is based on a blend of what Collins (1998) has described as two basic types of models of change: rational—emphasizing logical planning, problem solving, and execution—and socialized—emphasizing the process of changing and the unique context and culture of each situation. In the Success in Sight approach, the rational and socialized models are blended in opportunities and structures for collaborative problem solving and systematic continuous improvement. This approach asks participants to form school leadership teams,

introduces tools for rational problem solving, gives assignments to teams to identify and solve problems together, and provides opportunities for teams to reflect together and compare and contrast their context and culture to that of other schools.

During the large-group professional development sessions, leadership teams are introduced to standardized processes and content that together serve as a "toolbox" for school improvement and tend to represent Collins's rational change model. Assistance provided during onsite mentoring sessions and through distance support, on the other hand, is tailored to individual school needs and focuses on helping schools adapt content from the large-group professional development sessions to their individual contexts—representing the more socialized model of change. This assistance is intended to build schools' capacity to solve problems in a way that also takes into account their existing strengths.

Throughout this work, school leadership teams have time to reflect on their experiences and plan next moves during large-group professional development sessions. In this manner, Success in Sight is designed to implement an approach to organizational change and development that "provides an adaptable and real-time discipline for living systems that require information sharing to govern next moves and adjustments…[and] is interactive, relational, participative, and engaging" (Rothwell, Stavros, and Sullivan 2010, pp. 821–24).

Delivery component 1: large-group professional development sessions. Success in Sight uses a consortium model to build leadership teams' capacity to implement school improvement efforts. Facilitators deliver professional development to consortia of school leadership teams of five to seven staff members (principals, teachers, and other staff) during three two-day sessions each year. The meetings are designed to provide opportunities for teams from different schools to collaborate and learn from one another by sharing successes and challenges in their efforts to implement school improvements.

The purpose of the large-group professional development sessions is to build participants' capacities, knowledge, and skills in the five capacity areas: data-based decisionmaking, purposeful community, shared leadership, research-based practices, and continuous improvement. Following the introductory model, each two-day session examines one or more stage of the continuous improvement process in depth while also addressing the other four capacity-building areas through large- and small-group activities. Sessions include time for each team to work with two Success in Sight facilitators to plan how they will use the information back at their school sites.

Success in Sight facilitators deliver six modules, approximately one module per session, during the large-group professional development sessions over the two-year period. Module 1 is delivered prior to the first school year. Modules 2 and 3 are delivered during the first school year. Module 4 is delivered prior to the second school year. Modules 5 and 6 are delivered during the second school year. Descriptions of each follow.

Module 1. Facilitators present the overall Success in Sight approach and focus specifically on the five-stage continuous improvement process. Participants design a manageable change initiative (fractal improvement experience) that can be implemented immediately while incorporating the five stages of the continuous improvement process. Teams are introduced to a set of research-

based school and teacher practices, and student characteristics that improve student achievement, and they discuss their roles as leadership teams. This session also introduces the concept of purposeful community.

Module 2. Teams explore stages 1 and 2 of the continuous improvement process—taking stock and focusing on the right solution. They are introduced to four types of data, gain experience analyzing and interpreting data, and practice setting goals for improvement. If teams have experience using data, facilitators modify this session to deepen participants' capacity for in-depth analysis and interpretation of data. Teams also learn how to identify research-based strategies for improvement and conduct a quality review of those strategies. This session includes activities aimed at improving understanding of two aspects of purposeful community: "outcomes that matter to all" and "collective efficacy."

Module 3. Teams focus on stage 3 of the continuous improvement process—taking collective action. They engage in activities to define and measure improvement progress and to build group effectiveness in improving student achievement. They work to expand their understanding of purposeful community by focusing on the "agreed-upon processes" and "use of all available assets" attributes. Facilitators also introduce teams to the concept of magnitudes of change, explaining that first-order changes are often an extension of past practice, are consistent with prevailing values and norms, are implemented with existing knowledge and skills, are incremental, and affirm existing paradigms (Waters and Cameron 2007). Second-order changes, which Success in Sight promotes, break with past practice, are complex, conflict with prevailing values and norms, are outside existing paradigms, and require new knowledge and skills to implement (Waters and Cameron 2007). During this session, Success in Sight facilitators help school leadership teams understand both types of change and identify specific leadership actions they can take to manage second-order change and ensure lasting results.

Module 4. Leadership teams explore how to establish structures, processes, and attitudes that help the staff engage in stage 3 of the continuous improvement process—taking collective action. Facilitators present the program's four aspects of school culture—trust, communication, collaboration, and participation in decisionmaking—and the role culture plays in implementing change initiatives. This session emphasizes how "if certain norms of school culture are strong, improvements in instruction will be significant, continuous, and widespread; if these norms are weak, improvements will be at best infrequent, random, and slow" (Saphier and King 1985, p. 67). In addition, facilitators present information and activities related to shared leadership and the role of the leadership teams when improvement initiatives have second-order implications for the majority of staff. Participants work to deepen their understanding of ways to enhance collective efficacy.

Module 5. As part of their investigation of stage 4 of the continuous improvement process—monitoring and adjusting—teams work to deepen their data analysis and interpretation skills and increase their ability to use formative and summative data to determine the effectiveness of their improvement strategies. Teams revisit shared leadership, learning more strategies for managing the transitions that accompany second-order change, and explore the use of tangible and intangible assets for accomplishing outcomes that matter to all.

Module 6. Teams address ways to sustain improvement efforts by maintaining momentum (stage 5 of the continuous improvement process). They engage in activities designed to help them examine the structures and processes they have put in place to support ongoing improvement, and they develop sustainability plans. This session provides opportunities for teams to reflect on what they have learned about purposeful community, use of data, shared leadership, influences on student achievement, and the systematic improvement process. The session intends to deepen participants' understanding of how a school's purpose and vision can guide future improvement initiatives.

Delivery component 2: onsite mentoring and support. Success in Sight facilitators meet with each school's leadership team approximately once per month, for a total of 10 onsite meetings, to support teams as they apply what they learned during the large-group sessions in ways that are tailored to their specific school improvement priorities. The focus of onsite meetings varies with each school but might include, for instance, helping a team develop norms for working together and communicating with other staff, plan professional development to help other staff understand the systematic improvement process, refine the school's plan for implementing the small change initiative they designed at the large-group session, or develop a vision for future success. Facilitators meet with leadership teams for four to six hours during these visits. The remaining time during the visit is spent meeting with administrators, facilitating and lending support to professional learning community groups, and meeting with individual teachers.

Delivery component 3: distance support for school leadership teams. Leadership teams in treatment schools receive additional support for implementing the continuous improvement process by participating in phone conferences and email exchanges with Success in Sight facilitators. These communications occur with leadership team members on an as-needed basis to provide timely mentoring support.

Delivery component 4: fractal improvement experiences. Fractal improvement experiences are short-term projects designed to obtain quick results while providing practice in the five stages of the continuous improvement process and in the five capacity areas. Fractal improvement experiences are expected to be mechanisms for teams to experiment and "learn by doing" over time (Argyris 1976; Argyris and Schol 1996; Beckhard 1969; Beckhard and Pritchard 1992; Dewey 1938; DiBella and Nevis 1998; Freire 1998; Fullan 2010; Senge 1990). By experiencing quick success through early, manageable fractal improvement experiences, teams build collective efficacy, or the belief that by working together they can make a difference in student achievement. This approach to building confidence, credibility, and momentum for further change is supported by several change theorists (see, for example, Adams 1997; Kouzes and Posner 1997; Lippitt, Watson, and Westley 1958; Warrick 2005). Through repeated applications of the continuous improvement process, teams are expected to increase their knowledge and skills in the five capacity areas and learn how to take on larger and more complex initiatives with confidence. Facilitators help teams design and implement fractal improvement experiences over the two-year intervention period, providing less guidance as teams develop their own capacity to accomplish improvement goals.

During the first two professional development sessions, Success in Sight facilitators guide teams through a fractal improvement experience process that incorporates all five stages of the continuous improvement process. Based on their schools' specific improvement needs, teams

identify a focus area for their fractal improvement experience, such as reading, mathematics, school culture, student engagement, or parent involvement. Teams design a fractal improvement experience in a two-hour workshop during which they look at their data and select strategies based on research-based or "best" practices (taking stock); they design fractal improvement experiences that are manageable in scope and can be accomplished in four to six weeks (focusing on the right solution and taking collective action); they implement their solutions and monitor and adjust their plan as needed (monitoring and adjusting); they document and share with facilitators and peers the things that helped and hindered their success, and they use this information to inform their next fractal improvement experience (maintaining momentum). Between large-group professional development sessions, facilitators help teams refine and implement fractal improvement experience plans through onsite mentoring and distance support.

Following the initial guided practice, leadership teams design and implement subsequent fractal improvement experiences and repeat the five-stage continuous improvement process. The time it takes teams to complete each stage of the process is expected to vary based on the nature and complexity of the fractal, access to the necessary information and resources, staff time to meet and implement tasks, and experience with the process itself. For example, it could take several weeks to gather data for stage 1 (taking stock) and several more days to make decisions about the right solution (stage 2). As teams become more sophisticated in their use of the continuous improvement process and the complexity of the problems increases, teams might require more time for stage 2 to research and select appropriate improvement strategies for their school's local context. Stages 3 and 4 together (taking collective action and monitoring and adjusting, respectively) might take three to five weeks for their initial fractals. Stage 5 (maintaining momentum) might be completed in one to two meetings for leadership team members. Leadership teams are expected to complete at least two fractal improvement experiences per year of increasing magnitude—that is, experiences requiring new knowledge and skills; departing from past practices, values, norms, or paradigms; and expanding to involve community members, parents, teachers, or school staff who do not serve on the leadership teams. For one year or more, depending on the needs and context of each school, Success in Sight facilitators guide leadership teams in focusing on school improvement practices related to a specific content area for student growth.

A key factor in developing schoolwide capacity for school improvement practices is to involve an increasing number of teachers in fractal improvement experiences. This exposes teachers to the "learning by doing" approach to increase their understanding of the five-stage continuous improvement process as well as data-based decisionmaking, purposeful community, shared leadership, and research-based strategies. Teachers then have the opportunity to apply what they learn from these experiences to other schoolwide improvement initiatives and classroom instruction. Teachers further develop their capacities in the five Success in Sight areas as they continue to collaborate with leadership team members in planning and implementing additional fractal improvement experiences.

The following is an example of a fractal improvement experience for a leadership team that wanted to address low reading test scores for specific student populations in the school.

The team began the continuous improvement process by taking stock—looking at state and district reading achievement data for identified student populations. This led the team to focus on

the right solution—in this case, building student academic vocabulary using research-based strategies—and then to plan collective action involving the entire staff. With guidance from Success in Sight facilitators, the team:

- Set a six-week timeline for the initiative.
- Worked with grade-level teams to choose academic vocabulary to be taught and assessed weekly.
- Planned for pre- and posttesting as summative evaluation.
- Taught vocabulary strategies to the rest of the staff.
- Set specific targets for student achievement.

Team members monitored the experience by meeting weekly to review data, and they adjusted their program goals and strategies when they saw that their original expectations were not being achieved and students were not learning as many new words as they had hoped. The team learned to identify key concept words and use research-based, direct instruction for those words. As part of the team's effort to maintain momentum, it asked teachers to reflect on what worked well, what did not work well, and what could be changed to improve the vocabulary fractal experience. Positive feedback from teachers helped the team decide to continue vocabulary development. Having strengthened their own individual and collective capacities by using the continuous improvement process with vocabulary content, team members were then able to extend the fractal improvement strategy to include all core content areas. The team implemented these strategies on its own while Success in Sight facilitators helped it focus on other achievement areas for future fractal initiatives.

In this example, the fractal improvement experience targeted a specific issue that was manageable in scope and duration, and it engaged participants in the continuous improvement process while also focusing on other capacity areas (data-based decisionmaking, fostering a purposeful community, building shared leadership in the school improvement process, and using research-based practices). It helped the leadership team develop the structures (grade-level teams, timeline, evaluation plan) and processes (target setting, data collection, staff training) that supported this particular school improvement effort. The example also depicts how the magnitude of the improvement efforts expanded to include all teachers and core content areas.

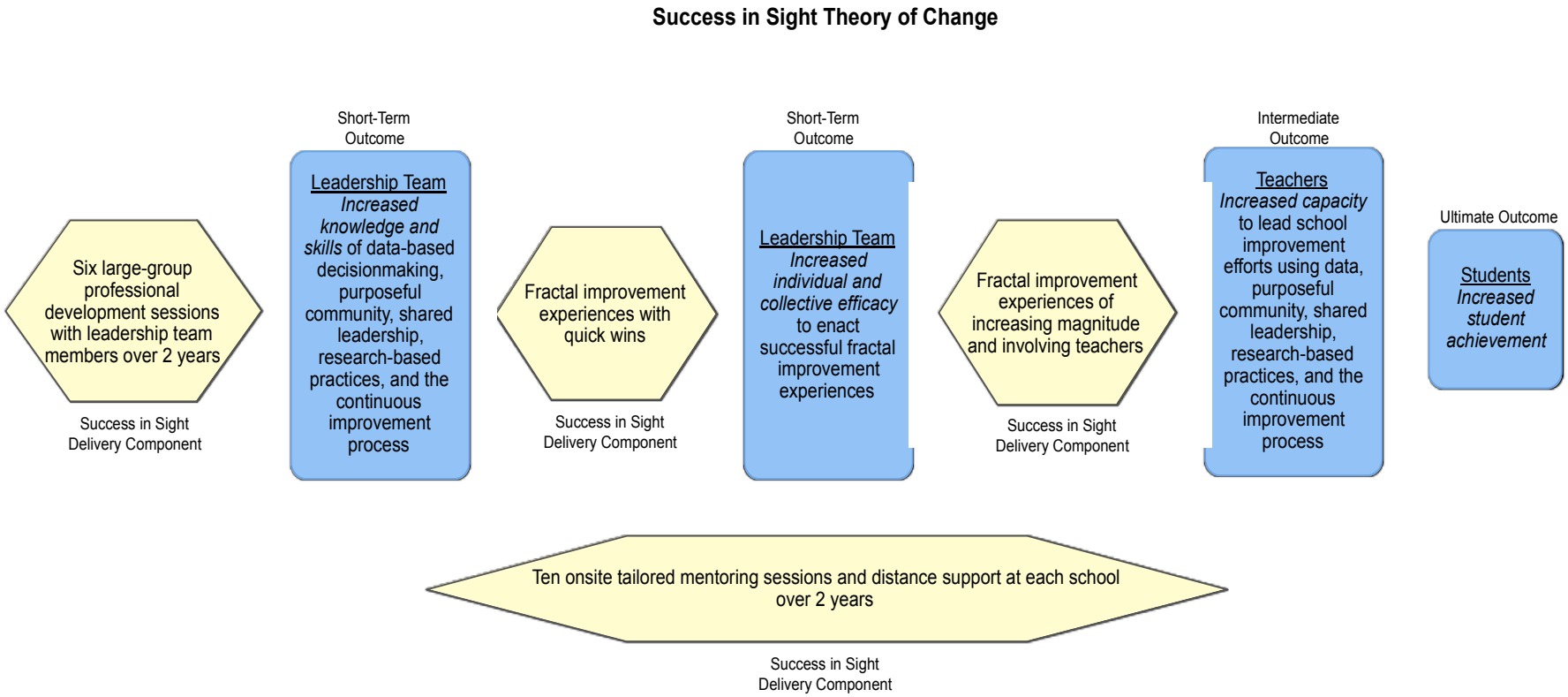
Success in Sight's theory of change

The Success in Sight theory of change posits that through large-group professional development in five key capacity areas (data-based decisionmaking, purposeful community, shared leadership, research-based practices, and the continuous improvement process), onsite tailored mentoring and distance support, and fractal improvement experiences, school leadership teams and teachers will over time be able to implement systemic changes that will result in improved student achievement. Large-group professional development is intended to result in short-term outcomes for leadership team members, including increased knowledge and skills in the five capacity areas (figure 1.2). With quick successes in fractal improvement experiences related to student

outcomes, the collective efficacy of leadership team members grows. Onsite mentoring and distance support is intended to expand leadership team members' capacities to increase the magnitude of fractal improvement experiences and increase teacher participation.

As teachers become involved in fractal improvement experiences, they participate in the same "learning by doing" approach that leadership team members experience during initial fractal improvement experiences. The fractal improvement experience takes teachers through the five-stage continuous improvement process which incorporates elements of data-based decisionmaking, purposeful community, shared leadership, and research-based strategies. By participating in fractal improvement experiences, it is intended that teachers' capacities for data-based decisionmaking, purposeful community, shared leadership, research-based strategies, and the continuous improvement process will increase. It is expected that teachers will apply what they learn from these experiences to school-level improvement efforts and classroom-level instructional practices geared toward increasing student achievement. As teachers increasingly join leadership team members in planning and implementing fractal improvement experiences, it is intended that teachers further enhance their capacities in the five areas.

Figure 1.2 Success in Sight theory of change



Source: Theory of change developed by authors, based on Success in Sight program materials.

It is expected that the increase in leadership team members' and teachers' capacities in data-based decisionmaking, purposeful community, shared leadership, research-based strategies, and the continuous improvement process will reflect an increased schoolwide capacity to implement improvement initiatives. Ultimately, the intended result of all school improvement initiatives is higher student achievement schoolwide.

This theory recognizes that the timeframe for realizing student results will vary based on schools' local conditions (such as level of trust among staff, how much experience the staff has working collaboratively, leadership capacity and support of the principal), contexts (such as student and teacher attrition, student demographics, budget stability, and policy changes) and salient issues (such as reading or mathematics achievement, teacher capacity, and school culture). Although leadership teams might improve schoolwide structures and processes that could impact instruction and learning across content areas, within the first two years of implementation impacts might be more detectable in the content area of primary emphasis. Success in Sight can extend into a third year of implementation for schools wanting to continue creating and implementing fractal improvement experiences. For struggling schools, a third year gives them more time to focus more attention and create more fractal improvement experiences for particularly weak areas related to student achievement (such as data use or shared leadership). Schools also can use a third year of implementation to sustain improvement efforts by increasing the magnitude of previously successful fractal improvement experiences to reach more school leaders, teachers, and staff, and to address other content areas.

Success in Sight developers and facilitators report that they have observed small-scale results measured by classroom assessments related to fractal focus areas within the first two years of the program and broad-scale results measured by district and state assessments after three to four years in schools that sustain fidelity of program implementation (personal communication, Danette Parsley, McREL Senior Director, December 7, 2010). This timeframe is consistent with research on educational change in elementary schools that states "moderately complex change takes from 2 to 4 years" (Fullan, 2007, p. 68). In their meta-analysis of comprehensive school reform initiatives, Borman et al. (2003) found a statistically significant effect ($d = 0.14$) on student achievement after two years of implementation. However, there is no efficacy research on Success in Sight that shows detectable changes in student achievement as measured by state assessments after two years of implementation.

Study overview

Although educators have used Success in Sight in selected sites across the nation since 2000 to assist with their school improvement efforts, the intervention lacks causal evidence of its effectiveness in improving student and teacher outcomes. The primary focus of this study was to provide an unbiased estimate of the impact of Success in Sight on student academic achievement in reading or mathematics.[2] The study also was designed to provide an unbiased estimate of the effects of Success in Sight on teacher capacity for school improvement practices in data-based decisionmaking, purposeful community, and shared leadership.

[2] The achievement outcome areas of reading and mathematics were chosen for this study based on the NCLB mandate that all students should be proficient in reading and mathematics by 2014. As a result of this mandate, all states assess students' reading and mathematics achievements in grades 3 through 5.

Study design

The study used an experimental design with 52 elementary schools randomly assigned to either the treatment (n = 26) or control (n = 26) condition for the 2008/09 and 2009/10 school years. The target population was low- to moderate-performing large and small elementary schools in rural, urban, and suburban settings. Participating schools were located in two states: Minnesota and Missouri. The period of implementation and data collection for the two-year intervention was March 2008–June 2010.

Schools in the treatment group participated in Success in Sight's six large-group professional development sessions, 10 onsite mentoring sessions, and a minimum of two fractal improvement experiences during the 2008/09 and 2009/10 school years. The large-group sessions included three consortia: Minnesota (12 treatment schools), Missouri Area 1 (7 treatment schools), and Missouri Area 2 (7 treatment schools). Missouri was divided into two areas to provide intervention participants a location close to their schools for the large-group professional development sessions. The control schools continued to use their usual school improvement practices. (The Success in Sight intervention is not intended to replace existing reform efforts but rather to engage schools in a process that incorporates existing and new improvement practices.) At the end of the study, control schools could elect to participate in the intervention at their own discretion and expense.

All school principals, leadership team members, classroom teachers, and instructional staff in treatment and control schools were required to participate in the study. Treatment and control participants received monetary stipends for their participation in the annual teacher survey, a baseline focus group, and a follow-up phone interview (chapter 2 presents stipend amounts for participants). Student reading and mathematics state assessment data from 2009/10 were collected for students in grades 3–5 for the impact analysis. The sample for the impact analysis included 8,182 students for reading achievement, 8,213 students for mathematics achievement, and 1,516 teachers.

Research questions

This study addresses five research questions—two primary and three secondary—that fall within two domains: student achievement and teacher capacity for school improvement practices.

Primary research questions: student achievement

1. Does implementation of Success in Sight have a significant impact on student achievement in reading?
2. Does implementation of Success in Sight have a significant impact on student achievement in mathematics?

The primary research questions examine the effect of participation in Success in Sight on student achievement in reading and mathematics.[3] Success in Sight does not inherently focus on any particular content area of learning or achievement, but rather focuses on building the overall functioning of a school in its capacity to implement continuous school improvement in areas of achievement important and relevant to them. Reading and mathematics were selected to measure the impacts of Success in Sight on student achievement across 52 elementary schools in two states. These outcomes were selected in part because the NCLB Act of 2001 holds low-performing schools accountable for improved reading and mathematics achievement based on state assessments.

Secondary research questions: teacher capacity for school improvement practices

1. Does implementation of Success in Sight have a significant impact on teacher capacity for engaging in data-based decisionmaking?
2. Does implementation of Success in Sight have a significant impact on teacher capacity for developing and maintaining a purposeful community?
3. Does implementation of Success in Sight have a significant impact on teacher capacity for shared leadership?[4]

The broad intent of Success in Sight is to strengthen school capacities to use improvement practices to increase student achievement. Success in Sight aims to build school capacity by working with school leadership teams comprised of principals, teachers, and other staff. The program developers theorize that as leadership teams increase teacher participation in their fractal improvement experiences, teachers will increase their capacity to implement school improvement practices. Although the intervention addresses school capacity broadly, this study measured teacher capacity because teachers are those most directly responsible for applying improvement practices with students. The three secondary research questions examine the intermediate effects of the intervention on teachers' capacity for data-based decisionmaking, purposeful community, and shared leadership, which provides information supporting interpretation of the main impacts on student outcomes.

[3] Success in Sight developers note that because schools usually focus on one achievement area (that is, reading or mathematics) during the first two years of Success in Sight implementation, any impacts on achievement might be uneven across content areas. Treatment schools chose to focus primarily on reading or mathematics based on their local needs, current initiatives, and areas of improvement once implementation had begun, and therefore researchers did not know which schools would focus on which content area before data collection began. Therefore, the study's primary research questions examine the effect of Success in Sight on either achievement in reading or achievement in mathematics, rather than a composite of both, after two years of implementation.

[4] The secondary research questions address three hypothesized short-term outcome areas (data-based decisionmaking, purposeful community, and shared leadership) and omit two (research-based practices and continuous improvement process). Although all five areas are important components of the change process, this study focused only on the selected three because they represent a requisite set of knowledge and skills for the other two areas (selecting research-based practices that address the most pressing problems, and enacting and managing the continuous improvement process). Therefore, it is possible that impacts on teacher capacity for engaging in data-based decisionmaking, developing and maintaining a purposeful community, and sharing leadership would emerge before impacts related to research-based practices and continuous improvement process.

Exploratory research questions

For this study, interpretations regarding the effectiveness of Success in Sight are based on the primary research question findings. However, because Success in Sight is intended to increase teacher capacity for data-based decisionmaking, purposeful community, and shared leadership, it is important to directly explore the relationship between these teacher capacities and student achievement in reading and mathematics. The study, therefore, poses exploratory research questions to address the hypothesized relationship between teacher capacity and student achievement outcomes:

1. What is the relationship between teacher capacity for data-based decisionmaking and student achievement in reading?
2. What is the relationship between teacher capacity for data-based decisionmaking and student achievement in mathematics?
3. What is the relationship between teacher capacity for purposeful community practices and student achievement in reading?
4. What is the relationship between teacher capacity for purposeful community practices and student achievement in mathematics?
5. What is the relationship between teacher capacity for shared leadership and student achievement in reading?
6. What is the relationship between teacher capacity for shared leadership and student achievement in mathematics?

Answers to the research questions are intended to inform educators about the effectiveness of the Success in Sight intervention for systemic school improvement. Study results will provide policymakers and state and district officials the knowledge they need to determine whether to invest in Success in Sight for their low- to moderate-performing schools.

Content and organization of this report

This report presents findings from a cluster randomized trial designed to estimate the impact of Success in Sight on student achievement and school improvement practices. Chapter 2 presents the study design and methodology, including sample characteristics, data collection procedures, and estimation approach. Chapter 3 describes the implementation of the intervention under study. Chapter 4 presents findings from the impact analysis. Chapter 5 presents findings from the exploratory analysis. Chapter 6 concludes the report by summarizing key findings.

Chapter 2. Study design and methodology

This study uses a cluster randomized trial to assess the impacts of Success in Sight on student achievement in reading or mathematics and teacher capacity for school improvement practices. Researchers determined that a cluster randomized trial with school-level random assignment was an appropriate design for this study because Success in Sight is a schoolwide intervention that is delivered at the school level rather than at the individual student or classroom level.

As a schoolwide intervention, Success in Sight is expected to improve overall school functioning regardless of student and teacher mobility. The intervention's theory of change posits that the effects of school functioning on student achievement should emerge in the overall student body regardless of how long individual students had been enrolled at a particular school at a given time. Likewise, the effects on teacher capacity for school improvement practices were expected to emerge at the school level regardless of how long individual teachers had been teaching at a particular school at a given time.

Consistent with the hypothesis that Success in Sight should affect overall school functioning regardless of individual student and teacher mobility, data collection efforts focused on students and teachers present within participating schools at each data collection point rather than following students and teachers longitudinally. Specifically, researchers collected student achievement data from state reading and mathematics assessments in 2010 to assess the primary research questions, and researchers collected teacher capacity data from a teacher survey administered in 2010 to assess the secondary research questions. Although it would also have been informative to examine how students themselves may have changed relative to each other in response to the intervention, the intent of this study was to estimate the main effect of the schoolwide intervention, which was delivered across grades.

For this study, implementation of Success in Sight occurred over the 2008/09 and 2009/10 school years. During this timeframe, schools in the treatment group participated in Success in Sight, and schools in the control group served as the comparison for the study and continued their regular school improvement activities, or "business as usual," as described in chapter 3.

One potential limitation of this study's design is the two-year timeframe. This study estimates the impact of Success in Sight on student achievement in reading or mathematics after two years of implementation. The Success in Sight developers assert that immediate, small-scale results can emerge (often on teacher-developed or curriculum-based assessments) during the technical assistance period. They also assert that broader scale results on district or state assessments should not be expected until school staff achieve and continue implementation fidelity regarding the Success in Sight structures and process and develop proficient knowledge and skills in all five program outcome areas. The timeframe for these developments to occur and continue varies. Therefore, it is unclear whether two years of implementation is sufficient to yield student achievement impacts measurable by state assessments.

This chapter describes the study's design and methodology, including the study timeline, study sample, data collection, and data analysis methods.

Study timeline

The study's main activities occurred from September 2007 to June 2010 (table 2.1). Researchers identified interested districts and schools beginning in September 2007 and secured district and school memoranda of understanding on a rolling basis until July 2008. Random assignment of schools within each district occurred before any data collection activities took place for each district. Implementation of Success in Sight occurred during the 2008/09 and 2009/10 school years, with the first training occurring in June 2008 (for treatment schools from Minnesota), July 2008 (for treatment schools from Missouri Area 1), and September 2008 (for treatment schools from Missouri Area 2). The division of Missouri into Area 1 and Area 2 was based on school location and proximity across seven districts.

Table 2.1 Success in Sight study timeline

Timeframe	Task
September 2007–July 2008	Site recruitment and collection of memoranda of understanding
March 2008–July 2008	Random assignment of schools to treatment and control conditions
March 2008–May 2008	Collection of baseline student achievement data
June 2008	First Success in Sight training for Minnesota treatment schools and start of baseline teacher survey data collection for Minnesota schools and Missouri Area 1 schools (Missouri Area 1 represents four districts close in proximity and similar in size)
July 2008	First Success in Sight training for Missouri Area 1 schools
August 2008	Start of baseline teacher survey data collection for Missouri Area 2 schools (Missouri Area 2 represents three districts close in proximity and similar in size)
September 2008	First Success in Sight training for Missouri Area 2 schools
September 2008–October 2008	Baseline teacher survey data collection closed. Collection of baseline principal interview data and focus group data with principals, leadership team members, and teachers.
March 2010–June 2010	Collection of posttest data (student achievement data, teacher survey data, and principal, leadership team, and staff phone interview data)
May 2010	Final Success in Sight training, end of Success in Sight program delivery

Data collection occurred from March 2008 through August 2010. Baseline student achievement data were collected from March 2008 through May 2008 according to state testing schedules. Baseline teacher survey data were collected from June 2008 through October 2008.[5] The extended survey administration period accounted for time to identify site coordinators and administer the survey when school was in session rather than over the summer of 2008. Eight out of 1,374 (0.58 percent) treatment teachers (all from Missouri) completed the baseline teacher survey after participating in the first Success in Sight training. Therefore, it is possible that the training affected their baseline survey responses.[6] Baseline principal interviews and school focus groups were conducted from September 2008 through October 2008.[7] Posttest student achievement data were collected from March 2010 through May 2010, posttest teacher survey data were collected from March 2010 through April 2010, and posttest phone interviews[8] with the school principal, the leadership team member, and a staff member from each school were conducted from April 2010 through June 2010.[9]

Study sample

This section presents information about the Success in Sight study sample, including a description of the site recruitment and randomization process, comparisons of the study schools at baseline, and documentation of student and teacher mobility and attrition.

Sample recruitment

The study's target population was low- to moderate-performing public elementary schools located in states served by McREL's Regional Educational Laboratory (REL) Central and North Central Comprehensive Center (NCCC) programs (Colorado, Iowa, Kansas, Minnesota, Missouri, Nebraska, North Dakota, South Dakota, and Wyoming). Low- to moderate-performing schools were defined as schools that did not make adequate yearly progress (AYP) for any of the three school years prior to the 2008/09 school year or that were at risk of not making AYP.[10]

---

[5] Baseline teacher survey data collection began in June 2008 for Minnesota and Missouri Area 1 schools and August 2008 for Missouri Area 2 schools. For all schools, the baseline survey administration closed in October 2008.

[6] It is possible that the first Success in Sight training positively or negatively influenced teachers' perceptions regarding teacher capacity for school improvement at their respective schools. However, baseline teacher data were not used as outcome variables in any impact estimates and were used only to establish baseline equivalence and to construct covariates used to increase the precision of the impact estimate.

[7] Although the principal interview data and baseline focus group data were collected after the first Success in Sight professional development session was implemented in treatment schools, these data were not included in the impact analyses and served only to provide information regarding contextual factors present at each school at baseline.

[8] Researchers conducted posttest phone interviews to collect information regarding contextual factors that might have contributed to school improvement efforts across the 2007/08 to 2009/10 school years.

[9] The participant categories overlapped in some cases, wherein the principal was also the leadership team member in a school.

[10] "Being at risk of not making AYP" was defined as having experienced recent changes in the composition of a school's student population that might challenge the capacities of a staff to address the specific needs of new students, such as an influx of students learning English as a second language. Judging whether a school was "at risk" of not making AYP was subjective, based on school personnel reporting an influx of English language learner students in the current or prior year.

Researchers identified these target schools as potential study participants that might need more support than higher-performing schools to achieve NCLB objectives.

Researchers chose public elementary schools serving grades 3–5 because that sample enabled the use of existing data from state-administered reading and mathematics assessments, which reduced the data collection burden for participating schools. The school eligibility criteria for selection and participation in the study were as follows:

1. Public elementary school serving at least grades 3–5 (including schools serving K–5, K–6, and 3–6).
2. Low or moderate performance as indicated by having not made AYP in any of the three years prior to the intervention or being at risk of not making AYP in the current or prior year.
3. At least two classrooms in each of grades 3, 4, and 5, to ensure adequate sample size within each participating school.
4. Not already implementing a comprehensive school reform intervention that includes an emphasis on the continuous improvement process and collective efficacy (two unique features of Success in Sight) and had no plans to do so for the 2008/09 and 2009/10 school years.
5. Not slated to be closed or restructured during the study period.
6. Able to adhere to all study requirements, including random assignment, forming leadership teams of at least five members, and completing all data collection activities.

Among the nine target states, the four states with the highest number of elementary schools not making AYP in 2004/05 were Minnesota (244), Colorado (144), Missouri (129), and Kansas (122) (American Institutes for Research 2005). From this set of four states, recruitment efforts were focused on schools in Minnesota and Missouri. Within these states researchers targeted large and small elementary schools in rural, urban, and suburban settings.

Recruitment efforts began at the district level, which afforded a number of advantages, including garnering support of district administration for the study and reducing the number of required school-level approvals. Researchers recruited sites through outreach at professional education conferences and through other professional networks, including contacts at state departments of education and school districts. Once researchers identified potential sites, the study team worked closely with districts to enlist participation from eligible elementary schools within districts. None of the participating districts required parent consent for student participation. In identifying eligible schools, the research team requested assurances from the district that potential schools were not slated to be closed or restructured during the study period.

Recruitment began in September 2007 and was concluded in July 2008. In Minnesota, the study team contacted two districts for recruitment, but only one expressed interest in participating. Within the interested district, the study team contacted 44 elementary schools for recruitment. In Missouri, the study team contacted 53 districts and a total of 113 elementary schools across the districts for recruitment. Districts that declined the opportunity to participate did so for a number of reasons, including lack of support from key district leadership, contractual concerns about the time teachers would be out of the classroom for professional development, and the need to prioritize initiatives already in place. Schools' reasons for declining to participate included the

need to focus on current initiatives and discomfort with random assignment. No schools were eliminated from the sample if they expressed interest in participating and met the eligibility criteria. The study required 50 elementary schools to ensure statistical power of .80 to detect a minimum standardized effect size of 0.20 for the benchmark impact estimates of primary outcomes. The study required 52 elementary schools to ensure statistical power of .80 to detect a minimum standardized effect size of 0.30 for benchmark impact estimates of secondary outcomes.[11] Researchers recruited 52 schools.

The 52 schools recruited for the study are located in eight districts across Minnesota, Missouri Area 1, and Missouri Area 2 (table 2.2). The division of Missouri into two areas was based on school location and proximity. The study schools represent a mix of city, town, suburb, and rural locales. School sizes ranged from 165 students to 726 students (mean = 392.12, standard deviation = 127.87). Dividing the school sample into quartiles based on number of students per school revealed that within the first quartile, school sizes ranged from 165.00 to 311.25 students, within the second quartile, school sizes ranged from 311.25–393.00 students, within the third quartile, school sizes ranged from 393.00–483.75 students, and within the fourth quartile school sizes ranged from 483.75–726.00 (table 2.3).

Table 2.2 Number of eligible and participating schools by area

Area	Eligible schools	Schools that declined	Participating schools
Minnesota	35	11	24
Missouri Area 1	20	7	13
Missouri Area 2	37	22	15
Total	92	40	52

Source: Study recruitment records.

Table 2.3 School size ranges falling within each quartile of the study sample

Quartile	Number of schools	School size range
First quartile	13	165.00–311.25
Second quartile	13	311.25–393.00
Third quartile	13	393.00–483.75
Fourth quartile	13	483.75–726.00

Source: U.S. Department of Education, National Center for Education Statistics, 2008.

Comparison of study sample schools to state populations of schools

Researchers compared baseline (2008) characteristics of the study sample schools with the larger populations of all Minnesota and Missouri elementary schools not making AYP in any of the three years prior to the study (tables 2.4–2.7). A larger population of "at-risk" schools could not be identified because the criteria for "at-risk" was subjective, based on school personnel reports regarding the influx of English language learner students in the current or prior year.

For Minnesota, there were several statistically significant differences between all Minnesota schools not making AYP and Minnesota study sample schools in reading and mathematics

[11] See appendix B for power analysis estimates.

achievement, students per teacher, students eligible for free or reduced-price lunch, and student population. Mean reading and mathematics achievement scores in 2008 across study sample schools were statistically significantly lower than the mean achievement scores across the larger population of all elementary schools in the state not making AYP (table 2.4). A statistically significantly greater percentage of students in Minnesota study sample schools qualified for free or reduced-price lunch, and Minnesota study sample schools had a statistically significantly lower number of students per teacher compared with the statewide population of elementary schools not making AYP. The population of all Minnesota elementary schools not making AYP included a statistically significantly greater percentage of White students and a statistically significantly lower percentage of Black and Asian students than did the Minnesota study sample schools.

Table 2.4 Baseline comparison of all Minnesota elementary schools not making adequate yearly progress and study sample schools on achievement, size, and student characteristics 2007/08

	Total Minnesota elementary schools not making adequate yearly progress (N = 368)		Minnesota study sample schools (n = 24)				
Characteristic	**Mean**	**Standard deviation**	**Mean**	**Standard deviation**	**Difference**	**Test statistic**	***p*-value**
Reading achievement[a]							
Grade 3	3,620.85	252.13	3,493.85	275.38	–127.00	–16.77	< .01***
Grade 4	3,729.99	268.92	3,605.64	293.89	–124.35	–15.06	< .01***
Grade 5	3,817.80	265.28	3,692.23	274.90	–125.57	–16.34	< .01***
Mathematics achievement[a]							
Grade 3	3,624.16	213.73	3,511.81	227.15	–112.35	–17.70	< .01***
Grade 4	3,704.12	208.70	3,596.15	232.03	–107.97	–16.36	< .01***
Grade 5	3,808.56	207.17	3,712.99	226.36	–95.57	–14.71	< .01***
Students per school[b]	465.77	210.83	413.29	117.87	–52.48	1.21	.23
Students per teacher[b]	16.13	4.00	14.15	1.76	–1.98	2.40	.02**
Students eligible for free or reduced-price lunch (percent)[b]	46.37	24.22	80.31	15.97	33.94	–6.77	< .01***
Student population (percent)[b,c]							
White	63.16	31.97	17.11	14.26	–46.05	7.00	< .01***
Black	15.32	21.85	35.70	20.54	20.38	–4.44	<.01***
Hispanic	9.89	13.83	13.97	9.68	4.08	–1.42	.16
Asian	7.70	14.67	30.58	17.92	22.88	–7.30	< .01***
American Indian	3.93	14.54	2.64	6.86	–1.29	–0.43	.67

**Significant at p = .05; ***significant at p = .01.

Note: Black includes African American, Hispanic includes Latino, Asian includes Native Hawaiian or Other Pacific Islander, and American Indian includes Alaska Native.

Note: Schools were classified as "not making adequate yearly progress" if they did not make adequate yearly progress in one or more of the three years prior to the study (2005/06, 2006/07, or 2007/08).

a. Analyses for reading and mathematics scale scores were one-sample *t*-tests with state-level mean score by grade as population values.

b. Analyses for school demographics were *t*-tests between group means.

c. Values may not sum to 100 percent because of rounding.

Source: Minnesota Department of Education 2008a, 2010a, 2010c; U.S. Department of Education, National Center for Education Statistics 2008.

There also were statistically significant differences between the Minnesota study sample schools and the larger population of Minnesota schools not making AYP in Title I status and urbanicity (table 2.5). Specifically, Minnesota study sample schools had a statistically significantly higher proportion of schoolwide Title I schools than did the larger population of Minnesota elementary schools not making AYP. In addition, Minnesota study sample schools were all located in cities, but the statewide population of Minnesota elementary schools not making AYP included schools from city, suburb, town, and rural locales. There were no statistically significant differences between the Minnesota study sample schools and the larger population of state elementary schools not making AYP with regard to number of Title I–eligible schools.

Table 2.5 Baseline comparison of all Minnesota elementary schools not making adequate yearly progress and study sample schools on Title I, urbanicity, and adequate yearly progress status, 2007/08

	Total Minnesota elementary schools not making adequate yearly progress (N = 368)		Minnesota study sample schools (n = 24)			
Characteristic	***N***	**Percent**	***n***	**Percent**	**Test statistic**	***p*-value**
Schools receiving Title I						
Title I–eligible school	292	79.35	22	91.67	2.15	.14
Schoolwide Title I	97	26.36	21	87.50	33.78	<.01***
School urbanicity						
City	95	25.82	24	100.00	58.65	<.01***
Suburb	109	29.62	0	0.00		
Town	60	16.30	0	0.00		
Rural	104	28.26	0	0.00		

***significant at p = .01.
Note: Schools were classified as "not making adequate yearly progress" if they did not make adequate yearly progress in one or more of the three years prior to the study (2005/06, 2006/07, or 2007/08).
Note: Analyses were chi-square tests between percentages.
Source: Minnesota Department of Education, 2010c; U.S. Department of Education, National Center for Education Statistics, 2008.

For Missouri, there were no statistically significant differences between study sample schools and the larger population of all Missouri elementary schools not making AYP regarding reading achievement in grades 3–5 or regarding mathematics achievement for grades 3 and 5 in 2008. For grade 4 mathematics achievement the study sample schools had statistically significantly higher scores than the larger population of Missouri elementary schools not making AYP (table 2.6).

The statewide population of Missouri elementary schools not making AYP had a statistically significantly lower percentage of American Indian students than did the Missouri study sample schools. There were no statistically significant differences between the statewide population of elementary schools not making AYP and the Missouri study sample schools regarding percentage of White, Black, Asian, or Hispanic students. Missouri study sample schools had a statistically significantly higher mean number of students per teacher than did the larger population of elementary schools not making AYP across the state. There were no statistically

significant differences between the larger population of Missouri elementary schools not making AYP and Missouri study sample schools in the number of students per school or the percentage of students eligible for free and reduced price lunch.

Table 2.6 Baseline comparison of all Missouri elementary schools not making adequate yearly progress and study sample schools on achievement, size, and student characteristics 2007/08

	Total Missouri elementary schools not making adequate yearly progress (N = 565)		Missouri study sample schools (n = 28)				
Characteristic	**Mean**	**Standard deviation**	**Mean**	**Standard deviation**	**Difference**	**Test statistic**	***p*-value**
Reading achievement[a]							
Grade 3	630.28	35.24	629.52	39.86	–0.76	–0.75	.45
Grade 4	648.93	31.95	650.19	36.34	1.26	1.35	.18
Grade 5	665.51	31.44	664.33	34.96	–1.18	–1.32	.19
Mathematics achievement[a]							
Grade 3	613.64	33.00	614.67	38.17	1.03	1.07	.29
Grade 4	636.10	30.98	638.61	35.33	2.51	2.76	< .01***
Grade 5	652.62	36.92	651.22	45.41	–1.41	–1.22	.23
Students per school[b]	367.41	176.86	373.96	135.32	6.55	–0.19	.85
Students per teacher[b]	13.00	2.42	14.95	2.86	1.95	–4.14	< .01***
Students eligible for free or reduced-price lunch (percent)[b]	54.60	24.62	61.62	24.55	7.02	–1.47	.14
Student population (percent)[b,c]							
White	63.91	36.01	60.09	38.87	–3.82	0.55	.59
Black	29.04	35.08	31.28	41.76	2.24	–0.33	.74
Hispanic	4.81	9.20	5.94	11.08	1.13	–0.63	.53
Asian	1.96	3.19	1.42	2.16	–0.54	0.90	.37
American Indian	0.28	0.46	1.28	1.76	1.00	–8.74	.01***

***significant at p = .01.

Note: Schools were classified as "not making adequate yearly progress" if they did not make adequate yearly progress in one or more of the three years prior to the study (2005/06, 2006/07, or 2007/08).

a. Analyses for reading and mathematics scale scores were one-sample t-tests with state-level mean score by grade as population values.

b. Analyses for school demographics were t-tests between group means.

c. Values may not sum to 100 percent because of rounding.

Source: Missouri Department of Elementary and Secondary Education 2008a, 2009a, 2010a; U.S. Department of Education, National Center for Education Statistics 2008.

Compared with the larger population of Missouri elementary schools not making AYP, Missouri study sample schools showed no statistically significant differences with regard to Title I and school urbanicity (table 2.7).

Table 2.7 Baseline comparison of Missouri elementary schools not making adequate yearly progress and study sample schools on Title I, urbanicity, and adequate yearly progress status, 2007/08

	Total Missouri elementary schools not making adequate yearly progress (*N* = 565)		Missouri study sample schools (*n* = 28)			
Characteristic	***N***	**Percent**	***n***	**Percent**	**Test statistic**	***p*-value**
Schools receiving Title I						
Title I–eligible school	470	83.19	20	71.43	2.57	.11
Schoolwide Title I	233	41.24	15	53.57	1.67	.20
School urbanicity						
City	167	29.56	10	35.71		
Suburb	161	28.50	11	39.29		
Town	52	9.20				
Rural	185	32.74	7[a]	25.00[a]	4.48	.21

Note: Analyses were chi-square tests between percentages. Schools were classified as "not making adequate yearly progress" if they did not make adequate yearly progress in one or more of the three years prior to the study (2005/06, 2006/07, or 2007/08).

a. All categories were analyzed separately, but for the Missouri study sample schools the categories of town and rural were combined to preserve anonymity.

Source: Missouri Department of Elementary and Secondary Education 2009a; U.S. Department of Education, National Center for Education Statistics 2008.

Results from this study suggest that the low-performing schools that volunteered to participate in the study differed from the target population of low-performing schools in both Minnesota and Missouri. Thus, this study's results may not represent how other low-performing schools in Minnesota and Missouri would be impacted if they chose to implement Success in Sight.

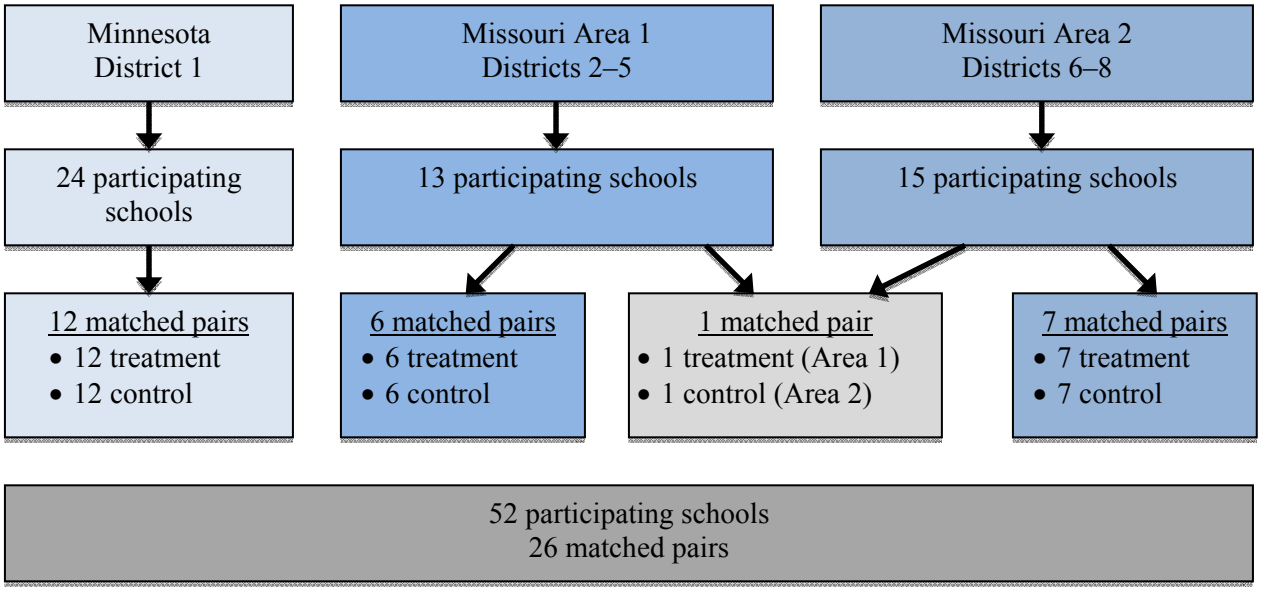
Random assignment of schools and baseline group equivalence

As part of the random assignment process, researchers created matched pairs of schools based on prior reading achievement and student eligibility for free or reduced-price lunch.[12] Researchers did not stratify the sample by AYP status failure or risk of failure before randomization because they thought matching on prior reading achievement would result in comparable mixes of failing and at-risk schools in the treatment and control samples. The matching process began when interested and eligible schools returned signed memoranda of understanding to the research team on a rolling basis. Researchers grouped participating schools by district, then ranked schools in each group according to student reading scores, first, and then by the percentage of students eligible to receive free or reduced-price lunch. Researchers then created matched pairs using the

[12] Researchers chose eligibility for free or reduced-price lunch as a matching variable based on previous research by Abbott and Joireman (2001) indicating that low income explains 12–29 percent of the variance in academic achievement.

nearest neighbor process in which data are matched based on the proximity of their values, in this case reading scores and free or reduced-priced lunch percentages. After schools with similar values were paired, researchers used the random sample procedure in SPSS to assign one school in each pair to the treatment group and its match to the control group. Each school had a 50 percent chance of assignment to the treatment or control group. Schools from Missouri Area 1 and Missouri Area 2 were grouped together because Missouri Area 1 and Missouri Area 2 had odd numbers of participating schools (figure 2.1). Schools completed baseline data collection following random assignment.[13] All schools remained in the study throughout the two-year intervention.

Figure 2.1 Random assignment of schools by area and matched pairs



As mentioned above, schools eligible for the study were those that had failed to make AYP, based on state AYP criteria, in at least one of the three years prior to the study (2005/2006, 2006/2007, and 2007/2008), or were at-risk of not making AYP in 2007/08 based on school personnel reports regarding changing student enrollment (such as an influx of English language learners). Researchers examined the distribution of schools according to prior AYP status by treatment and control condition. The difference between treatment and control schools in the distribution of schools based on their AYP category (at risk of not making AYP, not making AYP for one of three years, not making AYP for two of three years, or not making AYP for three years) was statistically significant ($p = .01$), indicating that the distribution of schools across these categories by treatment or control condition was not equal. Some 92 percent of treatment schools and 77 percent of control schools failed to make AYP in at least one of the three years prior to the study. Although the analytic models did not account for differences in AYP status (which could fluctuate within individual schools over the three years prior to pretest), each

[13] Because teachers completed the baseline teacher survey after random assignment had taken place, it is possible that their knowledge of group assignment impacted their responses. However, baseline teacher data were not used as outcome variables in any impact estimates. They were used to establish baseline equivalence and to construct school-level covariates used to increase the precision of the impact estimate.

benchmark impact estimate model included a cluster-level pretest covariate corresponding to the outcome of interest.

Although researchers randomly assigned schools to treatment and control conditions, it was possible that the two groups would differ on relevant characteristics at baseline. To test this, researchers compared baseline group data on school size, student free or reduced-price lunch eligibility, student ethnicity, and student reading and mathematics achievement scores (tables 2.8 and 2.9). Researchers also examined group equivalence for the three secondary outcomes related to teacher capacity for school improvement: data-based decisionmaking, purposeful community, and shared leadership (table 2.10). Comparisons were made at the school level because this was the level of random assignment and the level at which groups were expected to be equal regarding both measured and unmeasured characteristics. Researchers converted scale scores from the two states to *z*-scores to make cross-state comparisons.[14]

For the baseline comparisons between treatment and control groups, multilevel modeling analyses revealed no statistically significant differences between groups on mean student achievement *z*-scores, and *t*-tests comparing group means revealed no statistically significant differences between groups on school demographics (see table 2.8). Specifically, for school demographics, there were no statistically significant baseline differences between groups based on student ethnicity, percentage of students eligible for free or reduced-price lunch, number of students per school, or number of students per teacher.

[14] *z*-scores are standardized scores expressed in standard deviation units. The data analysis section in this chapter describes the process for converting scale scores to *z*-scores.

Table 2.8 Baseline comparison of treatment and control schools on achievement, size, and student characteristics 2007/2008

Characteristic	Treatment (schools = 26) Mean	Treatment (schools = 26) Standard deviation	Control (schools = 26) Mean	Control (schools = 26) Standard deviation	Total Mean	Total Standard deviation	Difference	Test statistic	*p*-value
Mean reading achievement (z-score) [a]									
Grade 3	–0.39	1.02	–0.44	1.10	–0.40	1.05	–0.05	0.46	.65
Grade 4	–0.32	1.10	–0.43	1.09	–0.37	1.09	–0.11	0.92	.36
Grade 5	–0.39	1.00	–0.37	1.10	–0.38	1.04	0.02	–0.13	.90
Total	–0.37	1.03	–0.41	1.10	–0.41	1.06	–0.04	0.38	.71
Mean mathematics achievement (z-score) [a]									
Grade 3	–0.39	1.03	–0.41	1.04	–0.40	1.04	–0.02	0.17	.87
Grade 4	–0.35	1.06	–0.41	1.13	–0.38	1.06	–0.06	0.45	.65
Grade 5	–0.37	1.09	–0.37	1.11	–0.37	1.09	0.00	–0.00	.99
Total	–0.37	1.06	–0.39	1.07	–0.39	1.07	–0.02	0.16	.87
Number of students per school (mean) [b]	410.88	121.29	373.35	133.84	392.12	127.87	37.53	1.06	.29
Number of students per teacher (mean) [b]	14.59	1.99	14.58	2.84	14.59	2.43	0.01	0.02	.98
Students eligible for free or reduced-price lunch [b] (percent)	69.89	22.95	70.63	23.23	70.26	22.86	–0.74	–0.12	.91
Student population (percent) [b, c]									
White	37.53	35.11	42.98	39.10	40.25	36.87	–5.45	–0.53	.60
Black	33.87	34.43	32.76	33.09	33.32	33.44	1.11	0.12	.91
Hispanic	10.54	12.90	8.76	9.18	9.65	11.11	1.78	0.57	.57
Asian	15.68	20.70	14.07	17.61	14.88	19.05	1.61	0.30	.76
American Indian	2.38	6.68	1.43	1.58	1.91	4.83	1.66	0.70	.49

Note: Black includes African American, Hispanic includes Latino, Asian includes Native Hawaiian or Other Pacific Islander, and American Indian includes Alaska Native.

a. Test statistics and p-values accounted for clustering of students within schools.

b. Analyses for school demographics were *t*-tests between group means.

c. Values may not sum to 100 percent because of rounding.

Source: Minnesota Department of Education 2008a; Missouri Department of Elementary and Secondary Education 2008a; U.S. Department of Education, National Center for Education Statistics 2008.

Chi-square tests between group percentages revealed no statistically significant differences between treatment and control groups regarding mean percentages for Title I eligibility, schoolwide Title I, or school urbanicity at baseline (table 2.9).

Table 2.9 Baseline comparison of treatment and control schools on Title I, urbanicity, and AYP status, 2007/08

	Treatment (schools = 26)		Control (schools = 26)		Total				
Characteristic	***n***	**Percent**	***n***	**Percent**	***N***	**Percent**	**Difference**	**Test statistic**	***p*-value**
Schools receiving Title I									
Title I–eligible school	21	80.77	21	80.77	42	80.77	0	0.00	1.00
Schoolwide Title I	18	69.23	18	69.23	36	69.23	0	0.00	1.00
School urbanicity									
City	16	61.54	18	69.23	34	65.38	–7.69		
Suburb	7	26.92	4	15.38	11	21.15	11.54	2.60	.46
Rural/Town[a]	3	11.54	4	15.38	7	13.46	–3.84		

Note: Values may not sum to 100 percent because of rounding. Analyses were chi-square tests between percentages.
a. All categories were analyzed separately, but the categories of town and rural were collapsed to prevent disclosure.
Source: U.S. Department of Education, National Center for Education Statistics 2008.

To assess potential baseline differences between treatment and control schools on teacher demographics, researchers conducted *t*-tests and multilevel modeling analyses (table 2.10). There were no statistically significant differences between groups on the percentage of teachers with a master's degree or on total years teaching. Additionally, there were no statistically significant differences between treatment and control school teacher groups on baseline scores for the school improvement practice measures in this study: data-based decisionmaking, purposeful community, or shared leadership. These measures were derived from two surveys: the Teacher Survey of Policies and Practices (Mid-continent Research for Education and Learning 2005) and the Collective Efficacy Scale (Goddard 2002). They are described fully in the Data Collection section of this report.

Table 2.10 Baseline comparison of treatment and control schools on teacher demographics, 2008

	Treatment (schools = 26, teachers = 819)		Control (schools = 26, teachers = 755)				
Characteristic	**Mean**	**Standard deviation**	**Mean**	**Standard deviation**	**Difference**	**Test statistic**	***p*-value**
Percent with a master's degree or higher[a]	64.93	15.24	65.64	15.24	–0.01	–0.17	.87
Total years teaching overall[b]	14.58	9.46	14.45	9.26	0.13	0.18	.86
Data-based decisionmaking score[b,c]	4.43	0.53	4.45	0.54	0.02	–0.28	.78
Purposeful community score[b,c]	3.32	0.65	3.34	0.62	-0.02	–0.32	.75
Shared leadership score[b,c]	3.81	0.83	3.90	0.83	0.09	–0.65	.52

a. Test statistics and *p*-values were from *t*-tests between group means.
b. Test statistics and *p*-values accounted for clustering of teachers within schools.
c. Scores based on the Teacher Survey of Policies and Practices (Mid-continent Research for Education and Learning 2005) and the 12-item Collective Efficacy Scale (Goddard 2002).
Source: 2008 teacher survey.

Student sample for impact analyses of primary outcomes

The impact analyses of primary outcomes examined the impact of Success in Sight on student reading or mathematics achievement in grades 3–5 after two years of implementation.[15] As described in chapter 1, Success in Sight is a systemic program aimed at improving school-level capacities, structures, processes, and attitudes that increase student reading and mathematics achievement. Although student movement into and out of schools occurs naturally, Success in Sight purports that school-level achievement is impacted regardless of student mobility. In other words, Success in Sight asserts that the program effects should emerge at the school level, regardless of how long individual students have been enrolled at any given time. Although individual students from treatment schools might be exposed to Success in Sight for varying durations (because some students move into and out of different grade levels or change schools over the study period), student mobility is not expected to undermine the overall school-level impacts of the program.

The impact analyses focused on school-level means of student achievement on 2009/10 state reading and mathematics assessments. The impact analyses did not track individual student performance from the 2007/08 school year to the 2009/10 school year, but instead included all students in grades 3–5 with available reading or mathematics achievement scores on the 2010 state reading and mathematics assessments. Available student 2008 baseline reading or mathematics achievement scores for grades 3–5 were used to create baseline covariates to increase the precision of the impact estimates.[16] Because impacts could have emerged for students enrolled in the same study schools throughout the study period before they emerged for the larger study sample (which includes students who have moved into the study sample or changed schools over the study period), researchers also estimated the impacts of Success in Sight on a subsample of students who did not change schools over the study period and who participated in baseline and posttest data collection.[17]

In keeping with Consolidated Standards of Reporting Trials recommendations for describing the flow of study participants from baseline to posttest (Campbell, Elbourne, and Altman 2004), this study describes how researchers established the impact analysis sample with regard to the 2008 baseline assessment, student mobility and missing data, and available 2010 posttest assessment data (figure 2.2). Students were nested in 52 participating schools, which were randomly assigned to treatment and control groups. All 26 treatment schools and all 26 control schools remained in the study from 2008 baseline assessment to 2010 posttest assessment. At the 2008

[15] Success in Sight developers note that because schools usually focus on one achievement area (that is, reading or mathematics) during the first two years of Success in Sight implementation, any impacts on achievement might be uneven across content areas. Therefore, the study's primary research questions examine the effect of Success in Sight on student achievement in either reading or mathematics, rather than both, after two years of implementation. One limitation of this study is that it did not examine the impact on reading achievement only in schools that selected reading as an area for improvement, and it did not examine the impact on mathematics achievement only in schools that selected mathematics as an area for improvement.

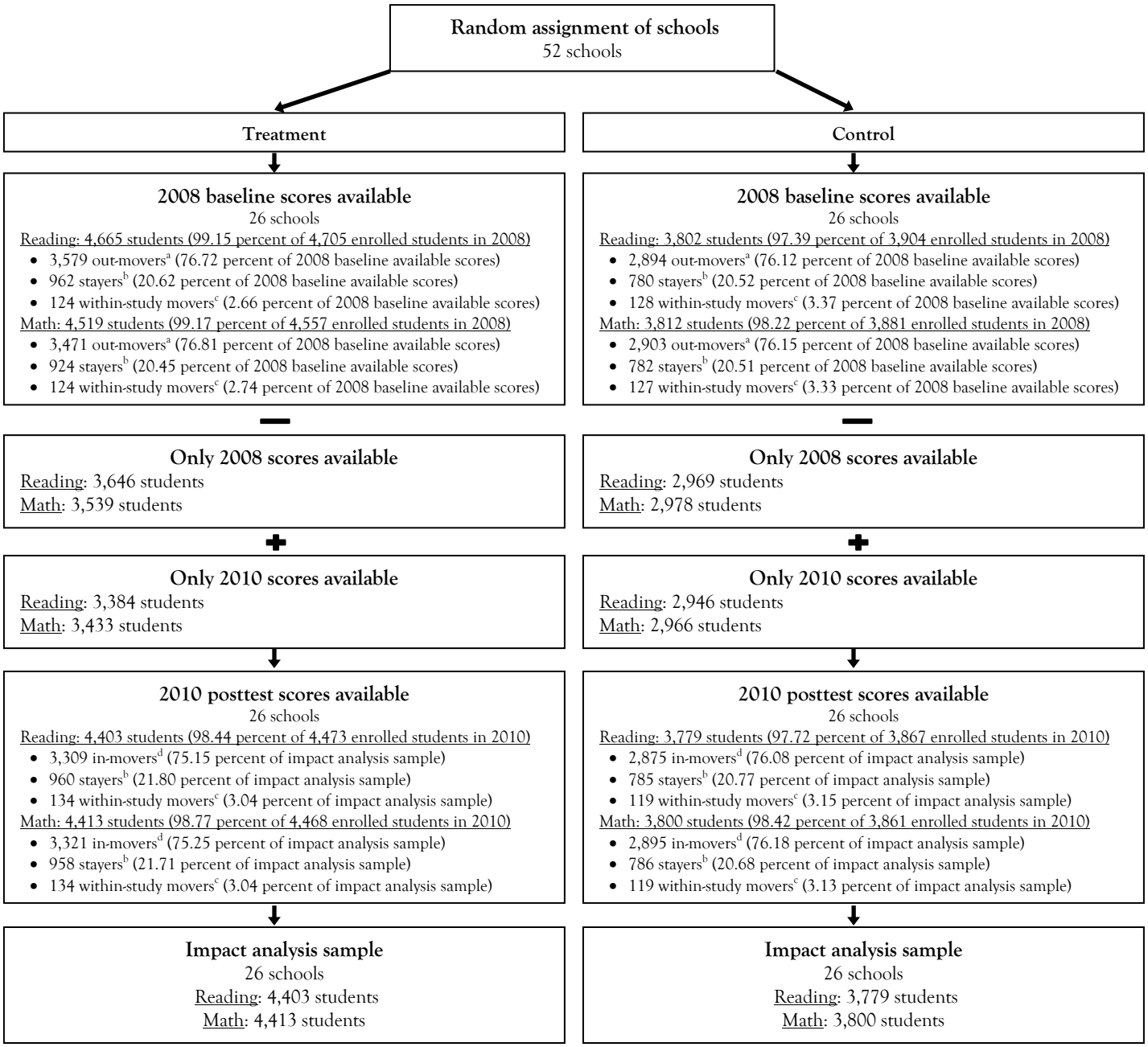
[16] The data analysis methods section of this chapter provides additional detail about the impact analyses, including the construction of the baseline achievement covariate.

[17] Appendix B presents results from a power analysis for estimating the impact for this subsample of students, and the data analysis methods section presented later in this chapter presents details about these analyses.

baseline assessment administration, 4,665 students in the 26 treatment group schools (99.15 percent of enrolled students) participated in the reading assessment, and 4,519 students (99.17 percent of enrolled students) participated in the mathematics assessment. And 3,802 students in the 26 control group schools (97.39 percent of enrolled students) participated in the baseline reading assessment, while 3,812 students (98.22 percent of enrolled students) participated in the baseline mathematics assessment. At the 2010 posttest assessment administration, 4,403 students in the 26 treatment group schools (98.44 percent of enrolled students) participated in the reading assessment, and 4,413 students (98.77 percent of enrolled students) participated in the mathematics assessment. And 3,779 students in the 26 control group schools (97.72 percent of enrolled students) participated in the 2010 posttest reading assessment, and 3,800 students (98.42 percent of enrolled students) participated in the posttest mathematics assessment (see figure 2.2).

The final student impact analysis sample for reading achievement includes 4,403 students in treatment schools and 3,779 students in control schools. The final student impact analysis sample for mathematics achievement includes 4,413 students in treatment schools and 3,800 students in control schools.

Figure 2.2 Student sample flow from baseline to posttest, 2007/08 to 2009/10



Note: Students with missing baseline or posttest scores were not included in analyses.

a. Out-movers were students who were in grades 3–5 within a study school at 2008 baseline, but moved out of the study by the 2010 posttest, by either moving out of the school or moving out of the eligible grade range.

b. Stayers were students who were enrolled in grades 3–5 in a study school at 2008 baseline and 2010 posttest, who did not change study schools between 2008 and 2010.

c. Within-study movers were students who were enrolled in grades 3–5 in a study school at 2008 baseline and 2010 posttest, who changed study schools between 2008 and 2010.

d. In-movers were students not in grades 3–5 within a study school at 2008 baseline, but moved into a study school or into grades 3–5 prior to the 2010 posttest.

Source: Adapted from the Consolidated Standards on Reporting Trials flow diagram (www.consort-statement.org).

Researchers created four categories to describe student movement during the study period: "out-movers," "in-movers," "within-study movers," or "stayers." Out-movers were students who were enrolled in grades 3–5 at a study school during the 2008 baseline data collection but moved out of the study before the 2010 posttest data collection, either because they were no longer in grades 3–5 or because they moved to a new school outside of the study. Thus, out-movers included students who moved out of the grade 3–5 study target range during the study period. For students whose scores contributed to the baseline reading covariate, out-movers accounted for 76.72 percent of the treatment group and 76.12 percent of the control group. For students whose scores contributed to the baseline mathematics covariate, out-movers accounted for 76.81 percent of the treatment group and 76.15 percent of the control group (table 2.11).

The impact analysis sample included in-movers, within-study movers, and stayers (see table 2.11). In-movers were students who were not enrolled in grades 3–5 at a study school during the 2008 baseline data collection, but moved into the study in 2009 or 2010 as students in grades 3–5 and therefore were eligible for the 2010 posttest assessments.[18] Thus, in-movers included students who moved into the grade 3–5 study target range during the study period. In-movers accounted for 75.15 percent of the treatment reading impact analysis sample, 76.08 percent of the control reading impact analysis sample, 75.25 percent of the treatment mathematics impact analysis sample, and 76.18 percent of the control mathematics impact analysis sample. Within-study movers were students in grade 3 at baseline and grade 5 at posttest who changed study schools during the study. Within-study movers accounted for 3.04 percent of the treatment reading impact analysis sample, 3.15 percent of the control reading impact analysis sample, 3.04 percent of the treatment mathematics impact analysis sample, and 3.13 percent of the control mathematics impact analysis sample. Stayers were students enrolled in study schools in grade 3 at 2008 baseline data collection and grade 5 at 2010 posttest data collection and who did not change schools over the course of the study. Stayers made up 21.80 percent of the treatment reading impact analysis sample, 20.77 percent of the control reading impact analysis sample, 21.71 percent of the treatment mathematics impact analysis sample, and 20.68 percent of the control mathematics impact analysis sample. A total of 99.06 percent of within-study movers and stayers were in grade 3 in 2008 and grade 5 in 2010.[19] There were no statistically significant differences between study conditions regarding the degree to which the impact analysis sample consisted of stayers, in-movers, or within-study movers (see table 2.11). Students moved into and out of grade levels each year. As a result, 66 percent of the reading and mathematics baseline student samples advanced grades and did not have posttest data because they were in grades 4 or 5 at pretest and grades 6 or 7 at posttest (out-movers). Likewise, 67 percent of the reading and mathematics posttest study samples were new students who did not have pretest data because

[18] In-movers included both within-study in-movers (students who were enrolled in study schools at baseline but not yet in eligible grades and therefore were exposed to the intervention for the entire study period if they were in the treatment group), as well as students who moved into study schools between baseline and posttest. Because researchers did not collect class rosters for students who were in grades 1 and 2 at baseline, it is not possible to distinguish in-movers who moved into eligible grades between 2008 baseline and 2010 posttest from in-movers who moved into study schools between 2008 baseline and 2010 posttest.

[19] A total of 17 stayers and two within-study movers moved by only one grade level between 2007/08 and 2009/10. The 17 stayers who moved by one grade level comprised 0.84 percent of all stayers and within study movers. The two within-study movers who moved by one grade level comprised 0.10 percent of all stayers and within-study movers.

they were in grades 1 or 2 at pretest and moved into grades 3 or 4 at posttest (in-movers). These samples reflect the natural fluctuation in sample populations because of student mobility into and out of grade levels, a common occurrence in most schools. Although the student sample fluctuated over the study period as students moved into and out of the grade 3–5 study target range, all classroom teachers in treatment schools were included in the Success in Sight intervention, regardless of the grade they taught.

Table 2.11 Student sample categories by study condition, 2007/08 and 2009/10

	Treatment		Control		Total		Test	*p*-
Student category	***n***	**Percent**	***n***	**Percent**	***n***	**Percent**	**statistic**	**value**
Reading								
Out-mover[a]								
Yes	3,579	76.72	2,894	76.12	6,473	76.45	0.39	.53
No	1,086	23.28	908	23.88	1,994	23.55		
In-mover[b,c]								
Yes	3,309	75.15	2,875	76.08	6,184	75.58	0.89	.35
No	1,094	24.85	904	23.92	1,998	24.42		
Within-study mover[c,d]								
Yes	134	3.04	119	3.15	253	3.09	0.05	.83
No	4,269	96.96	3,660	96.85	7,929	96.91		
Stayer[c,e]								
Yes	960	21.80	785	20.77	1,745	21.33	1.23	.27
No	3,443	78.20	2,994	79.23	6,437	78.67		
Mathematics								
Out-mover[a]								
Yes	3,471	76.81	2,903	76.15	6,374	76.51	0.46	.50
No	1,048	23.19	909	23.85	1,957	23.49		
In-mover[b,c]								
Yes	3,321	75.25	2,895	76.18	6,216	75.68	0.91	.34
No	1,092	24.75	905	23.82	1,997	24.32		
Within-study mover[c,d]								
Yes	134	3.04	119	3.13	253	3.08	0.03	.85
No	4,279	96.96	3,681	96.87	7,960	96.92		
Stayer[c,e]								
Yes	958	21.71	786	20.68	1,744	21.23	1.22	.27
No	3,455	78.29	3,014	79.32	6,469	78.77		

Note: Researchers calculated the percentages of student sample categories by classifying each student in the database as belonging to a mutually exclusive category based on their school enrollment during 2008 and 2010. Analyses were chi-square tests between percentages.

a. The out-movers consist of students enrolled in grades 3–5 at baseline who either left their schools or moved out of grades 3–5 prior to the posttest. These between-group comparisons refer to the 2008 baseline groups contributing to the baseline covariate.

b. The in-movers consist of students who were not enrolled in grades 3–5 within study schools at baseline. These students moved into grades 3–5 or moved into a study school after the baseline and were only eligible for the posttest assessments.

c. These between-group comparisons refer to the 2010 posttest impact analysis sample.

d. The within-study movers consist of students who were enrolled students in grade 3 at baseline and in grade 5 at posttest and changed study schools during the study.
e. The stayers consist of students who were enrolled students in grade 3 at baseline and in grade 5 at posttest and did not change schools over the course of the study.
Source: Minnesota Department of Education 2008a, 2010b; Missouri Department of Elementary and Secondary Education 2008a, 2010b.

To determine whether student mobility between study schools was related to original study condition, researchers examined the mobility patterns of within-study movers in the student analytical sample. For the impact estimates on reading and mathematics achievement, students moving to the same study condition before the posttest made up 46.27 percent of the within-study movers in the treatment group and 47.06 percent of the within-study movers in the control group. Students moving to a different study condition before the posttest made up 53.73 percent of the within-study movers in the treatment group and 52.94 percent of the within-study movers in the control group. The total number of within-study movers who completed posttest assessments did not differ by assessment type. There were no statistically significant differences between study conditions regarding the mobility patterns of within-study movers (table 2.12).

Table 2.12 Within-study student mobility patterns by study condition, 2009/10

	Treatment		Control		Total		Test	*p*-
Within-study mover	***n***	**Percent**	***n***	**Percent**	***n***	**Percent**	**statistic**	**value**
Reading								
Moved to same study condition at posttest	62	46.27	56	47.06	118	46.64		
Moved to different study condition at posttest	72	53.73	63	52.94	135	53.36	0.00	1.00
Total	134	100.00	119	100.00	253	100.00		
Mathematics								
Moved to same study condition at posttest	62	46.27	56	47.06	118	46.64		
Moved to different study condition at posttest	72	53.73	63	52.94	135	53.36	0.00	1.00
Total	134	100.00	119	100.00	253	100.00		

Note: Analyses were chi-square tests between percentages. The within-study movers consist of students who were enrolled students at baseline and at posttest and changed study schools during the study.
Source: Minnesota Department of Education 2008a, 2010b; Missouri Department of Elementary and Secondary Education 2008a, 2010b.

The Success in Sight schoolwide approach is designed to support student achievement irrespective of the natural inflow and outflow of students.[20] Researchers examined the mobility patterns of within-study movers to determine the number of students moving from a treatment school to a different treatment school, from a control school to a different control school, from a treatment school to a control school, or from a control school to a treatment school. There were 253 within-study movers, accounting for 3.09 percent of the reading student impact analysis

[20] As indicated previously, Success in Sight purports that school-level impacts should emerge even though individual students might be exposed to the program for different timeframes because of student mobility.

sample and 3.08 percent of the mathematics student impact analysis sample (table 2.13). Within-study movers who changed study conditions between pretest and posttest account for 1.65 percent of the student reading impact analysis sample and 1.65 percent of the mathematics student impact analysis sample (n = 135).

Table 2.13 Contribution of within-study student mobility to student impact analysis sample, 2009/2010

	Reading assessment		Mathematics assessment	
Within-study mobility pattern	Number of students	Percentage of sample (N = 8,182)	Number of students	Percentage of sample (N = 8,213)
Treatment to treatment	62	0.76	62	0.75
Control to control	56	0.68	56	0.68
Treatment to control	63	0.77	63	0.77
Control to treatment	72	0.88	72	0.88
Total within-study mobility	253	3.09	253	3.08

Note: The student impact analysis sample is comprised of stayers, in-movers and within-study movers.
Source: Minnesota Department of Education 2008a, 2010b; Missouri Department of Elementary and Secondary Education 2008a, 2010b.

Teacher sample for impact analyses of secondary outcomes

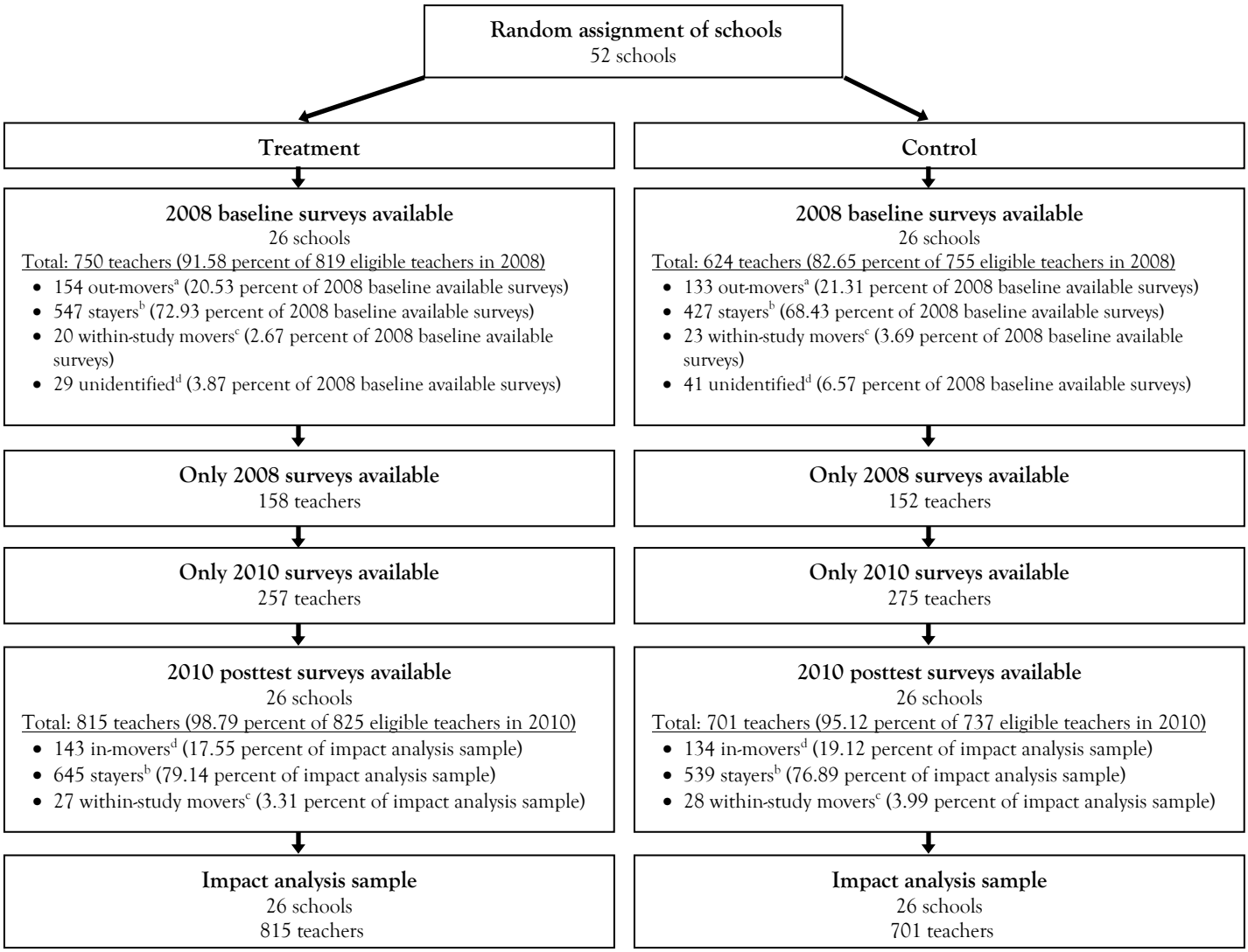
The impact analyses of secondary outcomes examined the impact of Success in Sight on teacher capacity for three school improvement practices—data-based decisionmaking, purposeful community, and shared leadership—after two years of implementation. Success in Sight aims to affect schoolwide teacher capacity regardless of naturally-occurring individual teacher mobility. This study focused on school-level teacher capacity as measured by a teacher survey in 2010. The impact analyses did not track changes in individual teacher capacity from the 2007/08 school year to the 2009/10 school year, but instead included all eligible teachers with available 2010 survey data. Teachers were considered eligible to participate in the 2010 posttest survey if they were members of the leadership team, classroom teachers, or specialists, and had appointments of 0.50 full-time equivalent or greater at the school at the 2010 posttest. Available teacher 2008 baseline survey scores were used to create baseline covariates to increase the precision of the impact estimates.[21]

The Consolidated Standards on Reporting Trials statement (Campbell, Elbourne, and Altman 2004) describes the flow of teachers in the study from baseline to posttest and documents how researchers established the teacher impact analysis sample with regard to the 2008 baseline survey, teacher mobility and missing data, and available 2010 posttest survey data (figure 2.3). In this study, teachers were nested in 52 participating schools, which the research team randomly assigned to treatment and control groups. As indicated previously, all 26 treatment schools and all 26 control schools remained in the study from 2008 baseline assessment to 2010 posttest assessment. At the 2008 baseline survey administration, 750 teachers in the 26 treatment group schools (91.58 percent of eligible teachers) participated in the survey. In the control group, 624 teachers in the 26 control group schools (82.65 percent of eligible teachers) participated in the

[21] The data analysis methods section of this chapter provides additional detail about the impact analyses, including the construction of the baseline teacher capacity for school improvement practices covariates.

baseline survey. At the 2010 posttest survey administration, 815 teachers in the 26 treatment group schools (98.79 percent of eligible teachers) participated in the survey. In the control group, 701 teachers in the 26 control group schools (95.12 percent of eligible teachers) participated in the 2010 posttest survey (see figure 2.3). The final impact analysis sample of 815 teachers from treatment schools and 701 teachers from control schools resulted from natural teacher mobility characteristic of all schools regardless of programming strategies.

Figure 2.3 Teacher sample flow from baseline to posttest, 2008 and 2010



Note: Teachers with missing baseline or posttest surveys were not included in analyses.
a. Out-movers were teachers who were in a study school at the 2008 baseline but moved out of the study by the 2010 posttest.
b. Stayers were teachers who were in a study school at the 2008 baseline and 2010 posttest, and who did not change study schools between 2008 and 2010.
c. Within-study movers were teachers who were in a study school at the 2008 baseline and 2010 posttest, and who changed study schools between 2008 and 2010.
d. Researchers did not track teacher responses at the baseline survey. Upon survey completion, researchers compared baseline survey timestamps and demographic information to identify baseline respondent matches between the eligible teacher roster and received surveys. A total of 1,304 baseline surveys were matched with individual names and 70 surveys could not be uniquely identified beyond the school name. As a result, researchers could not accurately break out missing survey data by group for the 2008 baseline survey.
e. In-movers were teachers who were not in a study school at the 2008 baseline, but moved into a study school prior to the 2010 posttest.
Source: Adapted from the Consolidated Standards on Reporting Trials flow diagram (www.consort-statement.org).

Researchers classified teachers, like students, into four categories: "out-movers," "in-movers," "within-study movers," and "stayers." Out-movers were teachers who were eligible at the 2008 baseline survey administration but moved out of the study before the 2010 posttest survey administration. In-movers were teachers who were not in a study school at baseline but moved into a study school by the 2010 posttest and were eligible to take the survey. Within-study movers were teachers who were eligible to participate at baseline and posttest but changed study schools over the course of the study. Stayers were eligible teachers at baseline and posttest who did not change schools over the course of the study.

The impact analysis sample for secondary outcomes included all four categories (table 2.14). Out-movers were only eligible for the baseline survey and accounted for 20.53 percent of the treatment group and 21.31 percent of the control group at baseline. In-movers were only eligible for the posttest survey and accounted for 17.55 percent of the treatment sample and 19.12 percent of the control sample at posttest. Within-study movers were eligible for the baseline and posttest surveys but changed study schools between surveys. At posttest, a total of 3.31 percent of the treatment sample and 3.99 percent of the control sample were categorized as within-study movers. The stayer sample was eligible for the baseline and posttest teacher surveys and represented 79.14 percent of the treatment sample and 76.89 percent of the control sample at posttest. There were no statistically significant differences between study conditions in the percentages of teachers representing out-movers, stayers, in-movers, or within-study movers. Additionally, there were no statistically significant differences between study conditions on patterns of teacher mobility (see tables 2.14 and 2.15).

Table 2.14 Teacher sample categories by study condition, 2008 and 2010

	Treatment		Control		Total		Test	*p*-
Teacher category	***n***	**Percent**	***n***	**Percent**	***n***	**Percent**	**statistic**	**value**
Out-mover[a]								
Yes	154	20.53	133	21.31	287	20.89	0.08	.77
No	596	79.47	491	78.69	1,087	79.11		
In-mover[b,c]								
Yes	143	17.55	134	19.12	277	18.27	0.52	.47
No	672	82.45	567	80.88	1,239	81.73		
Within-study mover[c,d]								
Yes	27	3.31	28	3.99	55	3.63	0.33	.57
No	788	96.69	673	96.01	1,461	96.37		
Stayer[c,e]								
Yes	645	79.14	539	76.89	1,184	78.10	0.99	.32
No	170	20.86	162	23.11	332	21.90		

Note: Researchers calculated the percentages of teacher sample categories by classifying each teacher as belonging to a mutually exclusive category on the basis of available teacher survey identifiers in 2008 and 2010. Analyses were chi-square tests between percentages.

a. The out-movers consist of teachers who were eligible at baseline but moved away before the posttest survey and were no longer eligible for the posttest. These between-group comparisons refer to the 2008 baseline groups contributing to the baseline covariate.

b. The in-movers consist of teachers who were not eligible at baseline. These teachers moved into one of the study schools after the baseline and were only eligible for the posttest survey.

c. These between-group comparisons refer to the 2010 posttest sample.

d. The within-study movers consist of teachers who changed study schools over the course of study, but were eligible for the baseline and posttest surveys.
e. The stayers consist of teachers who were eligible at baseline and were also eligible at posttest, and did not change schools over the course of the study.
Source: 2008 baseline and 2010 posttest teacher surveys.

To examine whether teacher movement between study schools was related to original study condition, researchers examined the mobility patterns of within-study movers in the impact analysis sample for secondary outcomes. Teachers moving to the same study condition before the posttest made up 48.15 percent of the within-study movers in the treatment group and 53.57 percent of the within-study movers in the control group. Teachers moving to a different study condition before the posttest made up 51.85 percent of the within-study movers in the treatment group and 46.43 percent of the within study movers in the control group. There were no statistically significant differences between study conditions in the mobility patterns of within-study movers (table 2.15).

Table 2.15 Within-study teacher mobility patterns by study condition, 2010

	Treatment		Control		Total		Test	*p*-
Sample type	***n***	**Percent**	***n***	**Percent**	***n***	**Percent**	**statistic**	**value**
Moved to same study condition at posttest	13	48.15	15	53.57	28	50.91		
Moved to different study condition at posttest	14	51.85	13	46.43	27	49.09	0.02	.90
Total	27	100.00	28	100.00	55	100.00		

Note: Analyses were chi-square tests between percentages. The within-study movers consist of teachers who changed study schools over the course of study, but were eligible for the baseline and posttest surveys.
Source: 2010 teacher survey.

Recognizing that teacher turnover occurs from year to year, the Success in Sight approach is designed to support systemic school improvement practices by introducing new teachers to improvement efforts as part of facilitators' mentoring of the leadership team. Researchers considered potential crossover effects and concluded that even if a teacher moved from a treatment school to a control school, the teacher would not be able to single-handedly implement the systemic schoolwide intervention program he or she was exposed to in the treatment school. Likewise, teachers (as well as students) might want to move from a control school into a treatment school if the latter is perceived as improving. A total of 55 teachers moved from a treatment school to a different treatment school, from a control school to a different control school, from a treatment school to a control school, or from a control school to a treatment school (table 2.16). Overall, within-study movers from the teacher sample comprised 3.63 percent of the teacher impact analysis sample.

Table 2.16 Contribution of within-study teacher mobility to impact analysis sample, 2010

Mobility pattern	Number of teachers	Percentage of teacher impact analysis sample (n = 1,516)
Treatment to treatment	13	0.86
Control to control	15	0.99
Treatment to control	13	0.86
Control to treatment	14	0.92
Total	55	3.63

Note. The teacher impact analysis sample consisted of stayers, in-movers, and within-study movers.
Source: 2008 and 2010 teacher survey.

Missing Data

This study experienced no missing data at the level of random assignment—the school level. All 52 participating schools provided baseline and posttest data. There was, however, student and teacher missing data in the study.[22] Because school was the unit of analysis, researchers explored differences in missing data rates at the school-level.

Student missing data occurred when students were enrolled in grades 3–5 in the study schools, and were therefore eligible to take the state assessments, but were missing reading or mathematics test scores. In this study, missing student scores for eligible students could be attributed to student absences on the day of testing. The overall school-level student missing data rate was less than 2 percent at baseline and posttest (table 2.17). At baseline, scores were missing from less than 1 percent of students in treatment schools and less than 3 percent of students in control schools. At posttest, scores were missing for less than 2 percent of students in treatment schools and less than 3 percent of students in control schools. (See the treatment of missing data section for details.) For reading and mathematics, there were no statistically significant differences between groups regarding student missing data rates (see table 2.17).

[22] As indicated previously in this chapter, this study's data collection focused on students and teachers present within a school at each data collection point rather than following students and teachers longitudinally because the effects of Success in Sight were intended to emerge regardless of the natural fluctuation of student and teacher populations.

Table 2.17 Comparison of the percentage of missing student assessment scores by condition, 2007/08 and 2009/10

	Treatment (schools = 26)		Control (schools = 26)		Total				
Characteristic	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation	Difference	Test statistic	*p*-value
Baseline missing scores (percent)									
Reading	0.89	1.33	2.68	5.14	1.78	3.83	–1.79	–1.72	.10
Mathematics	0.89	1.31	2.02	4.92	1.45	3.61	–1.13	–1.13	.26
Posttest missing scores (percent)									
Reading	1.46	1.45	2.33	4.45	1.89	3.30	–0.87	–0.94	.35
Mathematics	1.16	1.18	1.71	4.31	1.44	3.31	–0.55	–0.63	.53

Note: Analyses were *t*-tests between school-level means for the percentage of missing scores.
Source: Minnesota Department of Education 2008a, 2010b; Missouri Department of Elementary and Secondary Education 2008a, 2010b.

Wave nonresponse (that is, complete missing data at baseline or posttest) caused teacher-level missing data when teachers were eligible to complete a survey at one time point but did not do so, perhaps because they were on leave during the survey completion window or did not have sufficient time in their schedule to complete the online survey. Wave nonresponse is different from item nonresponse, wherein teachers complete a survey but choose not to answer individual survey items (Graham, Cumsille, and Elek-Fisk 2003; Puma et al. 2009). (See the treatment of missing data section in this chapter for details.)

Teacher missing data occurred when teachers were full-time staff members at a school and were therefore eligible for the teacher survey but did not complete it. Teacher missing data rates differed by measure and by time point. The overall school-level teacher missing data rates were 13.04 percent at pretest and 3.17 percent at posttest (table 2.18). Regarding differential missing data rates, at baseline, completed surveys were missing from 8.15 percent of eligible teachers from treatment schools and 17.92 percent of eligible teachers from control schools. Missing data rates for teachers from control schools were statistically significantly greater than for teachers from treatment schools at baseline. At posttest, completed surveys were missing from 1.38 percent of eligible teachers from treatment schools and 4.96 percent of eligible teachers from control schools. Missing data rates for teachers from treatment and control schools did not statistically significantly differ at posttest.

Table 2.18 Comparison of the percentage of missing teacher surveys by condition, 2008 and 2010

	Treatment (schools = 26)		Control (schools = 26)		Total				
Characteristic	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation	Difference	Test statistic	*p*-value
Baseline missing surveys (percent)	8.15	9.27	17.92	18.49	13.04	15.30	-9.77	-2.41	.02**
Posttest missing surveys (percent)	1.38	3.60	4.96	11.60	3.17	8.69	-3.58	-1.50	.14

** Significant at $p = .05$.
Note: Analyses were *t*-tests between school-level means for the percentage of missing surveys.
Source: 2008 and 2010 teacher survey.

As mentioned previously, missing data analyses examined mean school-level differences. Appendix C provides additional information regarding participation eligibility, nonresponses, and response rates by data collection instrument and administration period at the participant-level.

Data collection

Data collection occurred from March 2008 until June 2010 (table 2.19). Specifically, state reading and mathematics assessments, administered in the spring of 2008 and 2010, were used to gauge student achievement in reading and mathematics at baseline and posttest. A teacher survey was administered in 2008 and 2010 to gauge teacher capacity for school improvement practices at baseline and posttest. To describe the fidelity of Success in Sight delivery and participation in treatment schools, researchers collected data from Success in Sight professional development facilitators throughout the study period. Researchers also collected interview and focus group data from principals, leadership team members, and teachers to provide information about the local contexts of the treatment and control schools. Except for fidelity measures for treatment schools, all data collection efforts represented staff in treatment and control schools. Appendix C includes response rates for each instrument and administration period. Appendix D presents data collection instruments used in the study.

Table 2.19 Data collection schedule

Data collection method	Baseline	Intermediate[a]	Posttest
Treatment and control schools			
Student achievement			
State reading and mathematics assessments	March–May 2008	March–April 2009	March–May 2010
Teacher capacity for data-based decisionmaking, purposeful community, and shared leadership			
Teacher survey	June–October 2008[b]	March–June 2009	March–April 2010
Local school context			
Principal interviews	September–October 2008		
School focus groups with principals, leadership team members and teachers from each school	September–October 2008		
Phone interviews with the principal, one leadership team member and one staff member from each school			April–June 2010
Treatment schools only			
Implementation fidelity			
Program records[c]	September 2008–May 2010		
Electronic logs[d]	September 2008–May 2010		

a. Intermediate wave data were collected in 2009 but were not analyzed for this study.
b. The extended survey administration period accounted for time to identify site coordinators and administer the survey while school was in session rather than during the summer.
c. Program records included site visit summaries and attendance records and were completed each time Success in Sight facilitators visited a treatment site or conducted a large-group professional development session.
d. Electronic logs completed by Success in Sight facilitators included data related to professional development, content delivery, and school fractal experiences.

To help staff understand the nature and timeline of study activities and encourage their participation, researchers conveyed study information verbally and through printed research orientation materials at the beginning of the study period. Each treatment and control school identified a site coordinator who was responsible for working with researchers to coordinate and facilitate data collection activities. Researchers conducted study orientations with each school in the fall of 2008 during site visits, providing principals, site coordinators (if different from the principal), and staff participants with a description, instructions, and timeline for the online teacher survey, focus groups, and principal interviews.

In February 2010, researchers presented the posttest spring data collection activities, schedule, and instructions to all treatment schools during Success in Sight large-group professional development sessions. During these onsite visits, researchers also met with site coordinators in control schools that had response rates less than 70 percent on the 2008 teacher survey. The purpose of these face-to-face visits was to present the spring 2010 data collection schedule and process, stress the importance of study participation, and answer any questions. Researchers conducted webinars covering the same information as the onsite visits with site coordinators in

all other control schools in early spring. To promote participation in posttest data collection activities further, school district superintendents provided a letter of support to treatment and control schools reinforcing the district support of their participation. Researchers made this request of school districts because posttest data collection occurred 17 months after baseline data collection, and researchers wanted to make sure that study schools remained committed to participating in all study-related activities.

Treatment and control staff received stipends for participating in surveys, focus groups, and interviews. Teacher survey participants received a $25 stipend for each survey that they completed, for up to a total of $75 across the three administrations. Fall 2008 focus group participants each received a $25 stipend. Principal interview participants received a $35 stipend for each interview (2008 and 2010), and other interview participants (leadership team interview participants and nonleadership team interview participants) received a $25 stipend for participating in the spring 2010 phone interviews. Site coordinators received a $75 stipend for their help in coordinating each of the three data collection waves (baseline 2008, intermediate 2009, and posttest 2010). To increase survey response rates on the 2010 administration of the teacher survey, schools reaching a 100 percent survey completion rate received a $100 gift card for a school celebration. Site coordinators, who distributed the survey and tracked survey completion, could earn a $25 gift card for an 80 percent completion rate in their school or a $50 gift card for a 100 percent completion rate.[23]

Student achievement measures

The study's primary outcomes were measured by student achievement data from state reading and mathematics assessments for grades 3–5 from the Minnesota Comprehensive Assessment II[24] and the Missouri Assessment Program.[25] Coefficient alphas for the 2008 assessment administration ranged from .88 to .91 across all domain-specific assessments in reading and mathematics on the Minnesota assessment. Coefficient alphas ranged from .91 to .92 across all domain-specific assessments in reading and mathematics on the Missouri assessment. Appendix F provides additional information about these assessments.

Researchers used state assessments because of their established reliability and validity, because of their alignment to NCLB goals and to the reading and mathematics content taught across study schools, and because of annual testing procedures already in place. Researchers collected these data directly from the participating school district in Minnesota and from the Missouri Department of Elementary and Secondary Education. In addition to student achievement data for the 2007/08 (baseline) and 2009/10 (posttest) school years, school district data files included

[23] School districts did not provide researchers with staff email addresses, and therefore, researchers emailed the online survey link to site coordinators, who were responsible for distributing the survey and tracking responses.

[24] The Minnesota Comprehensive Assessment II Technical Manual and Yearbook (Minnesota Department of Education 2008b, 2008e) are available online at http://education.state.mn.us/MDE/Accountability_Programs/Assessment_and_Testing/Assessments/MCA/TechReports/index.html

[25] The Missouri Assessment Program Technical Reports (Missouri Department of Elementary and Secondary Education 2008b, 2010c) are available online at http://dese.mo.gov/divimprove/assess/tech/

student demographic information such as ethnicity, English language learner status, and special education status.

Researchers used vertically scaled scores by grade level and subject area from both state assessments to create *z*-scores for the primary outcome analysis (see data analysis section in this chapter for a full description of how researchers calculated *z*-scores). Upon receiving the state assessment data, researchers examined the means, standard deviations, and ranges of the data to identify potential erroneous values, outliers, ceiling effects, and floor effects. Before creating *z*-scores, researchers addressed the assumptions recommended by May et al. (2009) for studies using rescaled scores to combine impact estimates across grades and states. Appendixes E and F provide an overview of these assumptions and the ways in which this study's data addressed these assumptions.

Teacher capacity for school improvement practices

For this study, participating teachers were administered a survey to collect data on the three intermediate outcomes for teacher capacity for school improvement practices: data-based decisionmaking, purposeful community, and shared leadership. The teacher survey was developed from the Teacher Survey of Policies and Practices (Mid-continent Research for Education and Learning 2005) and the Goddard Collective Efficacy Scale (Goddard 2002). Appendix G provides detailed information about the development of the teacher survey as well as the psychometric properties of the scales used to measure the three intermediate outcomes. The teacher survey also included items for teacher background information such as years of teaching, education degree, and certification.

The Teacher Survey of Policies and Practices was used as a basis for measuring the intermediate teacher outcomes because it was designed for use in high-need schools, includes questions worded appropriately for the school as the unit of analysis, demonstrated high quality with regard to its technical characteristics, and addressed two of the three outcomes for this study in their entirety. The Collective Efficacy Scale (Goddard 2002) was included because it assesses collective efficacy of teachers at the group level and addresses an important aspect of purposeful community not covered in the Teacher Survey of Policies and Practices. Researchers ensured data quality for the teacher survey through the administration and format of the online survey. Researchers conveyed the eligibility criteria for survey recipients to site coordinators to ensure the correct staff received the survey. As mentioned previously, these criteria related to job position and employment status. To reduce missing critical item-level data, researchers designed the survey so that respondents were required to complete items identifying their district and school (for analysis purposes) and confirming that they work with students in an instructional capacity (to ensure that the participant was eligible to take the survey). Researchers divided the survey into multiple pages with a minimum of two and a maximum of four matrix questions per page so that respondents would not need to scroll down the page (potentially skipping items). Given the length of the survey, researchers also provided a bar at the bottom of each page indicating the percentage of the survey they had completed so that teacher participants could gauge their progress in completing the survey. This confirmed for teacher participants when they had completed all items and helped to avoid respondents submitting the survey without accessing all items.

Researchers administered this survey in an online format to all classroom teachers, instructional specialists, and leadership team members with employment status of 0.50 full-time equivalent or greater within each participating treatment and control school. Researchers set these criteria to ensure that teacher survey participants were in a position to influence student learning instructionally and implement school improvement practices. Starting in June 2008 (baseline) and March 2010 (posttest), researchers sent a link to the online survey to each site coordinator, along with a list of survey recipients. Site coordinators distributed the survey link to teacher survey participants who then completed and submitted the survey online. Researchers worked closely with site coordinators to monitor the completion rate of the survey and follow up with staff who did not respond to initial survey completion requests. Site coordinators followed up with respondents until they submitted a completed survey or the data collection window closed. At the 2010 posttest, 815 teachers from treatment schools (98.79 percent of those eligible) and 701 teachers from control schools (95.12 percent of those eligible) completed surveys.

Teachers responded to individual survey items using a 5-point scale (1 = strongly disagree, 2 = somewhat disagree, 3 = neither agree nor disagree, 4 = somewhat agree, and 5 = strongly agree).[26] The outcomes were all self-reports. Self-reports are limited by respondents' accuracy in recalling their practices or activities. Self-report measures are also susceptible to response sets or response styles. The two response sets most problematic to self-reports are social desirability and acquiescence. Social desirability occurs when respondents choose the response that they think will be seen as more socially desirable or more socially favorable. For example, teachers may over-report their capacity for school improvement practices because they think that this capacity is socially desirable. Acquiescence occurs when respondents choose the positive responses to items regardless of content. To some degree, acquiescence can be countered though the use of negatively valenced items, such as those in the Goddard Collective Efficacy scale.

The teacher survey and its administration were the same for the treatment group and the control group. Any limitations of the survey as a self-report likely would be relevant to both groups. Given that the treatment and control groups were formed via random assignment, the two groups can be expected to be equal in terms of the degree to which either response set—social desirability or acquiescence—were present. Comparisons between the two groups on these three self-report outcomes, therefore, were unlikely to be adversely impacted by the self-report nature of the survey.

Scores used in the impact analysis for each of the three intermediate outcomes (data-based decisionmaking, purposeful community, and shared leadership) were calculated by averaging the ratings for items comprising each scale. The coefficient alphas for the intermediate outcomes were .76 for data based decisionmaking, .89 for purposeful community, and .96 for shared leadership. These alpha coefficients exceed the What Works Clearinghouse standards for reliable outcome measures (What Works Clearinghouse 2008). Results from the confirmatory factor analysis showed that the three intermediate outcomes were highly related, with correlations between the three latent variables representing each outcome at .89 or higher. The confirmatory factor analysis results also suggested that the items measuring shared leadership provided a reliable and valid measure, but that the items for data-based decisionmaking and shared community may not provide reliable and valid measures of their respective constructs. Taken

[26] As described in appendix G, six of the survey items were reverse coded to adjust for negatively valenced items.

together, psychometric analysis results suggest that the teacher survey likely provided a measure sufficient for the purpose of this study, an estimation of the impact of Success in Sight on school level teacher capacity for school improvement practices. The limitations of the instrument, however, should be kept in mind when interpreting results.

Local context measures

To gain a better understanding of the local contexts that might influence school improvement practices in treatment and control schools, researchers collected baseline contextual data from September to October 2008 and end-of-study contextual data from April to June 2010. Baseline contextual data included site visit interviews with school principals and focus groups with the school leadership teams and a cross-section of school staff. End-of-study contextual data included phone interviews with the principal, a member of the leadership team, and a classroom teacher not on the leadership team in each treatment and control school. Response rates were 100.00 percent for 2008 baseline principal interviews and school focus groups and 99.36 percent for 2010 phone interviews. See appendix C for more information about response rates for these data collection activities.

The site visit interview and focus group protocols contained parallel questions designed to document the nature of school improvement activities, the local education context, and circumstances that might have helped or hindered school improvement efforts such as changing demographics and enrollments, or changes in state education policy. (See appendix D for instruments.) The focus groups that researchers conducted with a cross-section of school staff at treatment and control sites provided additional feedback on the extent to which school improvement activities and any subsequent changes had spread beyond the school leadership team. With the assistance of the site coordinator, researchers recruited teachers representing different grade levels, subject areas, and instructional duties (such as classroom teachers and counselors) who were not part of the school leadership team to participate in the focus groups in each treatment and control school. Focus groups with the school leadership teams did not include school principals, even though each principal was a member of the team, so that staff participants could share their experiences and perceptions freely without fear of jeopardizing their job or relationship with the principal. Instead, researchers conducted a separate interview with the school principal.

Near the end of the study, in spring 2010, researchers conducted 15-minute follow-up interviews with key contacts in each participating school to determine whether the local conditions documented during the baseline site visits had stayed the same or changed, and if so, how. Key contacts included the principal, a member of the leadership team, and a teacher not on the leadership team. Researchers randomly selected the leadership team member and teacher from the list of focus group staff members from 2008. If the selected leadership team member or teacher was no longer at the school or not responsive to the interview request after three email invitations and two phone messages, researchers randomly selected another staff member from the focus group roster. The interview items aligned with questions from the 2008 focus groups and principal interviews. Researchers provided staff with the interview questions prior to the phone call, so they could prepare for the interview.

Delivery and participation fidelity measures

Implementation fidelity for this study focused on the Success in Sight facilitators' delivery of professional development and leadership teams' participation in the program's professional development components (see chapter 1 for a description of the professional development delivery components).

Researchers developed fidelity indicators based on the Success in Sight delivery components and requirements for participation. Indicators of facilitators' fidelity to delivering the program as intended include conduct of six large-group professional development sessions, coverage of the required content during those sessions, and conduct of 10 onsite mentoring sessions. School requirements for fidelity of participation include forming leadership teams with a minimum of five members, attending six large-group professional development sessions, attending 10 onsite mentoring sessions, and completing at least two fractal experiences.

The purpose of the fidelity indicators was to ensure that the five key intervention activities were implemented: formation of a leadership team, attendance at six large-group professional development sessions, conduct of 10 onsite mentoring sessions with leadership team members, conduct of 10 onsite mentoring sessions with school principals, and implementation of a minimum of two fractal improvement experiences. (See table 3.1 in chapter 3 for criteria and indicators for adequacy of fidelity.)

Researchers gathered data on delivery and participation fidelity from Success in Sight program records and electronic logs. Success in Sight facilitators maintained program records throughout the two-year implementation period, which included documentation of the composition of the school leadership team, including the number of members and their roles in the school; leadership team attendance records for onsite professional development sessions; and site visit summaries. Site visit summaries included a record of the dates, duration, participation in and nature of the onsite mentoring sessions with leadership teams and school leaders, as well as the fractal improvement experiences that each school completed.[27] Because these data were self-reported by Success in Sight facilitators, they are not considered objective data. This study did not examine relationships among implementation fidelity and primary or secondary outcomes, but the reliance upon only self-report data to document implementation fidelity is still a limitation of this study.

Success in Sight facilitators completed electronic logs developed by researchers to document the content they delivered at each large-group professional development session and the fractal improvement experiences at each school. The first electronic log consisted of the Large-Group Professional Development Fidelity Checklist (see appendix D) to measure the extent facilitators delivered each professional development module (fully, partially, or not at all). The second electronic log tracked the topic and number of staff participants for each fractal improvement experience during the two-year period.

[27] As discussed in chapter 1, a fractal improvement experience is a small, manageable, and deliberate experience that enables participants to practice school improvement skills in areas of local need.

Data analysis methods

Researchers used benchmark and sensitivity analyses to address the study's primary and secondary research questions.[28] The analyses included data from all participating schools as they were randomized at the onset of the study. Consistent with the random assignment of schools to either the treatment or control group, researchers estimated impact analyses at the school level using multilevel modeling to account for the sources of variability in the data that result from the nested structure of the school environment. Analyses included two-tailed *t*-tests ($p = .05$) to assess the significance of the impact estimates as well as procedures to correct for multiple comparisons across impact estimates.[29] Researchers used HLM (Version 6.08) to conduct all multilevel modeling analyses.

Analyses of primary outcomes: impact on student achievement

The impact analyses of primary outcomes examined the effect of assignment to the Success in Sight intervention on student achievement after two years. The outcome variables were *z*-scores derived from student achievement scale scores in reading and mathematics from the spring 2010 administration of the Minnesota Comprehensive Assessment II and Missouri Assessment Program. The student sample for the impact analyses of primary outcomes included students enrolled in participating schools in grades 3–5 with available achievement data from the reading or mathematics state assessments at posttest.

Researchers estimated the intervention effects using two multilevel random-intercept models (one for each achievement domain) to account for sources of variability of students nested within schools. To create the student achievement *z*-scores used as outcomes, researchers followed May et al.'s (2009) guidance and transformed all achievement data into *z*-scores, separately for each grade, state, and assessment content area. The approach to transforming achievement data across multiple states into *z*-scores also was similar to procedures used by Carlson, Borman, and Robinson (2010).[30] First, researchers obtained the 2009/10 statewide means and standard deviations for reading and mathematics scale scores for grades 3–5 from the Minnesota and Missouri state departments of education. For each student in the study sample, researchers subtracted the appropriate grade-level state mean from each student's reading and mathematics scale score and divided it by the corresponding standard deviation to derive each student's reading *z*-score and mathematics *z*-score. Researchers carried out these procedures separately for each grade, content area (reading and mathematics), and state.

Each random-intercept multilevel model (that is, a model in which only the school intercept was allowed to vary randomly across schools) included the level 1 intercept as a random effect and

[28] The benchmark analyses are the analyses that determine whether Success in Sight has a statistically significant impact on the primary and secondary outcomes. Sensitivity analyses are the analyses that examine the robustness of the benchmark impact analyses to variations in the analytic models and samples.

[29] Before conducting analyses, researchers conducted several data cleansing and preparation procedures, including calculating and examining descriptive statistics, examining data ranges, and looking for outliers.

[30] In the Carlson, Borman, and Robinson (2010) study, outcome data were collected at the school level rather than at the student level, so they created standardized school level scores rather than standardized student level scores. This study uses student-level data to estimate the impact of Success in Sight.

the level 1 coefficients on the covariates as fixed effects. Level 1 (the student level) of each model included two dummy-coded indicator variables for posttest grade level (*GRADE 4* or *GRADE 5*), with grade 3 as the reference group.[31] These variables were grand mean–centered. The following equation represents each level 1 model:

$$Y_{ij} = \beta_{0j} + \beta_{1j}(GRADE\ 4)_{ij} + \beta_{2j}(GRADE\ 5)_{ij} + r_{ij}$$

where Y_{ij} is the posttest reading or mathematics performance of student i in a particular school j, β_{0j} is the mean posttest performance of students in school j, β_{1j} is the coefficient for the fixed effect for grade 4, β_{2j} is the coefficient for the fixed effect for grade 5, and r_{ij} is the random error for student i in school j.

Each level 2 (school-level) model included a dummy-coded variable to indicate assignment to treatment or control group (TREATMENT) to estimate the impact of the intervention on student achievement; researchers coded this variable as 0 for control and 1 for treatment. Each level 2 model also included baseline school size (SIZE) as a grand-mean centered integer variable, as well as indicator variables for blocks (that is, matched pairs) used in random assignment (BLOCK). *BLOCK 1* served as the reference group. The block variables were grand mean–centered. To account for baseline school-level differences in achievement, each level 2 model included a baseline achievement score (*PREACHIEVE*) as a grand-mean centered covariate.[32] The following equation represents each level 2 model:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(TREATMENT)_j + \gamma_{02}(SIZE)_j + \gamma_{03}(BLOCK\ 2)_j \ldots + \gamma_{029}(BLOCK\ 26)_j + \gamma_{030}(PREACHIEVE)_j + u_{0j},$$

$$\beta_{1j} = \gamma_{10},$$

$$\beta_{2j} = \gamma_{20}$$

[31] Although the inclusion of student-level demographic variables as covariates might have improved model specification, researchers did not include them for two reasons. First, prior to randomization researchers matched schools on student reading achievement scores and free or reduced-priced lunch status to account for variability in these areas. Second, the power analyses revealed that the models used would have the power of .80 to detect a standardized effect size of .20 or larger for benchmark analyses of primary outcomes and .30 or larger for the benchmark analyses of secondary outcomes. Therefore, student-level demographic variables were not included in these models.

[32] The model examining the impact of Success in Sight on reading achievement included a baseline reading covariate, and the model examining the impact of Success in Sight on mathematics achievement included a baseline mathematics covariate. To create these achievement covariates, researchers followed May et al.'s (2009) guidance and transformed achievement data into *z*-scores, separately for each grade, state, and assessment content area. Specifically, for each school in the study, researchers calculated grade-level means for reading and mathematics from the 2007/08 state assessments. From these school-specific grade-level reading and mathematics means, researchers subtracted the appropriate statewide mean and divided the resulting values by the corresponding standard deviations, yielding an overall mathematics and reading *z*-score for each school, which served as the baseline achievement covariates. Researchers did not include a student-level baseline covariate in the primary impact analyses because the students contributing to the baseline covariate were in grades 3–5 in 2008, and the posttest student sample included students in grades 3–5 in 2010. Thus, individual-level baseline covariate data were only available for the students who were in grade 3 at baseline and grade 5 at posttest.

where γ_{00} is the adjusted mean posttest performance for average-size, average-performing schools in the control group, while controlling for assignment block; γ_{01} is the effect of being in the treatment or control group, the treatment–control difference in adjusted mean school performance; γ_{02} is the regression coefficient for school size; γ_{03}–γ_{029} are the regression coefficients for the random assignment blocks; γ_{030} is the regression coefficient for school mean baseline achievement; and u_{0j} is the random error term for school j.

Researchers calculated effect sizes using Glass's *d* approach (Glass, McGaw, and Smith 1981). Appendix H provides additional information about the calculation of effect sizes. For each effect size resulting from the benchmark impact estimates of primary and secondary outcomes, researchers calculated a corresponding What Works Clearinghouse (2008) Improvement Index. This value characterizes the difference between the percentile ranks corresponding to the treatment and control-group means in the control-group distribution. It reflects the expected change in percentile rank for an average student in the control group if that student had participated in the treatment.

Analyses of secondary outcomes: impact on teacher capacity for school improvement practices

The impact analyses of secondary outcomes examined the impact of assignment to the Success in Sight intervention on teacher capacity for school improvement practices after two years. The outcome variables were teacher survey scores for the following three capacities for school improvement practices: data-based decisionmaking, purposeful community, and shared leadership. The teacher sample for the analyses included classroom teachers, specialists, and leadership team members from participating schools with employment status of 0.50 full-time equivalent or greater and who had available baseline or posttest survey data.

Researchers estimated the intervention effects on teacher capacity for school improvement practices using three separate multilevel models (one for each capacity for school improvement practice outcome) to account for sources of variability of teachers nested within schools. In each model, level 1 represented posttest teacher-reported capacity for school improvement in a specific practice (that is, data-based decisionmaking, purposeful community, and shared leadership). Each random-intercept multilevel model included the level 1 intercept as a random effect. The following equation represents each level 1 model:

$$Y_{ij} = \beta_{0j} + r_{ij}$$

where Y_{ij} is the posttest data-based decisionmaking, purposeful community, or shared leadership score of teacher i in a particular school j, β_{0j} is the mean posttest data-based decisionmaking, purposeful community, or shared leadership score of teachers in school j, and r_{ij} is the random error for teacher i in school j.

Each level 2 model included a dummy-coded variable to indicate assignment to treatment or control group (TREATMENT) to estimate the impact of the intervention on capacity for school improvement practices; researchers coded this variable as 0 for control and 1 for treatment. Each level 2 model also included baseline school size (SIZE) as a grand-mean centered integer variable as well as indicator variables for blocks (i.e., matched pairs) used in random assignment (BLOCK). BLOCK 1 served as the reference group. The block variables were grand-mean

centered. To account for baseline school-level differences in teachers' reported capacity for school improvement practices, the level 2 model included a baseline teacher score for data-based decisionmaking, purposeful community, or shared leadership (PRECAPACITY) as a grand-mean-centered covariate.[33] The following equation represents each level 2 model:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(TREATMENT)_j + \gamma_{02}(SIZE)_j + \gamma_{03}(BLOCK\ 2)_j \ldots + \gamma_{029}(BLOCK\ 26)_j + \gamma_{030}(PRECAPACITY)_j + u_{0j}$$

where γ_{00} is the adjusted mean posttest teacher capacity score for average-size control schools with average teacher capacity scores, while controlling for assignment block; γ_{01} is the effect of being in the treatment or control group, the treatment–control difference in adjusted mean teacher-reported capacity for school improvement practice (in data-based decisionmaking, purposeful community, or shared leadership); γ_{02} is the regression coefficient for school size; γ_{03}–γ_{029} are the regression coefficients for the random assignment blocks; γ_{030} is the regression coefficient for school mean baseline capacity for school improvement practice score; and u_{0j} is the random error term for school j.

Treatment of missing data

Although there was no school-level attrition in this study, there were missing student- and teacher-level data. Student-level missing data resulted from attrition when one or more assessment scores were not available for a student at either baseline or posttest. As indicated in table I1 (in appendix I), the amount of missing student-level data at any specific data point was less than 3 percent. Researchers used listwise deletion to address missing student data because it was not expected to bias the findings or result in a statistically significant loss of power, since the rate of missing data was less than 5 percent (Graham, Cumsille, Elek-Fisk 2003; Graham 2009).

Teacher-level missing data resulted from item-level nonresponse and attrition (wave nonresponse). The item-level missing data rates ranged from 25.37 percent to 42.63 percent (see table I2 in appendix I). To address item-level nonresponse at the teacher level, researchers implemented multiple imputation procedures for missing baseline and outcome data. Specifically, researchers implemented multiple imputation with chained equations (Van Buuren and Groothuis-Oudshoorn forthcoming) because it was a flexible procedure that handled data with different levels of measurement. Appendix I provides a more detailed discussion of the specific multiple imputation procedures used.

For wave nonresponse, attrition led to missing posttest teacher data for less than 5 percent of the cases. Therefore, researchers used listwise deletion to address missing posttest teacher data. At baseline, attrition led to missing data for 8.42 percent of treatment group cases and 17.35 percent of control group cases, making listwise deletion inappropriate for the baseline teacher data.

[33] The model examining the impact of Success in Sight on data-based decisionmaking included school mean baseline data-based decisionmaking score as a cluster-level covariate. Likewise, the model examining the impact of Success in Sight on purposeful community included school mean baseline purposeful community score as a cluster-level covariate. Finally, the model examining the impact of Success in Sight on shared leadership included school mean baseline shared leadership score as a cluster-level covariate. Researchers constructed the school-level covariates by calculating the mean scores for eligible teachers within each school. These means were calculated using available data from all eligible teachers who participated in the baseline survey.

Multiple imputation was also inappropriate because teacher baseline and posttest responses were not linked, making it unfeasible to use available data in an imputer's model. Furthermore, the use of a cluster-level, rather than individual-level, covariate precluded the use of the dummy variable method to address these missing data. Therefore, the impact models for secondary outcomes included cluster-level covariates calculated from available data. Appendix I provides more details about the extent of missing data and the methods used to address missing data for this study.

Corrections for multiple comparisons

This study's benchmark analyses of primary and secondary outcomes included multiple hypothesis tests. Testing multiple hypotheses within a domain can lead to an inflated Type I error, which can contribute to inaccurate conclusions about the study's findings (see Schochet 2008a).[34] Therefore, for this study, researchers followed Schochet's (2008a) and the What Works Clearinghouse's (2008) recommendations and made statistical corrections for multiple comparisons for this study's benchmark impact analyses (see appendix J). Specifically, based on the What Works Clearinghouse's protocol for addressing multiple comparisons, researchers used the Benjamini-Hochberg method. The Benjamini-Hochberg approach to multiple comparisons controls the false discovery rate, which is the probability that a statistically significant finding is a falsely rejected null hypothesis (Benjamini and Hochberg 1995; Schochet 2008a; What Works Clearinghouse 2008). Williams, Jones, and Tukey (1999) suggest that the Benjamini-Hochberg method is an appropriate method for addressing multiple comparisons across a wide range of situations.

Sensitivity analyses

Researchers conducted several sensitivity analyses to test the robustness of the benchmark estimates derived from the analyses of primary and secondary outcomes described above.[35] Researchers tested the robustness of benchmark estimates of primary and secondary outcomes to the use of a baseline covariate by running the analytic models for primary outcomes with no baseline achievement covariate and by running the analytic models for secondary outcomes with no baseline school improvement practice covariate.

In addition, researchers tested the robustness of the benchmark estimate of primary outcomes to the student sample by estimating the analytic model for student achievement using only the students who remained in the same school throughout the study.[36] This group of students is referred to as stayers. Although it would have been useful to test the sensitivity of findings to the inclusion of within-study in-movers (students who were enrolled in grades 1 and 2 at baseline and remained in the same school throughout the study), researchers did not have access to baseline enrollment rosters for students in grades 1 and 2, so it was not possible to identify

[34] A Type I error occurs when one incorrectly rejects a null hypothesis.

[35] For this study, the benchmark impact estimates were derived from the benchmark impact analysis models described previously in this chapter. The benchmark impact estimate determined whether Success in Sight was successful at impacting the specified outcomes for this study.

[36] Because random assignment occurred at the school level, and because each school shared unique within-school variance, these analyses included only students who remained in the same school throughout the study period.

students who were in grades 1 and 2 at baseline and remained in the same study school over the study period.

Finally, analyses were conducted to test the robustness of the benchmark analysis of primary outcomes to the methods used to combine estimates across state samples. For the benchmark analysis, researchers calculated *z*-scores from student achievement scale scores and included data from both states in each model. The sensitivity analyses used student achievement scale scores (instead of *z*-scores), estimated separate models for each state, and combined the results from the state-specific models meta-analytically. Appendix K provides details about the meta-analytic methods used to combine the results from these sensitivity analyses.

Chapter 3. Implementation of intervention

To supplement findings from the primary impact analyses, researchers measured fidelity to Success in Sight program delivery and participation criteria. Researchers also collected qualitative data on the local context of each participating school, including a description of "business as usual" school improvement efforts in control schools as compared with those in treatment schools during the study period. This supplemental information is intended to aid in interpreting impact results regarding the effects of participation in Success in Sight and to support discussion of whether contamination occurred between schools in the treatment and control groups.

This chapter describes Success in Sight implementation and presents implementation fidelity criteria and findings. The chapter continues with a depiction of schools' local contexts and concludes with a presentation of the cost to implement the Success in Sight intervention over two years.

Success in Sight implementation

Success in Sight is a systemic school improvement intervention intended to build the capacity of leadership team members, teachers, and school staff in five areas: data-based decisionmaking, purposeful community, shared leadership, research-based practices, and continuous improvement. Facilitators address these five capacity-building areas through four program delivery components: six large-group professional development sessions, 10 onsite mentoring sessions, ongoing distance support through phone or email, and fractal improvement experiences. In the Success in Sight approach, a pair of facilitators spends two years guiding each school leadership team through fractal improvement experiences using the five-stage continuous improvement process, which involves a "learning by doing" approach (see program overview in chapter 1). The early stages of capacity building focus on identifying needs and starting points, building relationships, learning about the context of sites, and deciding how improvement efforts will be organized and sequenced. Later stages involve assisting sites with development, implementation, and evaluation of improvement efforts, and planning for sustainability.

Success in Sight was delivered in this study by facilitators who worked directly with sites, rather than through a train-the-trainer model. When interventions have a strong, unique (preordained) technology, direct facilitation approaches increase the chances of near-term implementation (Reynolds et al. 2000). Direct facilitation by Success in Sight staff was an intentional design feature to reliably communicate expectations and the foundational concepts and practices of the program's components. This approach also follows a consortium model in which schools in the same geographic area come together to participate in the large-group professional development sessions where they can share experiences, accomplishments, and lessons learned. For this study, there was one consortium in Minnesota and two consortia in different geographic regions of Missouri. The onsite mentoring sessions, distance support, and fractal improvement experiences occurred in schools in between the large-group professional development sessions.

During the study, Mid-continent Research for Education and Learning (McREL) facilitators delivered the Success in Sight intervention to treatment schools as they typically would, with one exception. Typically, facilitators would work with district administrators to build their capacity

for systemic reform in the Success in Sight five capacity-building areas. In this study, because districts included both treatment and control schools, facilitators withheld district-level capacity building to prevent any potential for contamination across conditions. Therefore, no district personnel were involved in any Success in Sight professional development components.

Program delivery and participation fidelity criteria

Meta-analytical research on school improvement initiatives suggests that it is possible for systemic interventions to impact student achievement within two years if they are fully implemented by schools (Borman et al. 2003). Other research states that moderately complex educational change in elementary schools takes two to four years (Fullan 2007). A study of Accelerated Schools found it takes three to four years to detect measurable student achievement impacts (Bloom et al. 2001). Given that schools confront different challenges and vary in their capacity and readiness to address those challenges, it is also likely that schools will not progress at equal rates through any school improvement process. Accordingly, the timeframe in which measurable results are detectable might take longer than two years in sites with greater challenges to implementation.

Researchers developed indicators of adequate fidelity based on the Success in Sight delivery components and requirements for participation. Indicators of facilitators' fidelity to delivering the program as intended include: conduct of six large-group professional development sessions, and coverage of the requisite content during those sessions; and conduct of 10 onsite mentoring sessions (table 3.1, indicators 2–4). School requirements for fidelity of participation include: forming leadership teams with a minimum of five members, attending six large-group professional development sessions, attending 10 onsite mentoring sessions, and completing at least two fractal experiences (see table 3.1, indicators 1 and 5).

To assess delivery and participation fidelity, researchers used data from the program records and electronic logs that Success in Sight facilitators use as part of their typical delivery and site management practices. These records include site visit summaries, leadership team membership databases, and attendance records for both professional development and onsite mentoring sessions. Researchers also developed electronic logs that facilitators used to document the topics and activities covered at the large-group professional development sessions and to track fractal improvement experiences and staff participants by school.

Table 3.1 Program delivery and participation fidelity indicators, data sources, and criteria

Program component and indicator	Data sources	Adequacy criteria
1. *Leadership team formation.* Each school forms a leadership team with guidance from facilitators.	Leadership team Membership database	100 percent of treatment schools form a leadership team with a minimum of five members at each school, including principal and staff representation across grade levels and student service subgroups.
2. *Large-group professional development sessions.* Facilitators deliver three two-day professional development sessions (for a total of six sessions) covering planned content and completing planned activities.	Electronic log: content coverage Attendance records	80 percent of each module is delivered to treatment schools. 80 percent of all treatment leadership team members attend all six professional development sessions.
3. *Onsite mentoring for leadership team.* Facilitators conduct 10 half-day onsite mentoring sessions with school leadership teams that provide support for understanding the professional development content and process.	Site visit summaries Attendance records	80 percent of treatment leadership teams receive mentoring support from facilitators during 10 onsite meetings. 80 percent of all treatment leadership team members attend all 10 mentoring sessions with facilitator.
4. *Onsite mentoring for principals.* Facilitators meet with the school principal during each of 10 half-day onsite mentoring sessions to support learning and application of the professional development content and process.	Site visit summaries Attendance records	80 percent of all treatment principals receive mentoring support from facilitators during 10 onsite meetings. 80 percent of all treatment principals attend all 10 onsite leadership team meetings with facilitator.
5. *Fractal improvement experiences.* Facilitators provide support to school leadership teams to plan and complete at least two fractal improvement experiences of increasing magnitude.	Electronic log: fractal experiences	100 percent of treatment schools complete at least two fractal improvement experiences involving staff members and leadership team members.

Source: Mid-continent Research for Education and Learning (2008); personal communication, Danette Parsley, McREL Senior Director, and Ceri Dean, McREL Vice President of Field Services, March 21, 2009.

As mentioned in chapter 1, because McREL developed and facilitated implementation of Success in Sight, external researchers designed and conducted this study. McREL instituted firewall procedures that prohibited Success in Sight developers and facilitators from influencing evaluation activities, analyses, or reporting, and prevented researchers from sharing data with McREL. In keeping with the firewall protocol, the data collection for delivery and participation fidelity required a McREL research liaison to provide researchers with the program records and electronic logs for the two-year study period. Despite these firewall procedures, Success in Sight

facilitators could have introduced bias through their self-reports of implementation activities by presenting the program and its implementation in as positive light as possible. Although their documentation of program attendance, membership, and activities reflects typical practice, caution is warranted when interpreting these findings.

Program delivery and participation fidelity findings

This section presents findings based on the five fidelity indicators and the corresponding eight criteria described in table 3.1. Researchers calculated descriptive statistics (counts and percentages) based on implementation data for each criterion supporting the five fidelity indicators.

Indicator 1: leadership team formation

As part of Success in Sight, schools form leadership teams with the following criteria to ensure a diverse team composition: minimum of five members at each school, including the principal, and staff representation across grade levels and services for student subgroups (e.g., special education, English language learners, and reading intervention). All 26 schools formed leadership teams with a minimum of five members each. (The actual composition of leadership teams ranged from 6 to 10 members, with a mean size of 8.) All 26 leadership teams included the school principal as well as members representing two or more grade levels and different student services. Based on the implementation criteria, Success in Sight program records indicate that 100 percent of the 26 schools implemented the fidelity requirements regarding the formation of leadership teams (table 3.2). This percentage remained consistent throughout the study period.

Table 3.2 Percent of schools meeting fidelity criteria for leadership team composition

Criteria	Percent of schools meeting criteria (n = 26)
Minimum of five leadership team members	100.00
Principal a member of leadership team	100.00
Staff representing two or more grade levels	100.00
Staff representing different student services	100.00

Source: Success in Sight facilitator program records.

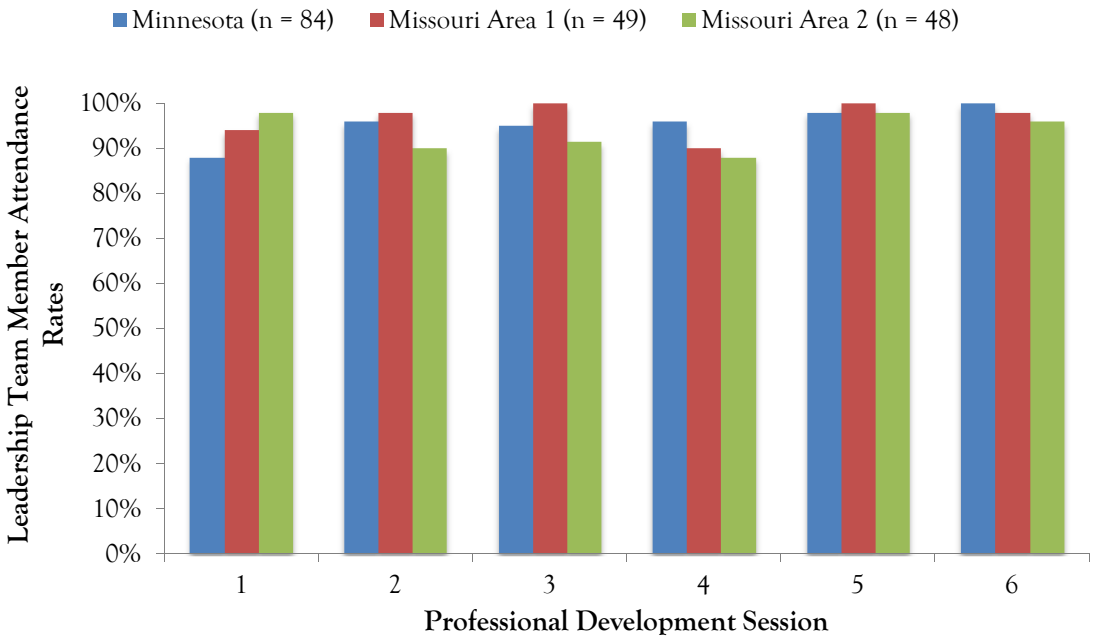
Indicator 2: large-group professional development sessions

Over the course of the two-year study, facilitators delivered three two-day professional development sessions, for a total of six sessions, to leadership teams within the three consortia (Minnesota, Missouri Area 1, and Missouri Area 2). Each session focused on one of the Success in Sight program modules described in chapter 1.

The program requires attendance from a minimum of five leadership team members, including the principal, at every session. In this study, 127 of 130 team members attended all six large-group professional development sessions (97.69 percent). Leadership team members met and exceeded the attendance fidelity criterion of 80 percent for the six large-group professional development sessions.

As presented previously, the actual size of leadership teams ranged from 6 to 10 members. Success in Sight facilitators requested that leadership teams limit the number of attending members to seven to keep group sizes manageable for collaborative activities within and across teams.[37] Based on attendance records for the actual number of team members attending sessions, researchers calculated attendance rates for each school and professional development session and then aggregated rates for each session by consortia. Because Success in Sight facilitators delivered this intervention component by consortium, results are presented accordingly. Missouri Area 1 had the highest attendance rate (96.60 percent) followed by Minnesota (95.63 percent) and Missouri Area 2 (93.48 percent) (figure 3.1). Team members attributed their absences from sessions to one of the following reasons: illness, vacation, left the school, or relinquished position on leadership team because of other time commitments and conflicting school responsibilities.

Figure 3.1 Percentage of leadership team members attending six professional development sessions by consortium



Source: Success in Sight attendance records.

[37] Attendance rates were calculated for schools based on the size of their leadership team, but capped at seven, because that was the requested limit set by Success in Sight facilitators. Therefore, attendance rates for team members with eight or more team members are based on a maximum expected of seven members (denominator) and maximum attended of seven members (numerator). Attendance rates for teams with fewer than seven members were calculated using the actual number of leadership team members in the denominator. For example, a team with five members would have a denominator of five and the number of members in attendance at each professional development session in the numerator. Five is the minimum number of participants required on a leadership team. No leadership teams had fewer than five members.

During each large-group professional development session, Success in Sight facilitators delivered one of six program modules (module 1 during the first session, module 2 during the second session, and so on). Each module is divided into segments that cover the module's content and activities. The number of segments per module range from 35 to 55, with 264 segments across all six modules. For each segment, lead facilitators used their electronic logs to report the level of content coverage by selecting one of four options: covered 80 percent or more, covered less than 80 percent, covered during a site visit session, or not at all. Researchers calculated the total segments for each response option by professional development session (1–6) and consortium (Minnesota, Missouri Area 1, Missouri Area 2), providing an indication of the extent to which facilitators delivered a module's overall content.

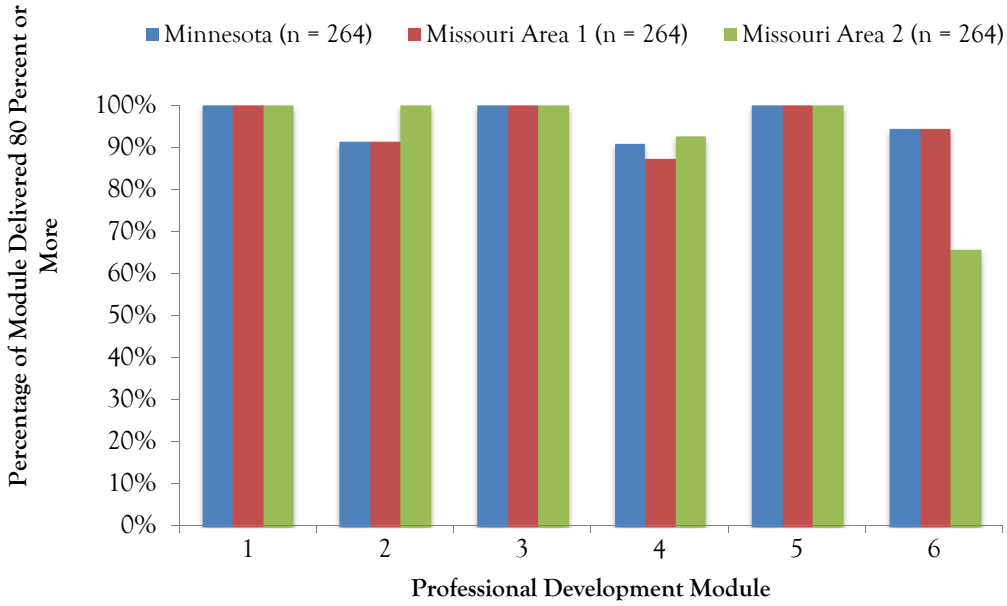
Results showed that facilitators delivered 80 percent or more of each module at 17 of the 18 professional development sessions across the three consortia (figure 3.2). By consortia, facilitators delivered 96.21 percent of the content (254 of 264 segments) in Minnesota, 95.45 percent of the content (252 of 264 segments) in Missouri Area 1, and 93.93 percent (248 of 264 segments) in Missouri Area 2. They delivered less than 80 percent (23 of 35 segments) of module 6 at session 6 for the Missouri Area 2 consortium.[38] Across consortia, Success in Sight facilitators delivered 100 percent of module 1, 94.29 percent of module 2, 100 percent of module 3, 90.30 percent of module 4, 100 percent of module 5, and 84.76 percent of module 6 (see figure 3.2).[39] These findings indicated that Success in Sight facilitators met and exceeded the criterion of delivering 80 percent of each module's content to leadership team members.

Across all six modules, facilitators delivered 85.60 percent of the content (226 of 264 segments). Facilitators delivered less than 80 percent of the intended content for 15 of the 264 segments, and they did not deliver eight segments during the large-group professional development sessions. Facilitators delivered 15 segments during follow-up onsite visits to schools.

[38] A severe winter storm prevented six of the seven schools in the Missouri Area 2 consortium from attending the sixth professional development session. Facilitators delivered the module content during subsequent onsite visits, ensuring that they delivered 80 percent of content for module 6.

[39] Percentages for each consortium are based on the number of segments in which facilitators delivered 80 percent or more of the content divided by the number of segments per session (module 1= 46 segments, module 2 = 35 segments, module 3 = 42 segments, module 4 = 55 segments, module 5 = 51 segments, and module 6 = 35 segments).

Figure 3.2 Percentage of module segments delivered at 80 percent or more by facilitators for each consortium



Source: Success in Sight electronic log for content coverage.

Indicator 3: onsite mentoring for leadership teams

As part of the Success in Sight delivery components, facilitators planned to conduct 10 half-day onsite mentoring sessions with each school leadership team. The purpose of these mentoring sessions was to support learning and application of the content delivered during the large-group professional development sessions. Facilitators reported through electronic logs that they delivered onsite mentoring sessions to 100 percent of schools, or 26 leadership teams, during the two-year study period; therefore, the fidelity criterion of 80 percent was met and exceeded. Facilitators' attendance records indicate that 100 percent of leadership team members (n = 130)[40] attended each of their school's 10 onsite mentoring sessions, which meets and exceeds the criterion that 80 percent of team members attend all 10 sessions.

Indicator 4: onsite mentoring for school principals

During the 10 half-day onsite mentoring sessions with school leadership teams, facilitators intended to meet separately with school principals to support their learning and provide guidance, demonstration, and feedback on fractal improvement experiences. The fidelity criterion required that 80 percent of the 26 principals receive one-on-one mentoring from Success in Sight facilitators during 10 onsite visits. Facilitators documented in electronic logs that they conducted 10 mentoring sessions each with all 26 principals in treatment schools (100

[40] The total number of leadership team members is based on the minimum fidelity requirement of five members per leadership team, which yields a total of 130 team members across 26 leadership teams.

percent). Therefore, the criterion of 80 percent of principals receiving 10 onsite mentoring sessions was met and exceeded. An additional fidelity criterion required principals to attend 10 half-day mentoring sessions with their leadership teams and Success in Sight facilitators. Of the 26 treatment principals, 25 principals (96.15 percent) attended all 10 mentoring sessions. One principal attended 9 of the 10 mentoring sessions. The criterion of 80 percent of principals attending 10 mentoring sessions with leadership team members and facilitators was met and exceeded.

Indicator 5: fractal improvement experiences

During large-group professional development sessions and onsite mentoring visits, facilitators provided guidance to school leadership teams as they planned and completed at least two fractal improvement experiences. In between sessions with facilitators, leadership team members would implement the plans for their fractal improvement experiences. The fidelity criteria indicator was met with 100 percent of schools completing at least two fractal improvement experiences that included staff members in addition to leadership team members. Schools completed three to eight fractal improvement experiences (mean = 5.46, standard deviation = 1.48) that included 7–115 staff participants per fractal. Leadership teams involved a mean of 29 staff participants (standard deviation = 15.26) in their fractals, with 13–80 participants by school. This indicates that leadership teams facilitated fractal improvement experiences of increasing magnitude by engaging staff participants outside of the leadership team.

Fractal improvement experiences focused on a broad range of areas, including reading, mathematics, teacher professional development, school culture, data-based decisionmaking, student behavior and engagement, parent involvement, and goals, among others (table 3.3). Of the 142 fractals completed across 26 schools, 39 fractals related to reading (27.46 percent), and 26 related to mathematics (18.31 percent). The other 77 fractal experiences (54.23 percent) focused on broader areas related to student achievement such as those mentioned previously. Within schools, the percentage of fractals focusing on reading ranged from zero to 100 percent. Twenty-five schools focused at least one fractal on reading. The percentage of fractals within schools focusing specifically on mathematics ranged from zero to 50 percent. Fifteen schools focused at least one of their fractal improvement experiences directly on mathematics.

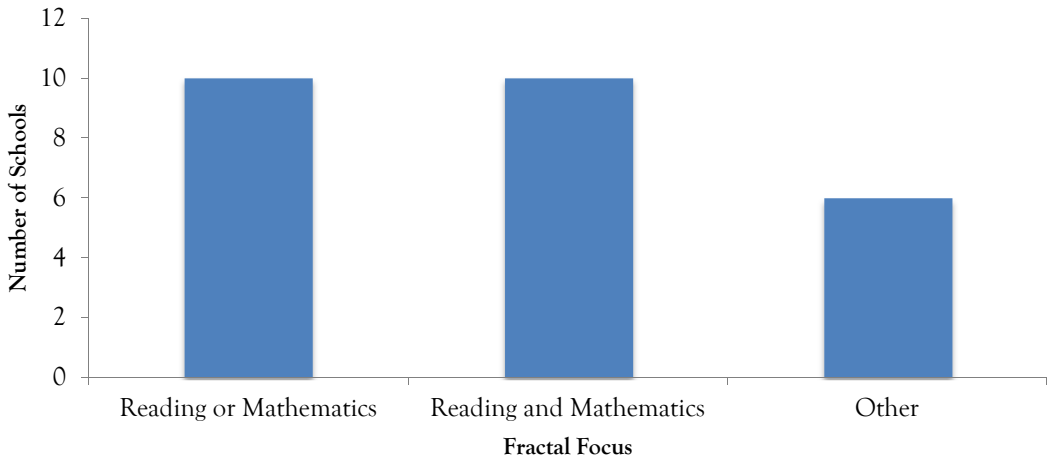
Table 3.3 Number of fractals completed with examples by category

Fractal category[a]	Number of fractals[b]	Examples of school focus for fractal improvement experiences
Reading	39	Improving instruction in guided reading, direct vocabulary instruction, reading comprehension strategies, vocabulary development, summarizing fiction and nonfiction, or higher order thinking; implementing Accelerated Reader, Reading Sight, Mondo Oral Language and Skill Blocks, Viva Vocabulary, or Words Their Way programs; establishing rituals and routines for Reader's Workshop; aligning pacing planning with student reading skills; using data to drive reading instruction; differentiating reading instruction; identifying essential reading skills.
Mathematics	26	Improving instruction in mathematics reasoning, mathematics facts, algebraic sense, mathematics concepts, mathematics games, direct vocabulary instruction in mathematics, or mathematics strand mathematical reasoning; flexible grouping in mathematics; differentiating mathematics instruction; differentiation in mathematics with peer observation and student feedback; using data to drive mathematics instruction.
Using data, assessments, and standards	15	Using data to inform assessment; data-driven decisionmaking; student-level data collection; targeting student performance based on formative assessment; how to examine student work; constructed response in assessment; using Mondo data to inform assessment; using standards to guide mini-lessons; using standards to guide innovations; response to intervention.
Student behavior	13	Behavior incentives and expectations; hallway behavior; Positive Behavioral Supports, Behavior Intervention Support Team.
Student engagement	12	Small-group engagement; student engagement using Concerns-Based Adoption Model Innovation Configuration; engagement during Viva Vocabulary; Service Learning; student engagement strategies; homework dues; individual student plans; differentiation during independent work time.
Teacher and faculty professional development and engagement	10	Professional Learning Communities; teacher team meetings; hot topic professional development; faculty engagements; oral language and academic vocabulary professional development; Accountable Talk; peer observations; walkthrough observations.
School culture	9	Setting and character; creating a culture of success; building community; hallway displays; fostering collaborative school culture; improved attendance; creating a responsive classroom.
Parent involvement	7	Parent contracts; Parents as Partners Night; parent volunteers; increasing parent involvement.
Goals	4	Goals and feedback, student goals, implement Mondo goals with fidelity, use data to inform assessment including grade-level goals, strategies, and agreements.
Other	7	Positive ending of Longfellow; HOT (Here on time); District Quality Improvement implementation with teachers; implementing mini-lessons; scaling up implementation schoolwide, using nonlinguistic representation.

a. Fractal categories were derived from a content analysis of fractal improvement experiences listed in facilitator electronic logs.
b. Number of fractal improvement experiences tallied across schools.
Source: Success in Sight facilitators' electronic logs.

Given that student achievement in this study is measured by student reading and mathematics achievement scores, researchers identified the number of treatment schools (n = 26) focusing 50 percent or more of their fractals on one specific content area (reading or mathematics) or both content areas. Ten schools focused 50 percent or more of their fractals on reading exclusively or mathematics exclusively with the majority focused on reading, and 10 schools focused 50 percent or more of their fractals on both reading and mathematics (figure 3.3). Six schools (categorized as "other" in figure 3.3) focused 50 percent or more of their fractals on areas such as student behavior, school culture, parent involvement, data collection, and teacher professional development.

Figure 3.3 Number of schools with 50 percent or more of fractals related to reading, mathematics, both reading and mathematics, or other focus area



Source: Success in Sight facilitators' electronic log.

Summary of Success in Sight delivery and participation fidelity findings

Success in Sight fidelity was defined as facilitators' delivery of the professional development components and leadership team members' participation in these components. Across five indicators, researchers measured fidelity of delivery and participation according to eight criteria. Success in Sight facilitators and leadership team members met all eight criteria (table 3.4):

- 100 percent of 26 schools formed a leadership team with a minimum of five members, including the principal and staff representing two or more grade levels and services for student subgroups.
- 80 percent of each module was delivered to all leadership teams.
- Of 130 leadership team members, 97.69 percent of leadership team members attended the six large-group professional development sessions.
- 100 percent of 26 leadership teams received facilitator mentoring during 10 onsite sessions.

- 100 percent of 130 leadership team members attended 10 onsite mentoring meetings.
- 100 hundred percent of 26 principals received facilitator mentoring during 10 onsite meetings.
- Of 26 principals, 96.15 percent attended the 10 onsite mentoring sessions with leadership team members and the facilitator.
- One hundred percent of 26 schools completed at least two fractal improvement experiences involving staff outside the leadership team.

Table 3.4 Number and percent of units meeting each of the eight required fidelity criteria

Program delivery and participation fidelity criteria	Number of units meeting criteria	Percent of units meeting criteria
Indicator 1: leadership team formation		
100 percent of schools (n = 26) meet four leadership team requirements	26 schools	100.00
Indicator 2: large-group professional development sessions		
80 percent of each module delivered (n = 6)	6 modules	100.00
80 percent of leadership team members (n = 130)[a] attend six large-group professional development sessions	127 members	97.69
Indicator 3: onsite mentoring for leadership teams		
80 percent of teams (n = 26) receive facilitator mentoring during 10 onsite meetings	26 teams	100.00
100 percent of team members (n = 130) [a] attend 10 onsite mentoring meetings	130 members	100.00
Indicator 4: onsite mentoring for principals		
80 percent of principals (n = 26) receive facilitator mentoring during 10 onsite meetings	26 principals	100.00
80 percent of principals (n = 26) attend 10 mentoring meetings	25 principals	96.15
Indicator 5: fractal improvement experiences		
100 percent of schools (n = 26) complete at least two fractal improvement experiences of increasing magnitude	26 schools	100.00

a. The total number of leadership team members is based on the minimum fidelity requirement of five members per leadership team.
Source: Researcher analysis.

In this study, the fidelity data findings indicate that facilitators and treatment schools met all the indicators of adequate delivery and participation. The analyses of fidelity confirm that the intended interactions between Success in Sight facilitators and leadership team members occurred as planned in treatment schools. The scope of this study did not include fidelity indicators to measure the thoroughness or quality of leadership team's implementation of their fractal improvement experiences and the continuous improvement process, both important components of the Success in Sight intervention for school improvement.

Local contexts and control schools

Researchers gathered qualitative data through focus groups and interviews with staff representatives at all treatment and control schools regarding the local contexts in which school improvement initiatives occurred. Focus group participants included teachers representing different grade levels, subject areas, and instructional duties (classroom teachers and counselors) who were not part of the school leadership team. Interview participants included the principal, a

leadership team member, and a teacher not on the leadership team (see chapter 2 for more information about participant selection). The contextual information provided by these interviewees helps with comparisons between treatment and control schools and aids in the interpretation of impact results regarding the effects of participation in Success in Sight. In addition, this information documents that contamination did not occur between treatment and control schools; that is, components and practices unique to Success in Sight were not implemented in control schools, and Success in Sight facilitators did not provide services to control schools.[41]

Researchers conducted baseline principal interviews and teacher focus groups in all treatment and control schools during the spring of 2008. At the end of the study period in the spring of 2010, researchers scheduled phone interviews with one principal, one leadership team member, and one classroom teacher in each of the 52 schools.[42] Researchers selected teacher participants for the phone interviews in 2010 if they had participated in the 2008 focus groups and were still at the same school, allowing researchers to capture teacher participants' observations of changes during the two-year study period (see appendix D for interview questions). All interviews lasted 10–20 minutes each. Researchers sent principals their interview questions prior to the phone interview so they could gather relevant information about student demographics, enrollment, and policy changes.

Comparisons between treatment and control schools included data on school characteristics (adequate yearly progress [AYP] status, student demographics, enrollment, and budget cuts during the study) and local education policy that could have influenced school improvement efforts (changes in school start times, grade-level configurations, curriculum, and assessment). Data from focus groups and interviews provided limited self-report information on the school improvement initiatives occurring in control schools during the study period.[43] Appendix L provides detailed information about these comparisons of the local context for treatment and control schools.

School characteristics

As presented in chapter 2, schools' AYP status during the three years prior to the Success in Sight intervention (2005/06, 2006/07, and 2007/08) was unequally distributed across treatment and control schools. During the three-year period preceding the study, 92 percent of treatment schools failed to make AYP at least one year, whereas 77 percent of control group schools failed to make AYP at least one year. Thirty-one percent of treatment schools did not make AYP all three years, whereas 12 percent of control schools did not make AYP all three years. Eight

[41] After year 1 of the study, the district administration transferred a principal in a treatment school to a control school for reasons unrelated to Success in Sight. During interviews with the principal and staff, it was clear that the principal was not using components of Success in Sight in the control school, nor did the principal intend to.

[42] Of the 156 possible phone interviews, researchers were able to conduct interviews with 155 participants for a participation rate of 99.35 percent. The missing interview was with a staff member who was nonresponsive to five interview requests up until the end of the data collection period. Researchers interviewed all principals.

[43] Because of budget limitations, researchers were unable to conduct independent or comprehensive measures of the nature and extent of school improvement practices in control schools.

percent of treatment schools and 23 percent of control schools had made AYP all three years but were reported by school personnel to be at risk of not making AYP.

Researchers also examined treatment and control schools' AYP status in reading and mathematics by state during the study period. Researchers collected schools' AYP status in reading and mathematics for the 2007/08 (baseline) and 2009/10 (posttest) school years from the Minnesota and Missouri state department of education websites (appendix L, table L1). In reading, 6 treatment schools and 11 control schools made AYP for 2007/08. By the end of the study period, 8 treatment schools and 12 control schools made AYP in reading for 2009/10. In mathematics, 11 treatment schools and 14 control schools made AYP during for 2007/08. By the end of the study period, 13 treatment schools and 16 control schools made AYP in mathematics.

Researchers documented student demographic and enrollment information for the 2007/08 and 2009/10 school years for treatment and control schools in each state (appendix L, table L2). The percentage of students in grades 3–5 qualifying for free or reduced-price lunch increased 3.27 percentage points in treatment schools (from 72.88 percent in 2007/08 to 76.15 percent in 2009/10) and 1.10 percentage points in control schools (from 75.77 percent in 2007/08 to 76.87 percent in 2009/10). The percentage of students in grades 3–5 in each ethnic group changed less than 1 percent from 2007/08 to 2009/10, with two exceptions for control schools. The percentage of White students increased 2.53 percentage points from 34.96 percent in 2007/08 to 37.49 percent in 2009/10, and the percentage of Hispanic students decreased 2.17 percentage points from 30.50 percent to 28.33 percent.

During interviews, principals reported if changes in student demographics, student enrollment or school budgets influenced their school improvement practices (appendix L, table L3). Among 26 treatment schools, two interviewees cited declining enrollment, three noted increased enrollment, and nine mentioned transiency issues with students moving in and out of their schools during each school year, which they believed influenced their school improvement efforts. Among 26 control schools, four interviewees thought that an increase in students qualifying for free or reduced-price lunch influenced their school improvements practices, two indicated that an increase in English language learner students influenced their school improvement practices, three perceived that increases in their Black student population influenced their school improvement efforts, and five reported that changes in student enrollment numbers influenced their school improvement efforts. Interviewees from six treatment schools and eight control schools indicated that budget cuts affected their school improvement efforts.

Local policy changes and restructuring

Consistent with the economic climate across the country, participating treatment and control schools dealt with districtwide budget cuts during the study period. Interviewees from six treatment schools and eight control schools indicated that budget cuts affected the availability of resources for school materials and equipment as well as full funding for support and instructional staff (such as music, art, and physical education teachers). The three main changes in local policies and practices that interviewees noted as affecting school improvement efforts involved curriculum, instruction, and assessment (appendix L, table L4). All 52 schools reported changes in curriculum between 2008 and 2010.

All treatment (n = 12) and control (n = 12) schools in Minnesota were mandated to implement a new reading curriculum, Mondo, during the study period as a supplement to their current reading curricula. Mondo is a K–5 comprehensive literacy program that includes guided reading, shared reading, intervention, and oral language curriculum materials. The program's Reader's and Writer's Workshops as well as its Skill Block require 150 minutes of instruction daily. The program also includes a train-the-trainer model of professional development in which national experts train literacy coaches, and they then train teachers. In Minnesota, Mondo is intended to support the district's goal of increasing student reading proficiency rates by 10 percentage points a year. It is also intended to support implementation of the district's Positive Schoolwide Behavior Model.

Beginning in 2009, all participating schools in Minnesota (12 treatment, 12 control) were required to administer the Phonological Awareness Literacy Screening assessment, a criterion-referenced assessment that educators can use as a screening, diagnostic, and progress-monitoring tool for K–3 students.

Changes in administration affected both treatment and control schools during the study period, according to feedback from all interviewees. Two treatment schools had a new principal for the 2008/09 school year, and one treatment school had a new principal for the 2009/10 school year. As members of their schools' leadership teams, these new principals participated in three large-group professional development sessions, five onsite mentoring sessions, and two to six school-level fractal improvement experiences each year. Three control schools in Minnesota had new principals for the 2009/10 school year. In Missouri, four control schools had new principals for the 2009/10 school year. Of these four, one control school had a new principal because it was targeted for turnaround and had 50 percent staff turnover during the 2009/10 school year. One of the Missouri control schools experienced the death of its principal and had a new principal for the 2009/10 school year. Two other Missouri control schools with new principals lost staff and funding because of a decrease in student enrollment. One interviewee reported that one school experienced a 27 percent student turnover. In addition to principal changes, the superintendent in one Missouri district was fired for illegal activities.

Three treatment and two control schools were affected by current or impending school reorganization. At the end of the 2009/10 school year, one treatment school was closing, and two treatment schools were co-locating with each other because of decreased student enrollment. Although this change did not take place during the study period, principals reported that the impending change influenced school culture and community and the way staff thought about school improvement efforts beyond the current school year. At the end of the 2009/10 school year, one Minnesota control school was closing, and one school was co-locating.

School improvement initiatives

Through interviews, researchers identified various initiatives that both treatment and control schools participated in to improve student achievement during the study period (appendix L, tables L5 and L6). The intent of Success in Sight is to offer a structure and process for implementing improvement initiatives rather than to supplant other school initiatives. The program encourages leadership teams to design their own improvement strategies and to tailor program components to the specific needs of a school. Participation in Success in Sight does not

preclude schools from participating in other improvement initiatives. Both treatment (n = 26) and control (n = 26) schools had school leadership teams that supported school improvement initiatives per the study's requirements (refer to eligibility criteria in sample recruitment section). Control schools were not required to participate in specific or formal school improvement initiatives, but conducted "business as usual." Interviewees representing treatment and control schools identified four common improvement initiatives: professional learning communities, leadership academies, Reading First, and response to intervention. As indicated in the prior section on fractal improvement experiences, treatment schools were able to focus their fractal improvement experiences on initiatives such as these.

Based on responses from all interviewed principals, 37 schools took part in professional learning communities: 12 treatment and 10 control schools in Minnesota and 7 treatment and 8 control schools in Missouri. Professional learning communities engage educators in working toward a common purpose with shared mission, vision, and values as well as high expectations for student learning. According to the district's website in Minnesota, professional learning communities involve students, teachers, and administrators in creating a school environment that promotes trust, risk taking, collegial exchange, conflict resolution, and continual learning with the ultimate goal of increasing student achievement.

Three principals at treatment schools and six principals at control schools reported participating in the Minnesota Leadership Academy, which trains principals in a research-based curriculum developed by the National Institute for School Leadership. The program began working with a cohort of principals in 2009 to build their capacities to be strategic thinkers, instructional leaders, and creators of a just, fair, and caring culture in which all students meet high standards. During interviews, principals in Missouri did not report participating in a leadership academy.

In Missouri, all treatment schools (n = 14) and control schools (n = 14) participated in Reading First as part of the No Child Left Behind Act of 2001 during the study period. This program addresses the five essential reading components—phonemic awareness, phonics, vocabulary, comprehension, and fluency—through explicit and systematic instruction for K–3 students. In addition to Reading First, 14 treatment and 14 control schools in Missouri implemented response to intervention, which the National Association of State Directors of Special Education defines as practices continually informed by student data and guided by scientifically proven instruction aligned to student needs and effective for the majority of students (National Association of State Directors of Special Education 2005). Although there are different interpretations of response to intervention implementation, it occurs through a multitiered service delivery process, which allows for an efficient allocation of classroom resources in which the students who need more focused instruction based on assessments receive it outside their core instruction.

Although no control schools participated in systemic school initiatives such as the Center for Effective Schools, Accelerated Schools, or Onward to Excellence, 11 interviewees from three control schools and eight treatment schools reported receiving services from the Missouri Regional Professional Development Centers, which provide professional development to educators in a variety of areas, including school improvement, assessment, professional learning communities, migrant and English language learner students, Reading First, and Positive Behavioral Supports, for example.

Researchers collected estimates of time investments for the various initiatives from published data. Based on these estimates, schools invest similar amounts of time in professional development activities for Success in Sight (26 treatment schools; 166 hours), leadership academies (three or fewer treatment and 6 control schools; 168 hours), and professional learning communities in Missouri (7 treatment and 8 control schools; 192 hours). In addition, Success in Sight schools spent an estimated 152 hours per school over two years implementing the intervention, compared with an estimated 343 hours per school per year implementing Reading First in Missouri (14 treatment and 14 control schools) and an estimated 340 hours per school implementing Mondo in Missouri (12 treatment and 12 control schools). More appropriate comparisons would be between implementation times for Success in Sight and the leadership academies, professional learning communities, and Regional Professional Development Centers because of their comprehensive nature, but published data were not available on the amount of time spent implementing strategies acquired through these professional development experiences (appendix L, table L7).

Cost of the intervention

The costs associated with the Success in Sight intervention reflect costs that would be incurred by a school if they chose to participate on a fee-for-service basis in the intervention as implemented for this study; that is, in the consortium model.[44] McREL estimated Success in Sight implementation costs according to three categories: large-group professional development, including the costs for substitute teachers and stipends to support teacher and principal participation in the large-group professional development sessions; materials and facilities for the large-group professional development sessions; and implementation costs associated with Success in Sight facilitator onsite mentoring sessions and principal and teacher participation time required during the onsite mentoring sessions.

The per-school, one-year cost of Success in Sight implementation was estimated at $99,702, (table 3.5). These costs might be underestimated because they do not include any additional planning time that teachers and other staff members might contribute after school.

[44] Success in Sight is appropriate for use in a single school. In this alternative model, schools do not interact with other schools, as they do in the consortium approach.

Table 3.5 Cost of the intervention per school with seven schools per consortium

Intervention component	Number of units	Cost per unit	Total cost per school
Large-group professional development			**$26,819**
Trainers			
Training time	6 days	$1,000	$6,000
Preparation and follow-up time	4.5 hours[a]	$125	$563
Transportation	3 trips	$1,000	$3,000
Teachers (seven per school)			
Training time	42 days	$240	$10,080
Principal			
Training time	6 days	$300	$1,800
Substitute teachers (seven per school)	42 days	$128	$5,376
Materials and facilities			**$5,620**
Materials	7 manuals	$680	$4,760
Facility			
Meeting rental space	.86 day[b]	$1,000	$860
Implementation			
Teacher participation time	30 hours	$272	$8,160
Principal participation time	30 hours	$4,910	$14,730
Mentoring			**$ 67,263**
McREL facilitator onsite visit	30 hours	$1,000	$30,000
Facilitator preparation and follow-up time	7.5 hours[c]	$125	$9,373
McREL facilitator onsite visit transportation costs	5 trips	$1,000	$5,000
Total cost per school for one year			**$ 99,702**

a. Preparation and follow-up time for trainers is 1.5 hours per professional development (1.5 x 3).
b. Facilities costs are split among seven schools (6 days/7 schools = .86).
c. Preparation and follow-up time for facilitator site visits is 1.5 hours per site visit (5 x 1.5).
Source: Mid-continent Research for Education and Learning budget.

Chapter 4. Impact results

To estimate the impact of Success in Sight, researchers conducted benchmark analyses of primary outcomes (student achievement in reading and mathematics) and secondary outcomes (teacher capacity for school improvement practices in data-based decisionmaking, purposeful community, and shared leadership). Researchers also conducted sensitivity analyses to assess the robustness of the benchmark impact analyses.

Benchmark analyses of primary outcomes: impact on student achievement

The benchmark analyses of primary outcomes examined the impact of Success in Sight on student achievement in reading and mathematics after two years, as measured by Minnesota and Missouri state reading and mathematics assessments. The outcome variables for these analyses were *z*-scores derived from 2010 student-level scale scores in reading and mathematics.[45] Chapter 2 describes the analytic models used for these analyses. Appendix M provides the raw vertically scaled means and standard deviations for baseline and posttest reading and mathematics achievement by grade, separately for each state. Appendix N displays the estimates of variance components from the null models, which allowed researchers to calculate intraclass correlation coefficients and confirm that multilevel modeling was an appropriate analytic approach for the impact estimates. Appendix O presents baseline means and standard deviations as well as complete results from the multilevel models.

The results from the benchmark impact analyses of primary outcomes on student achievement in reading and mathematics (table 4.1) indicate that Success in Sight did not have a statistically significant impact on student achievement in reading (adjusted posttest mean difference = –0.01, standard error = 0.03, $p = .75$) or mathematics (adjusted posttest mean difference = –0.06, standard error = 0.04, $p = .10$). The effect size for the impact on student achievement in reading was –0.01, which corresponded to a What Works Clearinghouse (2008) Improvement Index of 0.00. The effect size for the impact on student achievement in mathematics was –0.06, which corresponded to a What Works Clearinghouse (2008) Improvement Index of –0.02. As indicated previously, a What Works Clearinghouse (2008) Improvement Index value characterizes the difference between the percentile ranks corresponding to the treatment and control-group means in the control-group distribution. It reflects the expected change in percentile rank for an average student in the control group if that student had participated in the treatment.

[45] A *z*-score is a standardized score expressed in standard deviation units. For each student in the study sample, researchers subtracted the appropriate grade-level state mean from each student's reading and mathematics scale score and divided it by the corresponding standard deviation to derive each student's reading *z*-score and mathematics *z*-score.

Table 4.1 Impact of Success in Sight on student achievement outcomes, 2009/10

Regression-adjusted posttest measure	Treatment			Control			Estimated difference					
	Mean	Standard deviation	Sample size	Mean	Standard deviation	Sample size	Value	Standard error	95 percent confidence interval	*p*-value	Effect size[a]	Improvement index
Reading *z*-score[b]	–0.42	1.03	4,403	–0.42	1.02	3,779	–0.01	0.03	–0.07–0.05	.75	–0.01	0.00
Math *z*-score[c]	–0.48	1.10	4,413	–0.42	1.09	3,800	–0.06	0.04	–0.14–0.02	.10	–0.06	–0.02

Note: Results are from multilevel models that account for the nesting of students in schools. Differences between group means may not equal estimated differences because of rounding.

a. Calculated by dividing the estimated difference in means by the control group standard deviation (see appendix H).

b. Covers 4,403 students in all 26 treatment group schools and 3,779 students in all 26 control group schools.

c. Covers 4,413 students in all 26 treatment group schools and 3,800 students in all 26 control group schools.

Source: Minnesota Department of Education 2010b; Missouri Department of Elementary and Secondary Education 2010b.

Sensitivity tests for impact analyses of primary outcomes

Researchers conducted three sets of sensitivity tests to assess the robustness of the impact estimates derived from the benchmark analyses described above. Appendix P displays the analytic models used for each analysis, and appendix Q displays the complete results from the multilevel models.

Use of baseline achievement covariate

Researchers tested the robustness of the benchmark impact estimates of primary outcomes to the use of a baseline achievement covariate by running models with no baseline achievement covariate. The findings from these sensitivity analyses supported the benchmark findings of no statistically significant impact of Success in Sight on student achievement in reading (adjusted posttest mean difference = 0.03, standard error = 0.04, p = .52) or mathematics (adjusted posttest mean difference = –0.05, standard error = 0.05, p = .29). These findings indicated that the differences in primary outcomes between treatment and control schools were consistently not statistically significant, regardless of whether a cluster-level baseline covariate was included in the analytic models.

Student sample

To test the robustness of the benchmark impact estimates of primary outcomes to the student sample, researchers ran the benchmark model with an impact analysis sample comprised only of students who remained in the same school throughout the study (student stayers). The sensitivity analysis on reading achievement supported the benchmark finding of no statistically significant impact of Success in Sight on student achievement in reading (adjusted posttest mean difference = –0.06, standard error = 0.03, p = .10). The sensitivity analysis on mathematics achievement indicated that Success in Sight had a statistically significant negative impact on mathematics achievement (adjusted posttest mean difference = –0.11, standard error = 0.04, p = .02.), with student stayers in treatment schools demonstrating mean posttest mathematics achievement lower than that of student stayers in control schools. The negative impact on mathematics achievement remained statistically significant after applying the Benjamini-Hochberg correction.[46]

These findings indicated that the difference in the reading achievement outcome between treatment and control schools was consistently not statistically significant, regardless of whether the impact analysis sample included the entire benchmark impact analysis sample or only the student stayer sample. However the statistical significance of the impact estimate on student mathematics achievement was sensitive to the student impact analysis sample. Therefore, readers should interpret the benchmark finding on mathematics achievement with caution.

[46] To apply the Benjamini-Hochberg correction, researchers multiplied the original p-value's rank by the study's alpha level of .05 and divided this result by the number of student achievement outcomes (two). The result was greater than the original p-value of .02. Therefore, after application of the Benjamini-Hochberg correction, the impact estimate was still statistically significant. Appendix Q provides additional information about this calculation.

Impact analysis methods across states

Finally, researchers tested the robustness of the benchmark impact estimates of primary outcomes to the methods used to combine estimates across state samples, by using student achievement scale scores in reading and mathematics (instead of *z*-scores), estimating separate models for each state, and combining the results from the state-specific models meta-analytically. Appendix K details the meta-analytic methods used. These findings supported the benchmark findings of no statistically significant impact of Success in Sight on student achievement in reading ($p = .98$) or mathematics ($p = .82$). Specifically, the weighted mean effect size was –0.01 for reading and –0.07 for mathematics. These findings revealed that the differences in primary outcomes between treatment and control schools were consistently not statistically significant, regardless of whether researchers used the alternative meta-analytic method or the benchmark method. Appendix Q displays additional findings from these sensitivity analyses, including the effect sizes calculated for each outcome area for each state.

Benchmark analyses of secondary outcomes: impact on teacher capacity for school improvement practices

The benchmark analyses of secondary outcomes examined the impact of Success in Sight on teacher capacity for school improvement practices after two years, as measured by a comparison of posttest teacher surveys. Teacher survey responses were scored to construct data for three outcome variables that Success in Sight had identified as school improvement practices: data-based decisionmaking, purposeful community, and shared leadership. Chapter 2 describes the models used for these analyses. Appendix M provides the raw means and standard deviations for baseline and posttest teacher capacity for school improvement practices. Appendix N displays the estimates of variance components and intraclass correlation coefficients which enabled researchers to calculate intraclass correlation coefficients and confirm that multilevel modeling was an appropriate analytic approach for the impact estimates. Appendix R displays the complete results from the multilevel models.

Results from benchmark impact analyses of secondary outcomes (table 4.2) indicate that Success in Sight did not have a statistically significant impact on teacher capacity for data-based decisionmaking (adjusted posttest mean difference = 0.03, standard error = 0.02, $p = .13$), purposeful community (adjusted posttest mean difference = 0.03, standard error = 0.04, $p = .49$), or shared leadership (adjusted posttest mean difference = 0.16, standard error = 0.07, $p = .02$, not statistically significant after applying the Benjamini-Hochberg correction[47]). The effect size for the impacts on teacher capacity for data-based decisionmaking, purposeful community, and shared leadership were 0.06, 0.04, and 0.19, respectively. These effect sizes corresponded to respective What Works Clearinghouse Improvement Indices of 0.02, 0.02, and 0.08.

[47] To apply the Benjamini-Hochberg correction, researchers multiplied the original p-value's rank by the study's alpha level of .05 and divided this result by the number of school improvement practices (three). The result was not greater than or equal to the original p-value of .02. Therefore, after application of the Benjamini-Hochberg correction, the impact estimate was no longer statistically significant. Appendix R provides additional information about this calculation.

Table 4.2 Impact of Success in Sight on capacity for school improvement practices, 2008 and 2010

Adjusted Posttest Measure	Treatment			Control			Estimated difference					
	Mean	Standard deviation	Sample size	Mean	Standard deviation	Sample size	Value	Standard error	95 percent confidence interval	*p*-value	Effect size[a]	Improvement index
Data-based decisionmaking score	4.51	0.48	815	4.49	0.51	701	0.03	0.02	–0.01–0.07	.13	0.06	0.02
Purposeful community score	3.47	0.66	815	3.45	0.62	701	0.03	0.04	–0.05–0.11	.49	0.04	0.02
Shared leadership score	4.03	0.73	815	3.90	0.86	701	0.16	0.07	0.02–0.30	.02[b]	0.19	0.08

Note: Results are from multilevel models that account for the nesting of teachers in schools. Includes 815 teachers in all 26 treatment group schools and 701 teachers in all 26 control group schools. Differences between group means may not equal the estimated differences because of rounding.

a. Calculated by dividing the estimated difference in means by the control group standard deviation (see appendix H).

b. The result of the Benjamini-Hochberg calculation to correct for multiple comparisons was < .02. Therefore, this finding was not statistically significant after applying the Benjamini-Hochberg method to correct for multiple comparisons.

Source: 2008 and 2010 teacher survey.

Sensitivity tests for impact analyses of secondary outcomes

Researchers tested the robustness of the impact estimates derived from the benchmark analyses of secondary outcomes by running models with no baseline capacity for school improvement practices covariate.[48] Appendix S displays the analytic models used for each analysis, and appendix T displays the complete results from the multilevel models. When the baseline covariate was not included in the analyses, Success in Sight did not have a statistically significant impact on data-based decisionmaking (adjusted posttest mean difference = 0.02, standard error = 0.02, $p = .27$), purposeful community (adjusted posttest mean difference = 0.02, standard error = 0.04, $p = .63$), or shared leadership (adjusted posttest mean difference = 0.14, standard error = 0.07, $p = .05$, not statistically significant after applying the Benjamini-Hochberg correction[49]). These findings indicated that the difference in secondary outcomes between treatment and control schools were consistently not statistically significant, regardless of whether or not a cluster-level baseline covariate was used in the analytic models.

Summary

The results of this study indicate that Success in Sight did not have a statistically significant impact on student achievement in reading or mathematics after two years, nor did it have a statistically significant impact on teacher capacity for school improvement practices after two years.

These findings were supported by sensitivity analyses with no baseline cluster-level covariate as well as by sensitivity analyses that estimated impacts separately by state and combined results meta-analytically. The sensitivity analysis with only student stayers supported the benchmark impact estimate finding that Success in Sight had no statistically significant impact on student achievement in reading but showed that the finding of no statistical significance regarding mathematics achievement was sensitive to the impact analysis sample. Specifically, the impact analysis that only included student stayers showed that, on average, students from schools participating in Success in Sight had posttest mathematics scores statistically significantly lower than those of students from control schools. This finding was still statistically significant after researchers applied the Benjamini-Hochberg correction for multiple comparisons.

Together, these findings indicate that the benchmark impact estimate finding of no statistically significant impact on reading achievement was not sensitive to the inclusion or exclusion of a cluster-level baseline covariate, the two impact analysis methods used, or the student sample. The finding of no statistical significance regarding Success in Sight's impact on mathematics achievement was sensitive to the student samples included in the benchmark and sensitivity analyses. It would have been prudent to run a sensitivity analysis including stayers as well as within-study in-movers (students who were enrolled in grades 1 and 2 at baseline and remained

[48] Chapter 2 describes how the baseline capacity for school improvement practices covariate was constructed.

[49] To apply the Benjamini-Hochberg correction, researchers multiplied the original *p*-value's rank by the study's alpha level of .05 and divided this result by the number of school improvement practices (three). The result was not greater than or equal to the original *p*-value of .05. Therefore, after application of the Benjamini-Hochberg correction, the impact estimate was no longer statistically significant. Appendix T provides additional information about this calculation.

in the same school throughout the study). However, because researchers did not have access to baseline enrollment rosters, it was impossible to identify those students who were in grades 1 and 2 at baseline and remained in the same study school over the study period.

Findings from analyses of teacher capacity for school improvement practices revealed that Success in Sight did not have a statistically significant impact on teacher capacity for data-based decisionmaking, purposeful community, or shared leadership after two years. The sensitivity analyses with no baseline teacher capacity covariate supported the findings of no statistically significant group differences in teacher capacity for school improvement practices in data-based decisionmaking, purposeful community, or shared leadership after two years. These findings indicated that the benchmark impact estimates of Success in Sight on teacher capacity for school improvement practices in these areas were not sensitive to the inclusion or exclusion of a cluster-level baseline covariate.

Chapter 5. Exploratory analyses of relationships between school improvement practice outcomes and student outcomes

Researchers conducted exploratory analyses of the relationships between each of the study's primary outcomes—student achievement in reading and mathematics—and each of its secondary outcomes: teacher capacity for school improvement practices in data-based decisionmaking, purposeful community, and shared leadership.

The exploratory analyses built on the primary and secondary research questions by addressing the underlying theory of Success in Sight. As indicated in chapter 1, Success in Sight is a systemic school intervention that aims to raise student achievement scores by building teacher capacity in data-based decisionmaking, purposeful community, and shared leadership. The intervention is delivered directly to members of school leadership teams who, over time, engage more staff in the Success in Sight process. The program is based on the theory that greater teacher self-efficacy in these practices will lead to an increase in teacher capacity, which will improve teaching and ultimately raise student test scores. Therefore, the exploratory analyses examined whether there was a statistically significant relationship between the program's intermediate outcomes—that is, teacher capacity for school improvement practices—and student achievement in reading and mathematics.

For the exploratory analyses outcome variables, researchers used 2010 student achievement data from the Minnesota and Missouri state reading and mathematics assessments for students in grades 3–5. The independent variables were teacher capacity for school improvement practices as measured by scores from the 2010 posttest teacher survey. For each student outcome content area, researchers ran one multilevel model that included a baseline measure of the outcome variable at level 2, posttest capacity for school improvement practice scores at level 2, and covariates for 2010 student grade (at level 1), school size (at level 2), and blocking variables used in random assignment (at level 2). Appendix U displays the analytic model used for these analyses. The models did not include a variable to indicate assignment to treatment or control group because the intent of the analyses was to examine the relationship between intermediate teacher outcomes and student achievement outcomes within the entire study sample. The models for each student outcome included all three teacher capacity practices (data-based decisionmaking, purposeful community, and shared leadership) to examine the relative importance of each practice in contributing to variance in student outcomes. These analyses were relational, not causal. Thus, results should be interpreted as describing relationships rather than causal effects. Appendix V presents the complete results from the multilevel models.

Results of the exploratory analyses (table 5.1) pertain to unique relationships between each secondary outcome (teacher capacity for data-based decisionmaking, purposeful community, or shared leadership) and each primary outcome (student achievement in reading or mathematics). Findings revealed a statistically significant negative association between posttest teacher capacity for shared leadership and posttest student reading achievement ($p = .03$), indicating that higher teacher capacity in shared leadership was statistically significantly associated with lower student reading achievement scores. Neither teacher capacity for data-based decisionmaking ($p = .60$) nor purposeful community ($p = .77$) was statistically significantly associated with posttest student reading achievement. For mathematics achievement, there was a statistically significant negative association between posttest teacher capacity for data-based decisionmaking ($p = .04$)

and shared leadership ($p < .01$) and posttest student mathematics achievement, indicating that higher teacher capacity in data-based decisionmaking was statistically significantly associated with lower student mathematics scores, and higher teacher capacity in shared leadership was statistically significantly associated with lower student mathematics scores. Findings also revealed a statistically significant positive association between posttest teacher capacity for purposeful community and posttest student mathematics achievement ($p < .01$), indicating that higher teacher capacity in purposeful community was statistically significantly associated with higher student mathematics scores.

The purpose of these analyses was to explore whether or not there were statistically significant relationships between each of the study's primary outcomes and each of its secondary outcomes. Because these analyses sought to explore rather than confirm relationships, researchers did not conduct follow-up sensitivity analyses. Additionally, based on Schochet's (2008) recommendations that a multiple comparison correction be applied only to confirmatory analyses, researchers did not apply a multiple comparison adjustment to statistically significant findings that emerged from exploratory analyses.

Table 5.1 Relationship between capacity for school improvement practice outcomes and student achievement outcomes, 2007/08 and 2009/10

Independent Variable	Estimate	Standard error	*t*-Ratio	Degrees of freedom	*p*-value
Reading[a]					
Data-based decisionmaking	0.10	0.19	0.53	21	.60
Purposeful community	0.04	0.15	0.30	21	.77
Shared leadership	–0.16	0.07	–2.37	21	.03**
Mathematics[b]					
Data-based decisionmaking	–0.63	0.28	–2.21	21	.04**
Purposeful community	0.53	0.15	3.53	21	<.01***
Shared leadership	–0.28	0.09	–3.06	21	<.01***

**Significant at $p = .05$; ***significant at $p = .01$.

Note: Results are from multilevel models conducted separately for reading and mathematics achievement outcomes.

a. Includes 815 teachers and 4,403 students in all 26 treatment group schools and 701 teachers and 3,779 students in all 26 control group schools.

b. Includes 815 teachers and 4,413 students in all 26 treatment group schools and 701 teachers and 3,800 students in all 26 control group schools.

Source: Minnesota Department of Education 2008a, 2010b; Missouri Department of Elementary and Secondary Education 2008a, 2010b; 2010 teacher survey.

Chapter 6. Summary of study findings and limitations

The purpose of this study was to provide unbiased estimates of the impact of Success in Sight on student reading or mathematics achievement and teacher capacities for school improvement practices. The study was conducted during the 2008/09 and 2009/10 school years with 52 schools randomly assigned to the treatment or control condition. Over the course of two school years, 26 treatment schools implemented Success in Sight and 26 control schools implemented their usual school improvement practices. The study describes the local contexts of all schools and documents Success in Sight program delivery and participation in intervention schools.

Intervention implementation

Success in Sight facilitators provided consortia of school leadership teams with six large-group professional development sessions and 10 onsite mentoring sessions, as well as distance support between site visits and assistance with fractal improvement experiences of increasing magnitude. The large-group professional development sessions focused on building the capacity of leadership teams in five areas thought to be associated with school improvement: data-based decisionmaking, purposeful community, shared leadership, research-based strategies, and the continuous improvement process. Sessions also focused on strengthening school structures, processes, and attitudes to support and sustain systemic school improvement. Through onsite visits and distance support, facilitators assisted leadership teams in creating and implementing fractal improvement experiences that addressed local needs and issues related to student achievement. During the fractal improvement experiences, leadership teams were encouraged to apply lessons from professional development sessions and to participate in a continuous improvement process involving five stages: taking stock, focusing on the right solution, taking collective action, monitoring and adjusting, and maintaining momentum.

Eight criteria were developed to gauge fidelity of program delivery and participation during the study period. Four criteria focused on Mid-continent Research for Education and Learning facilitators' fidelity to delivering Success in Sight as intended: conducting six large-group professional development sessions, implementing a content module at each session, facilitating 10 onsite mentoring sessions and distance support with leadership teams, and providing principals with mentoring during the 10 onsite visits and ongoing distance support between sessions. Four criteria focused on school participation requirements for fidelity: forming leadership teams with a minimum of five members representing different student support and instructional areas, attending the six large-group professional development sessions, attending 10 onsite mentoring sessions, and completing at least two fractal improvement experiences involving staff participants not on leadership teams. Success in Sight facilitators' program records and electronic logs provided the data used to assess adequate program delivery and participation. Researchers were unable to conduct independent measures of implementation fidelity because of a decrease in the project scope of work during planning phases. Although the fidelity data reflect facilitators' typical documentation of their work with schools, it was self-report and was not validated by an independent source.

All fidelity indicators were met by all 26 treatment schools for this study. All treatment schools formed leadership teams with at least five members including the principal and staff representing two or more grade levels and services for student subgroups. Of the required 130 leadership team members (five per team), 97.69 percent of leadership team members attended all six large-group professional development sessions at which Success in Sight facilitators delivered a minimum of 80 percent of each program module (one module per session, six modules total). Success in Sight facilitators provided 10 of 10 onsite mentoring sessions to the 26 schools in which 100 percent of leadership team members and 96 percent of principals attended. All principals in each treatment school received at least 9 of 10 one-on-one mentoring sessions with a Success in Sight facilitator during these site visits.

The Success in Sight fractal improvement experiences offered leadership team members and school staff opportunities to apply lessons from the professional development sessions regarding data-based decisionmaking, purposeful community, shared leadership, research-based practices, and the continuous improvement process. The 26 treatment schools completed a mean of 5.46 fractal experiences (standard deviation = 1.48) per school focusing on salient local issues, with a mean of 29 staff participants (standard deviation = 15.26) per experience. Across schools, each treatment school completed three to eight fractal improvement experiences with a range of 7–115 staff participants.

Of the 142 total fractal experiences completed across 26 schools, 39 experiences related specifically to reading (27.46 percent), and 26 related specifically to mathematics (18.31 percent). The other 77 fractal experiences (54.23 percent) represented broader areas related to student achievement such as teacher professional development, school culture, data-based decisionmaking, student behavior and engagement, and parent involvement. Of the 26 treatment schools, 10 focused 50 percent or more of their fractal improvement experiences on reading exclusively or mathematics exclusively with the majority focused on reading, 10 focused 50 percent or more of their fractal experiences on both reading and mathematics, and 6 focused 50 percent or more of their fractal experiences on multiple areas not directly targeting reading or mathematics, such as student behavior, school culture, parent involvement, and teacher professional development.

Both treatment and control schools had leadership teams and participated in other education initiatives as part of their school improvement process during the two-year period. Control schools implemented "business as usual" as their participation in the study did not require that they conduct specific or formal school improvement initiatives. In treatment schools, Success in Sight is meant to support rather than supplant other school improvement initiatives. Through fractal improvement experiences, leadership teams can focus on implementing, evaluating, and improving other initiatives, such as those involving curriculum and assessment. Based on interview feedback from 155 school representatives, 7 treatment schools and 8 control schools spent comparable amounts of time participating in professional learning communities and Success in Sight. Three or fewer treatment schools and 6 control schools spent comparable amounts of time participating in leadership academies and Success in Sight. Of the 28 Missouri schools participating in the study, 8 treatment schools and 3 control schools received professional development services from the Regional Professional Development Centers. All treatment and control schools in Missouri implemented Reading First and response to intervention during the study period. In Minnesota, all 24 treatment and control schools

implemented the Mondo literacy program and the Phonological Awareness Literacy Screening assessment. It is important to note that control schools' "business as usual" condition included school improvement professional development opportunities similar to components of the Success in Sight program, although no control schools implemented a systemic school improvement program similar to Success in Sight. This study did not seek to measure or describe school improvement practices related to reading and mathematics outside of the Success in Sight intervention in either treatment or control schools. Information regarding the nature and extent of these other education initiatives was limited to interviews with a sample of three participants from each school at the end of the study period.

Impact of Success in Sight on student achievement

This study's results revealed that Success in Sight did not have a statistically significant impact on student achievement in reading or mathematics after two years. Researchers conducted sensitivity analyses to test the robustness of the benchmark impact estimates to the use of a baseline achievement covariate, the student sample, and methods used to estimate impacts across the two states in the study sample. The sensitivity analyses with no covariate, as well as the sensitivity analyses that estimated impacts separately by state and combined results meta-analytically, supported the benchmark findings of no statistically significant effect of Success in Sight on student achievement in reading or mathematics. Sensitivity analyses conducted with a smaller sample of students who remained in the same school throughout the study period (student stayers) revealed that those students who stayed in schools participating in Success in Sight averaged posttest mathematics scores statistically significantly lower than those of students from control schools.

These findings indicated that the statistical significance of the benchmark impact estimate of Success in Sight on reading was not sensitive to the inclusion or exclusion of a cluster-level baseline covariate, the two impact analysis methods used, or the student sample. Findings also indicated that the statistical significance of the benchmark impact estimate of Success in Sight on student achievement in mathematics was not sensitive to the use of a cluster-level baseline covariate or the two impact analysis methods used, but it was sensitive to the student benchmark and stayers samples. It would have been useful to test the sensitivity of findings to the inclusion of stayers and within-study in-movers, but researchers did not have baseline enrollment data and could not identify students who were grades 1 and 2 at baseline and remained in the same study school over the study period.

Impact of Success in Sight on teacher capacity for school improvement practices

Results from this study indicated that Success in Sight did not have a statistically significant impact on teacher capacity for data-based decisionmaking, purposeful community, or shared leadership after two years. Researchers conducted sensitivity analyses to test the robustness of the benchmark impact estimates to the use of a baseline covariate for teacher capacity for school improvement practices. The results from these sensitivity analyses were consistent with the results from the benchmark analyses indicating that Success in Sight did not have a statistically significant impact on teacher capacity for school improvement practices in data-based decisionmaking, purposeful community, or shared leadership. This indicates that the benchmark

impact estimate was not sensitive to the inclusion or exclusion of a cluster-level baseline covariate.

Relationship between intermediate outcomes and primary outcomes

Exploratory analyses revealed a statistically significant negative relationship between the following intermediate teacher capacity outcomes and the primary student achievement outcomes: teacher capacity for shared leadership and student reading achievement, teacher capacity for data-based decisionmaking and student mathematics achievement, and teacher capacity for shared leadership and student mathematics achievement. Therefore, higher teacher capacity scores were associated with lower student achievement scores in these areas. There was a statistically significant positive relationship between teacher capacity for purposeful community and student mathematics achievement, indicating that higher teacher capacity scores for purposeful community were associated with higher student achievement in mathematics.

It is unclear why the exploratory analyses revealed statistically significant negative relationships between some of the intermediate outcomes and primary outcomes and why they did not find statistically significant positive relationships between all of the intermediate outcomes and primary outcomes. Regarding the findings that were not statistically significant, it is possible that this study did not find statistically significant positive relationships between all the teacher capacity intermediate outcomes and student achievement outcomes because these relationships do not exist as hypothesized or because they exist but the study did not measure either the intermediate outcomes or student achievement outcomes properly.

Study implications and limitations

Although educators have implemented Success in Sight over the past 11 years, there has been no systematic evaluation of its effectiveness until this study. This cluster randomized trial used rigorous methodology to yield objective evidence of Success in Sight's impact on student achievement in reading and math as well as on teacher capacity for school improvement practices. The study was adequately powered to detect an effect size of 0.20 for the primary outcomes of student achievement and an effect size of 0.30 for the secondary outcomes of teacher capacity for school improvement practices. Although this study incorporated rigorous methodology and was adequately powered, there are limitations to consider when interpreting these study findings.

The study's external validity is limited because of the specific sample selection criteria and characteristics of schools that volunteered to participate. Participating schools were located in Minnesota and Missouri. Thus, the study's findings do not generalize to schools located in other states. In addition, the study schools were specifically selected because they were low- to moderate-performing schools, defined as not having made adequate yearly progress (AYP) for any of the three years prior to the study or being at risk of not making AYP in the current or prior year. Therefore, the study's findings do not generalize to schools in Minnesota or Missouri that made AYP for the three years prior to the study and were not at risk for not making AYP in the current or prior year. Another limitation of this study is that schools' AYP status during the three years prior to the study was unequally distributed across treatment and control schools. Although the analytic models each included a cluster-level pretest covariate corresponding to the outcome

of interest to account for baseline differences across treatment and control groups, they did not account for differences in AYP status. Furthermore, baseline (2008) comparisons revealed statistically significant group differences between study sample schools and the larger population of Minnesota and Missouri elementary schools not making AYP in any of the three years prior to the study. Specifically, comparisons revealed that Minnesota study schools were statistically significantly different from the larger population of Minnesota elementary schools not meeting AYP regarding reading and mathematics achievement, student eligibility for free or reduced-price lunch, students per teacher, ethnicity, Title I status, and school urbanicity. For Missouri, comparisons revealed that the study sample schools were statistically significantly different from the larger population for Missouri elementary schools not meeting AYP regarding grade 4 mathematics achievement, number of students per teacher, and school size. Therefore, the study's findings are not generalizable to the larger population of low- to moderate-performing elementary schools in Minnesota or Missouri, defined as those not having made AYP for any of the three years prior to the study. Because of the voluntary nature of study participation, it is unclear whether the study's findings would generalize to schools that declined the opportunity to participate or to schools that systematically differ from those that chose to participate.

The study's external validity also is limited because of the specific student achievement content areas and teacher capacities assessed, the type of assessments used, and the population of student participants for the primary achievement outcomes. Because this study assessed only student reading and mathematics achievement in grades 3–5 using the Minnesota and Missouri state assessments administered to students in grades 3–5, the results are not generalizable to other achievement content areas, student achievement as measured by other assessments, or students in grades other than 3–5. Likewise, the findings related to Success in Sight's impact on teacher capacity for data-based decisionmaking, purposeful community, and shared leadership cannot be applied to Success in Sight's impact on other areas of teacher capacity for school improvement practices.

This study also has limitations related to the self-report nature of the implementation data and teacher survey data. Specifically, the implementation data were self-report data collected from Success in Sight facilitators. Success in Sight facilitators documented the nature and frequency of schools' fractal improvement experiences as part of their routine practice, but because of budget limitations, researchers were unable to confirm these data by conducting independent and objective fidelity measures. Although this study did not explore relationships among implementation fidelity and primary or secondary outcomes, the reliance upon only self-report data to document implementation fidelity is a limitation of this study. In addition, the data collected pertaining to this study's secondary outcomes, teacher capacity for data-based decisionmaking, purposeful community practices, and shared leadership, consisted solely of self-report data. Therefore, readers should use caution when interpreting findings regarding the study's secondary outcomes.

Results from the exploratory analyses, which revealed that none of the teacher outcomes were positively associated with reading and mathematics student achievement outcomes and that some of the teacher outcomes were statistically significantly negatively related to student achievement outcomes, suggest that there may be additional limitations to this study. One limitation is that the exploratory analyses were not based on the experimental design of the study and are subject to

selection bias, which might have contributed to the findings. Furthermore, it is possible that the teacher survey did not measure the constructs of data-based decisionmaking, purposeful community, and shared leadership as intended. The coefficient alphas for the three subscales were well above acceptable levels, but additional analyses suggested that the data-based decisionmaking and purposeful community subscales lacked sufficient reliability and validity. Related to Success in Sight's underlying theory, it is possible that the teacher survey assessed intended teacher outcomes reliably but that there is not a relationship between teacher capacity in the areas assessed and the student achievement outcomes assessed. It is also possible that these factors contributed to the finding of no statistically significant positive relationship between the teacher and student outcomes assessed in this study.

Other limitations relate to implementation of Success in Sight and variations in fractal improvement experiences. Regarding implementation, the study findings do not generalize to schools that implement Success in Sight for more than two years. As cited previously, it can take two to four years of implementing an improvement initiative before detecting statistically significant student impacts (Fullan 2007). The study findings also do not generalize to schools that do not participate in the consortium approach, which brings clusters of schools in the same geographic area together to participate in the large-group professional development sessions. The study did not examine the relationship between student or teacher outcomes and variations in fractal improvement experiences, including content focus area (reading only, math only, reading and math only, or other focus area) and magnitude (number of fractal improvement experiences completed and number of staff participants involved in fractal improvement experiences within each school). Therefore, it is unknown whether the focus area and magnitude of fractal improvement experiences had a positive or negative relationship with student or teacher outcomes. It also is unknown whether schools whose fractal improvement experiences focused on reading or math instructional changes had different student achievement outcomes from schools whose fractal improvement experiences focused on other areas unrelated specifically to reading or math.

Appendix A. Regional Educational Laboratory Central firewall procedures

The school improvement intervention investigated, Success in Sight, was developed by McREL. As such, assessing its effectiveness posed a potential conflict of interest. To mitigate this threat to the integrity of the study, McREL hired two external research organizations, Magnolia Consulting and ASPEN Associates, to conduct the research study, as approved by the Institute of Education Services (IES) and U.S. Department of Education (ED). ASPEN Associates was responsible for study design, recruitment, management, and data collection during study Year 1. Magnolia Consulting was responsible for data collection during Year 2, analysis, and reporting. With input from the IES, McREL built a "firewall" between the researchers conducting the study and the McREL staff implementing the intervention. The firewall consisted of a set of policies, structures, and procedures that functioned analogously to a network system firewall. The firewall limited communication between external researchers and Success in Sight mentors and access to data to maintain security of the information collected, for the purpose of providing unbiased answers to the research questions.

The dual purpose of the firewall was to ensure that McREL (a) did not intentionally or unintentionally obtain feedback or data from the external research firms, Magnolia Consulting and ASPEN Associates, that could have been used after implementation of the intervention (i.e., inform mid-course corrections), and (b) did not inform Magnolia Consulting's interpretation of study results in a manner that may have resulted in a biased presentation of the findings.

McREL established a policy on the structures and procedures needed to construct a firewall that separated the research and intervention components of this field-based study. This policy aligned with IES conflict of interest policies. McREL established structures to keep separate the research and intervention components, including a subcontract with Magnolia Consulting, to design, conduct, and manage the study of Success in Sight's effectiveness in changing school practices and raising student achievement. As stipulated in the external researchers' subcontract, and in accordance with IES requirements for Task 2 Rigorous Studies, external researchers randomly assigned schools, collected and analyzed data, and formulated interpretations of findings using its own facilities and organizational resources, which are independent of and geographically separate from McREL. Additionally, an IES-approved Technical Working Group (TWG) reviewed the research design, instruments, data analysis plan, and reporting strategies for this study, and advised Magnolia Consulting and ASPEN Associates on how to meet technical standards for cluster randomized trials. Finally, to facilitate and monitor communications as necessary between external researchers and the Success in Sight implementation at McREL, McREL assigned a staff member as a liaison between McREL and Magnolia Consulting, clarifying roles and responsibilities and scheduling data collection sessions.

Although ASPEN Associates and McREL worked together to recruit schools, in order to maintain objectivity and integrity of the study, the subcontractor conducted the random assignment of schools to treatment and control groups. McREL did not have access to the data collected for this study during the study period. McREL and external researchers avoided direct communication that did not include IES with respect to data quality, analyses, appropriateness of interpretations, or other technical issues that might have affected the outcome of the study.

Magnolia Consulting contracted with an external editor to conduct a substantive edit of the report. In addition, external technical advisors conducted a methodological review of the report.

Per the firewall procedure approved by IES, four McREL employees with editing and research expertise were authorized to review the report to ensure that it adhered to McREL's organizational quality standards. These employees were prohibited from disclosing or sharing any aspects of the study, including the results, with other McREL employees prior to the study's publication. Suggested edits focused on the flow and clarity of the report; no edits were made to the study findings. All suggested edits were tracked within the document and submitted concurrently by McREL to Magnolia Consulting and IES. Changes agreed to by Magnolia Consulting were accepted in the tracked changes document.

Changes in reporting not agreed to by Magnolia Consulting would have been noted as points of non-agreement requiring further discussion; if needed, these discussions would have been coordinated by IES. However, there were no non-agreement changes. If non-agreement would have occurred, McREL and Magnolia Consulting would have consulted with a panel comprised of one TWG member identified by McREL, one TWG member identified by Magnolia Consulting, the Analytical Technical Services monitor, and IES as an observer. Magnolia Consulting would have revised reports based on the panel's feedback, including preparation of a response to recommendations. McREL's project directors and study liaison monitored implementation of this set of policies, structures, and procedures and reported on and discussed their implementation with ED as part of the standard monitoring process and with the TWG at least annually.

Appendix B. Power analyses

Researchers conducted three power analyses for this study: one for the benchmark analyses of the impact of Success in Sight on student achievement using the entire student sample (students with baseline or posttest data available), one for the sensitivity analyses of the impact of Success in Sight on student achievement using only the student stayer sample (students who remained in the same school at baseline and posttest), and one for the benchmark analyses of the impact of Success in Sight on teacher capacity for school improvement practice outcomes.

For this study, researchers created blocks of school pairs by matching schools on 2006 reading achievement and student eligibility for free or reduced-price lunch before randomly assigning one school from each matched pair to a treatment or control group. In some cases, when blocking prior to random assignment has occurred, the block can be considered a "site," and the power analysis can be run as a multisite cluster randomized trial in Optimal Design (Liu et al. 2006). However, because the analytic model for this study did not treat blocks as sites, but rather used blocks to reduce variability and included them in the model as fixed effects, site was not considered a third level of the model. Therefore, researchers conducted this study's power analyses using Optimal Design software (Liu et al. 2006) for cluster randomized designs with treatment at level 2. In these power analyses, the effective school-level sample size reflects the number of matched pairs required to achieve .80 power to detect the specified standardized effect sizes. The discussion below provides rationales for the estimates for effect size, intraclass correlation, and the reduction in between-school variance by the matching variable and pretest covariate.

The final sample included 52 schools. This sample size supported the primary and secondary benchmark analyses, but it did not support the sensitivity analyses conducted only with the student stayer sample, and allowed for school-level attrition (table B1).

Table B1. Parameter estimates for power analyses

Analysis	Effect size	Intraclass correlation coefficient	R^2	Minimum power	Students or teachers per matched pair	Matched Pairs of Schools
Main effects on student achievement						
Benchmark sample	.20	.10	.75	.80	300	4225
Sensitivity stayer sample	.20	.10	.75	.80	100	29
Main effects on school improvement practices	.30	.10	.55	.80	6	26

Main effects on student achievement

The assumed minimum detectable effect size for the main effect of student achievement is 0.20 (see table B1), a conservative estimate based on the literature on the effects of whole-school reform on student achievement. No empirical evidence was available from field trials of the intervention itself. However, estimates of effect sizes were available from other studies of whole-school reform. These estimates vary according to the type of intervention and the outcome measure. In their meta-analysis, Borman et al. (2003) report that the average effects of comprehensive school reform on student achievement range from 0.09 for third-party studies

using comparison groups to 0.15 for all evaluations of the achievement effects. When using all available studies, the effects of four comprehensive school reform models most closely aligned with Mid-continent Research for Education and Learning and Success in Sight were 0.09 for Accelerated Schools, 0.13 for the Center for Effective Schools, 0.15 for the School Development Program, and 0.25 for Onward to Excellence (Borman et. al. 2003). Based on the documented size of the effect on student achievement for Onward to Excellence, the model most closely aligned with Success in Sight, a minimum detectable effect size of 0.20 is reasonable and reflects the effects of Success in Sight when implemented with fidelity over two years by the highly trained McREL mentors.

Researchers selected a value of 0.10 for the intraclass coefficient based on the following sources. Liu et al. (2006) cite typical intraclass coefficients for educational achievement to be between 0.05 and 0.15. Schochet (2005) states that intraclass coefficients for standardized test scores often range between 0.10 and 0.20. Schochet also found intraclass coefficients in grade 3 and 4 reading and math ranging from 0.06 to 0.08 (adjusted for district effects) across 71 Title I schools in 18 school districts engaged in whole-school reform.

Researchers selected prior achievement as a cluster-level covariate, and the proportion of postintervention variance explained by preintervention test scores of 0.50 was deemed an appropriately conservative estimate based on prior research. Schochet (2008b) concludes that the proportion of variance explained by baseline measures is at least 0.50 when student-level data are used. Bloom, Bos, and Lee (1999) found similar values. Bloom, Richburg-Hayes, and Black (2005) found values ranging from 0.33 to 0.81 across five districts for school-level baselines. Researchers conservatively estimated that creating matched pairs (based on 2006 reading achievement and student eligibility for free or reduced-price lunch) before random assignment would explain .25 of the variance in primary and secondary outcomes.

The sample sizes for the benchmark impact analyses on primary outcomes assume 300 students (that is, two classrooms of 25 students per classroom per grade across each of the two schools in the matched pair) to be nested within each school. The sample sizes for the sensitivity impact analyses on primary outcomes assume 100 students (that is, 2 classrooms of 25 students who were in grade 3 at baseline and grade 5 at posttest within each of the two schools in the matched pair) to be nested within each school.

Given the above assumptions and two-level cluster randomized trial, Optimal Design software (Liu et al. 2006) calculated that 50 schools (25 matched pairs) were necessary to achieve the desired power of 0.80 for the student achievement outcomes for the benchmark sample and 58 schools (29 matched pairs) were needed for the sensitivity stayer sample.

Main effects on teacher capacity for school improvement practices

The three outcomes related to teacher capacity for school improvement practices include data-based decisionmaking, purposeful community, and shared leadership. Researchers found little empirical evidence regarding estimates of effect size, intraclass correlation, and proportion of posttest variance explained by baseline measures of these school improvement practices. Consequently, researchers chose the parameter estimates for these analyses, shown in table B1, for the following reasons.

With regard to the minimal detectible effect size, it can be assumed that an effect exists, but the magnitude of that effect is unknown. Rigorous studies of comprehensive school reform to date have not been designed to examine changes in school practices over time. However, in one study of comprehensive school reform models and distributed (shared) leadership (Camburn, Rowan, and Taylor 2003), the authors "tentatively" claim that the comprehensive school reform programs they studied (Accelerated Schools, America's Choice, and Success for All) configure leadership in their schools differently than non–comprehensive school reform schools, but do not give direct estimates of those effects for individual programs. Given that Success in Sight is a systemic school improvement approach that deals directly with changing school practices, a conservative estimated effect size of 0.30 was deemed appropriate. Researchers found no evidence of intraclass correlation estimates for school improvement practices. Thus, researchers used a value of 0.10 as the estimate of the intraclass correlation based on the increased variability when schools from two states are part of the sample.

Estimates for post-intervention variance explained by preintervention measures of school improvement practices were conservatively set at 0.30 for three reasons. First, researchers assumed that school practices would vary because of differences in implementation, specifically, the manner and timelines by which the leadership teams at each school would "scale up" to involve the whole school in the change process. Second, the intervention itself anticipates variations in the choice of school improvement goals to be selected by participating schools as their focus for improvement. Third, the baseline scores on school practices were used as the covariate for school practices measures at the end of years 1 and 2, and the correlation between these measures were not directly known, but only hypothesized based on intercorrelations. A study of the Effective Schools comprehensive school reform model, which included 38 high schools, 32 middle schools, and 134 elementary schools across 22 school districts, reported high intercorrelations among school environment (that is, school practices) variables (Witte and Walsh 1990). This study examined four scales related to effective schools, teacher control or influence, and parent involvement. Intercorrelations for teacher control and teacher ratings of school effectiveness ranged from 0.52 at the elementary level to 0.82 at the middle school level.

Researchers used an assumption of 40 teachers per matched school pair (20 teachers per school) to estimate final sample size for the power analysis for reports of school practices with an effect size of 0.30, the proportion of postintervention variance explained by preintervention test scores of 0.30, and an intraclass correlation of 0.10. Using the above parameter estimates, Optimal Design calculated that 52 schools (26 matched pairs) were necessary to achieve a power greater than 0.80.

Appendix C. Response rates by time point, measure, and experimental group

Table C1. Response rates by time point, measure, and group, 2007/08 and 2009/10

	Total			Treatment			Control				
Measure	**Number of eligible participants**	**Number of actual participants**	**Response rate (percent)**	**Number of eligible participants**	**Number of actual participants**	**Response rate (percent)**	**Number of eligible participants**	**Number of actual participants**	**Response rate (percent)**	***p*-value**	**Effect size[a]**
Baseline											
Student reading assessments	8,609	8,467	98.35	4,705	4,665	99.15	3,904	3,802	97.39	<.01***	0.07
Student mathematics assessments	8,438	8,331	98.73	4,557	4,519	99.17	3,881	3,812	98.22	<.01***	0.04
School Improvement Practices Teacher Survey	1,574	1,374	87.29	819	750	91.58	755	624	82.65	<.01***	0.13
Principal interviews[b]	52	52	100.00	26	26	100.00	26	26	100.00	na	na
School focus groups[b]	52	52	100.00	26	26	100.00	26	26	100.00	na	na
Posttest											
Student reading assessments	8,340	8,182	98.11	4,473	4,403	98.44	3,867	3,779	97.72	.02**	0.03
Student mathematics assessments	8,329	8,213	98.61	4,468	4,413	98.77	3,861	3,800	98.42	.21	0.02
School Improvement Practices Teacher Survey	1562	1516	97.06	825	815	98.79	737	701	95.12	<.01***	0.11
Phone interviews	156	155	99.36	78	77	98.72	78	78	100.00	1.00	–0.08

**Significant at $p = .05$; ***significant at $p = .01$.

na is not applicable.

Note: Analyses conducted were 2 by 2 chi-square tests between the frequency of eligible and actual participants for treatment groups compared with control groups.

a. Effect sizes were calculated for chi-square square tests using the phi coefficient.

b. Chi-square tests not computed because 100 percent of participants completed the measure.

Source: Minnesota Department of Education 2008a, 2010b; Missouri Department of Elementary and Secondary Education 2008a, 2010b; principal interviews 2008; phone interviews 2008, 2010; school focus groups 2008; teacher surveys 2008, 2010.

Appendix D. Data collection instruments

Large-group professional development fidelity checklist (segment sample)

Session 1 Fidelity Checklist

Instructions: This document is to be completed the Intervention Team Members that participated in this segment. *Only one checklist needs to be completed per segment (i.e., have the team members complete it as a group at the conclusion of the large group session).*

Please note the extent to which you covered the segment as a whole and each component of this segment "as planned." Check the box that most closely reflects the coverage.

Your name(s)	**Date of first day of this PD session:**			
Area/consortium:				
Segment and Segment Components (in order of intended presentation)	**To what extent did you cover this segment and each component "as planned" (select one)?**			
	Covered all or almost all of it (80 percent or more)	**Covered part of it**		**Did not cover this at all AND did not intend to cover as needed during site visits**
Segment 1.1: Overview of Success in Sight				
Welcome and Introductions				
▪ Activity: Walk and Talk				
▪ Activity: School Success Stories				
▪ Activity: Our Work Together				
▪ *Segment 1.1 Learning Targets*				
Overview of Success in Sight				
▪ Goals of Success in Sight				

▪ *The Science of School Improvement*				
▪ *The Art of School Improvement*				
▪ *Activity: Why is Change So Hard?*				
▪ *Barriers to School Improvement*				
▪ *Overcoming Barriers to School Improvement*				

Site visit principal interview protocol 2008

INTRODUCTION TO INTERVIEW

- Good morning/afternoon. Thanks for taking the time today to talk with me about school improvement initiatives in your school.
- My name is _____ .
- I am assisting ASPEN Associates, a research organization, with the data collection for this study.
- The study is sponsored by the US Dept of Education and it examines school improvement initiatives and their impact on increasing student achievement. Your school is one of those participating in the study.
- This discussion is one of a series on school improvement. In each school we are talking to teachers, leadership teams and principals about their school improvement initiatives.
- We want to get your perspectives because you are on the front lines and working with students on a daily basis.
- Today, I have few questions that ask for your perceptions. Perceptions may vary and your experiences and thoughts may be different from others in your school. We don't expect that everyone will have the same views, and we encourage you to share your views, even if they differ from others' views.
- I will be recording the session because I don't want to miss any of your comments. No one else besides researchers will be listening to this tape recording and your responses to my questions will be kept confidential. By that I mean that your name will never be associated with any comment you make, nor will your answers be presented in a manner that a reader would be able to identify you.
- Okay, let's begin.

(Note: The text below is required on all data collection protocols per OMB and IES)

The U.S. Department of Education wants to protect the privacy of individuals who participate in data collection. Your answers will be combined with other respondents, and no one will know how you answered the questions. This data collection is authorized by law (1) Sections 171(b) and 173 of the Education Sciences Reform Act of 2002, Pub. L. 107-279 (2002); and (2) Section 9601 of the Elementary and Secondary Education Act (ESEA), as amended by the No Child Left Behind (NCLB) Act of 2001 (Pub. L. 107-110). Responses to this data collection will be used only for statistical purposes. The reports prepared for this study will summarize findings across the sample and will not associate responses with a

specific district or individual. We will not provide information that identifies you or your district to anyone outside the study team, except as required by law.

INTERVIEW QUESTIONS

Notes to Interviewer:

- Pay attention to whether the principal can readily respond to the questions (i.e., has a clear opinion) or he/she struggles to respond (i.e., does not have a clear opinion).
- In preparation for the summary at the end, listen to whether the principal believes there is shared understanding on both how to achieve (Q2) and can achieve (Q3), only one, or neither.

(2 mins.) General Introduction

(1 min.) Framing the Questions: As I mentioned, the purpose of this interview is to learn more about the nature of school improvement efforts currently underway at your school. By "***school improvement***" we mean everything your school is doing to improve teaching and learning.

Today, I would like you to focus on the ***school as a whole***, rather than on your individual role at the school.

(3 mins.) Vision for Success:

1. Does your school have a vision for success? What are some words you would use to describe success at your school?

   [Listen for then probe:

   - What does your school want to see change for students? (Note: The school goals may differ from the district goals, which most schools feel are reflected in the school improvement plan.)

(15 mins.) Working Together: Schools often talk about their vision for success in terms of their goals.

2. Do you feel there is a *shared or common understanding* among your staff ***about how*** your school will to achieve its goals or vision for success?

   What makes you say that? [Listen for and probe:

   - Do all staff believe this is the right strategy or approach to achieve the goals?

   - Do all staff feel there is a clearly articulated plan for moving forward?]

3. Do you feel there is a *shared or common belief* among the staff that your school ***can achieve*** its goals or vision for success?

   What makes you say that? [Listen for and probe:

   - Do all staff feel they have the resources, skills, and support they need to move forward to achieve the school's goals or vision for success?]

(10 mins.) Types of School Improvement Initiatives: Schools are engaged in many different initiatives, all of which are aimed at improving teaching and learning in some way.

4. When you consider *all* of the initiatives underway at your school this year, would you say that for the most part they were:

   - an extension of or building on what you've done in the past,
   - a real break with what's been done in the past, or…
   - a little of both?

   [Note: Do not force schools into one category. Some schools may have been in a holding pattern this year; that is, no change but more status quo.]

   4a. Can you tell me the initiatives you were thinking of in formulating your response?

(5 mins.) Summarize Themes

5. So, I'd like to summarize my understanding of what you have shared today.

   1. What I heard is that your school [has/doesn't have/you don't know if it has/mixed opinions] a shared or common understanding of ***how*** it will achieve its goals or vision for success.

   2. I also heard that your school [has/doesn't have/you don't know if it has/mixed opinions] a shared understanding that it ***can achieve*** its goals or vision for success.

   3. And, finally, that when considering all of the initiatives underway at your school this year, I heard that overall, you would **characterize these initiatives** as [an extension of the past/a break with the past/some of both/mixed opinions/none of the above].

   Have I adequately captured your perceptions?

(5 mins.) Final Question:

6. Before we end is there anything else you feel would be important for me to know – anything you feel may have helped or hindered your school's improvement efforts this year?

Thank you for your time today. If you have any other comments or any questions you'd like to share, I can give you the phone number of the Project Manager for the Study.

Focus group protocol 2008

INTRODUCTION TO 2008 FOCUS GROUP WITH KEY SCHOOL STAFF

- Good morning/afternoon. Thanks for taking the time today to join our discussion about school improvement initiatives in your school.
- My name is _____ and my partner is _____.
- We are assisting ASPEN Associates, a research organization, with the data collection for this study.
- The study is sponsored by the US Dept of Education and it examines school improvement initiatives and their impact on increasing student achievement. Your school is one of those participating in the study.
- This discussion is one of a series on school improvement. In each school we are talking to teachers, leadership teams and principals about their school improvement initiatives.
- We want to get your perspectives because you are on the front lines and working with students on a daily basis.
- Today, we have few questions that ask for your perceptions. Perceptions may vary and your experiences and thoughts may be different from others in the group. We don't expect that everyone will have the same views, and we encourage you to share your views, even if they differ from others' views.
- Before we get started, here are just a few ground rules:
    - If you have your cellular phone with you, please turn the volume off so that it will not disturb the group.
    - If you must leave the session for a meeting or appointment, we hope that you are able to return and continue in our discussion.
    - We will be recording the session because we don't want to miss any of your comments. No one else besides researchers will be listening to this tape recording and your responses to my questions will be kept confidential. By that I mean that your name will never be associated with any comment you make, nor will your answers be presented in a manner that a reader would be able to identify you.
    - We also want you to respect each others' confidentiality. In other words, what's said here, stays here.
    - Finally, we have five (5) questions to cover today, so I will keep us moving along.
    - Okay, let's begin.

(Note: The text below is required on all data collection protocols per OMB and IES)

The U.S. Department of Education wants to protect the privacy of individuals who participate in data collection. Your answers will be combined with other respondents, and no one will know how you answered the questions. This data collection is authorized by law (1) Sections 171(b) and 173 of the Education Sciences Reform Act of 2002, Pub. L. 107-279 (2002); and (2) Section 9601 of the Elementary and Secondary Education Act (ESEA), as amended by the No Child Left Behind (NCLB) Act of 2001 (Pub. L. 107-110). Responses to this data collection will be used only for statistical purposes. The reports prepared for this study will summarize findings across the sample and will not associate responses with a specific district or individual. We will not provide information that identifies you or your district to anyone outside the study team, except as required by law.

According to the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number. The valid OMB control number for this information collection is 1850-0838. The time required to complete this information collection is estimated to average 45 minutes per respondent, including the time to review instructions, gather the data needed, and complete and review the information collected. If you have any comments concerning the accuracy of the time estimate(s) or suggestions for improving this form, please write to: U.S. Department of Education, Washington, DC 20202. If you have comments or concerns regarding the status of your individual submission of this form, write directly to: U.S. Department of Education, Institute of Education Sciences, 555 New Jersey Avenue, NW, Washington, DC 20208.

FOCUS GROUP QUESTIONS 2008

Notes to Moderator and Assistant Moderators:

1. Pay attention to whether the group can readily respond to the questions (i.e., they have a clear opinion) or they struggle to respond (i.e., they do not have a clear opinion).
2. In preparation for the summary at the end, listen to whether they have shared understanding on both how to achieve (Q2) and can achieve (Q3), only one, or neither.
3. If you observe or hear disagreement, be sure to ask "What do others think?"

(2 mins.) General Introduction

(1 min.) Framing the Questions: As I mentioned, the purpose of this interview is to learn more about the nature of school improvement efforts currently underway at your school. By "***school improvement***" we mean everything your school is doing to improve teaching and learning.

Today, I would like you to focus on the ***school as a whole***, rather than on your individual role at the school. And, it is especially important for this study that we hear about ***different points of view.*** So, please feel free to share your views even if they differ from what others have said. You don't have to address all your comments to me. Feel free to follow-up on what someone else has said.

(3 mins.) Vision for Success:

1. Does your school have a vision for success? What are some words you would use to describe success at your school?

[Listen for then probe:

- What does your school want to see change for students? (Note: The school goals may differ from the district goals, which most schools feel are reflected in the school improvement plan.)

(15 mins.) Working Together: Schools often talk about their vision for success in terms of their goals.

2. Do you feel there is a *shared or common understanding* among your staff ***about how*** your school will to achieve its goals or vision for success?

What makes you say that? [Listen for and probe:

- Do all staff believe this is the right strategy or approach to achieve the goals?
- Do all staff feel there is a clearly articulated plan for moving forward?]

3. Do you feel there is a *shared or common belief* among the staff that your school ***can achieve*** its goals or vision for success?

What makes you say that? [Listen for and probe:

- Do all staff feel they have the resources, skills, and support they need to move forward to achieve the school's goals or vision for success?]

(10 mins.) Types of School Improvement Initiatives: Schools are engaged in many different initiatives, all of which are aimed at improving teaching and learning in some way.

4. When you consider *all* of the initiatives underway at your school this year, would you say that for the most part they were:

- an extension of or building on what you've done in the past,
- a real break with what's been done in the past, or…
- a little of both?

[Note: Do not force schools into one category. Some schools may have been in a holding pattern this year; that is, no change but more status quo.]

4a. Can you tell me the initiatives you were thinking of in formulating your response?

(5 mins.) Summarize Themes

5. So, I'd like to summarize my understanding of what the group has shared today.

4. What I heard is that your school [has/doesn't have/you don't know if it has/mixed opinions] a shared or common understanding of ***how*** it will achieve its goals or vision for success.

5. I also heard that your school [has/doesn't have/you don't know if it has/mixed opinions] a shared understanding that it ***can achieve*** its goals or vision for success.

6. And, finally, that when considering all of the initiatives underway at your school this year, I heard that overall, you would **characterize these initiatives** as [an extension of the past/a break with the past/some of both/mixed opinions/none of the above].

Have I adequately captured the perceptions of this group?

(5 mins.) Final Question:

6. Before we end is there anything else you feel would be important for me to know – anything you feel may have helped or hindered your school's improvement efforts this year?

Thank you for your time today. If you have any other comments *or any questions* you'd like to share, see me afterwards and I can give you the phone number of the Project Manager for the Study.

Spring 2010 interview protocols

INTRODUCTION TO SPRING 2010 INTERVIEWS

- Good morning/afternoon. Thanks for taking the time today to talk with me about school improvement initiatives in your school.
- My name is _____.
- I am assisting Magnolia Consulting, a research organization, with the data collection for this study of Success in Sight.
- The study, which is sponsored by the US Dept of Education, examines school improvement initiatives and their impact on student achievement. Your school is one of those participating in the study.
- In each school we are talking to principals, a member of the school leadership team, and a staff member about their perceptions of what has helped or hindered your school's improvement efforts in the last two years.
- We expect that perceptions may vary and your experiences and thoughts may be different from others in your school. We don't expect that everyone will have the same views, and we encourage you to share yours, even if they differ from others'.
- I will be recording the session because I don't want to miss any of your comments. No one else besides researchers will be listening to this tape recording and your responses to my questions will be kept confidential. By that I mean that your name will never be associated with any comment you make, nor will your answers be presented in a manner that a reader would be able to identify you.
- Okay, let's begin.

(Note: The text below is required on all data collection protocols per OMB and IES)

The U.S. Department of Education wants to protect the privacy of individuals who participate in data collection. Your answers will be combined with other respondents, and no one will know how you answered the questions. This data collection is authorized by law (1) Sections 171(b) and 173 of the Education Sciences Reform Act of 2002, Pub. L. 107-279 (2002); and (2) Section 9601 of the Elementary and Secondary Education Act (ESEA), as amended by the No Child Left Behind (NCLB) Act of 2001 (Pub. L. 107-110). Responses to this data collection will be used only for statistical purposes. The reports prepared for this study will summarize findings across the sample and will not associate responses with a specific district or individual. We will not provide information that identifies you or your district to anyone outside the study team, except as required by law.

According to the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number. The valid OMB control number for this information collection is 1850-0838. The time required to complete this information collection is estimated to average 45 minutes per respondent, including the time to review instructions, gather the data needed, and complete and review the information collected. If you have any comments concerning the accuracy of the time estimate(s) or suggestions for improving this form, please write to: U.S. Department of Education, Washington, DC 20202. If you have comments or concerns regarding the status of your individual submission of this form, write directly to: U.S. Department of Education, Institute of Education Sciences, 555 New Jersey Avenue, NW, Washington, DC 20208.

Spring 2010 Principal Interview Questions

Steps:

1. Mail questions to principal
2. Principal faxes back to Magnolia prior to interview
3. Interviewer reviews transcript of baseline principal interview for background prior to interview (see folder 13 in transfer file)
4. Interviewer conduct follow-up interview to ask questions about any information that needs clarifying

School Name: _____ State: _____

Principal Name: _____ Telephone Number: _____

AYP STATUS:

Q1. What was your school's recent AYP status? (please circle one response for each year, subject, and student group)

YEAR	ALL STUDENTS?		SUBGROUPS? (ethnic, special education, etc.)
BASELINE 2007-2008 READING	Made AYP	Did not make AYP	At least one (1) subgroup did not make AYP
BASELINE 2007-2008 MATH	Made AYP	Did not make AYP	At least one (1) subgroup did not make AYP

YEAR ONE **2008-2009** **READING**	Made AYP	Did not make AYP	At least one (1) subgroup did not make AYP
YEAR ONE **2008-2009** **MATH**	Made AYP	Did not make AYP	At least one (1) subgroup did not make AYP
YEAR TWO **2009-2010** **READING**	Made AYP	Did not make AYP	At least one (1) subgroup did not make AYP
YEAR TWO **2009-2010** **MATH**	Made AYP	Did not make AYP	At least one (1) subgroup did not make AYP

(Interviewer: Confirm all AYP status reported)

SCHOOL IMPROVEMENT INITIATIVES:

Q2. In the last two years (2007-08 and 2008-09), has your school participated in any major school improvement initiatives? (please check all that apply and add others)

Systemic Reform Initiatives	**Yes, our school participated**
1. McREL's Success in Sight	
2. Center for Effective Schools	
3. Comer's	
4. Accelerated Schools	
5. Onward to Excellence	
6. Early Reading First	

7. Reading First	
8. Other (please specify)	
9. Other (please specify)	
10. Other (please specify)	
Supplemental Initiatives	**Yes, our school participated**
1. Missouri's RPDCs	
2. Principal's Leadership Academy	
3. Professional Learning Communities (PLC) (not included in Success in Sight)	
4. Other (please specify)	
5. Other (please specify)	
6. Other (please specify)	
7. Other (please specify)	
8. Other (please specify)	
9. Other (please specify)	
10. Other (please specify)	

(Interviewer: Probe for others and clarify their nature, i.e., reading focus, etc.)

CHANGES IN STATE AND LOCAL EDUCATION POLICIES

Q3a. In the last two years (2007-08 and 2008-09), what if any changes in state and local education policies and practices occurred that you feel had an effect on your school's improvement efforts?

Some examples are:

school start times

grade level configurations

other school reorganization

curriculum

instruction

assessment

Q3b. When you consider these changes, would you say that for the most part they were (circle one response)?

1. an extension of or building on what you've done in the past,
2. a real break with what's been done in the past
3. a little of both

Q3c. Which changes in particular were you thinking of in formulating your response?

Interviewer:

- **Discuss changes**
- **Identify whether they occurred in Year 1 or Year 2**
- **Whether helped or hindered school improvement**
- **Whether mostly first- or second-order change (Q3b)**

OTHER BARRIERS & SUPPORTS TO SCHOOL IMPROVEMENT

Q4a. In the last two years (2007-08 and 2008-09),what else has changed at your school that you feel has had an effect on your school's improvement efforts?

Some examples are:

- **Changing student demographics**
- **Changing student enrollment**
- **Changes to school facilities (e.g., air conditioning)**
- **Other changes specific to budget cuts (e.g., staffing, materials)**

Q4b. When you consider these changes, would you say that for the most part they were (circle one response)?

1. an extension of or building on what you've done in the past,
2. a real break with what's been done in the past
3. a little of both

Q4c. Which changes in particular were you thinking of in formulating your response?

Interviewer:

- **Discuss changes**
- **Identify whether they occurred in Year 1 or Year 2**
- **Whether helped or hindered school improvement**
- **Whether mostly first- or second-order change (Q3b)**

SPRING 2010 LEADERSHIP TEAM & STAFF MEMBER

INTERVIEW QUESTIONS

Steps:

1. Conduct interview with principal first

2. Interviewer reviews Observation Notes from baseline leadership team and staff focus groups for background prior to interview (see folder 13 in transfer file)

3. Then conduct interviews with a member of the leadership team (LT) and a member of the school staff (ST)

School Name: _____ State: _____

Staff Name: _____ Telephone Number: _____

Interview: LT ST

CHANGES IN STATE AND LOCAL EDUCATION POLICIES

Q3a. In the last two years(2007-08 and 2008-09), what if any changes in state and local education policies and practices occurred that you feel had an effect on your school's improvement efforts?

Some examples are:

school start times

grade level configurations

other school reorganization

curriculum

instruction

assessment

Q3b. When you consider these changes, would you say that for the most part they were (circle one response)?

4. an extension of or building on what you've done in the past,
5. a real break with what's been done in the past
6. a little of both

Q3c. Which changes in particular were you thinking of in formulating your response?

Interviewer:

- **Discuss changes**
- **Identify whether they occurred in Year 1 or Year 2**
- **Whether helped or hindered school improvement**
- **Whether mostly first- or second-order change (Q3b)**

OTHER BARRIERS & SUPPORTS TO SCHOOL IMPROVEMENT

Q4a. In the last two years(2007-08 and 2008-09),what else has changed at your school that you feel has had an effect on your school's improvement efforts?

Some examples are:

- **Changing student demographics**
- **Changing student enrollment**
- **Changes to school facilities (e.g., air conditioning)**
- **Other changes specific to budget cuts (e.g., staffing, materials)**

Q4b. When you consider these changes, would you say that for the most part they were (circle one response)?

7. an extension of or building on what you've done in the past,
8. a real break with what's been done in the past
9. a little of both

Q4c. Which changes in particular were you thinking of in formulating your response?

Interviewer:

- **Discuss changes**
- **Identify whether they occurred in Year 1 or Year 2**
- **Whether helped or hindered school improvement**
- **Whether mostly first- or second-order change (Q4b)**

Teacher school improvement online survey

Welcome to the School Improvement Study! Your school is involved in a study of *Success in Sight*, a school improvement intervention. The research portion of this study is being conducted by Magnolia Consulting.

This survey asks about the educational practices engaged in at your school. Please select the answer that most closely represents your views. The survey has several sections and will take approximately 25-30 minutes to complete. (Note: You will not be able to exit the survey and return at a later time, so please plan to complete the survey in one sitting.)

Staff members who complete the survey will receive a $25 check. After you complete the survey, you will be directed to another page (separate from your survey responses) where you will be asked to provide the information needed to process your check. This information will not be attached to your survey responses. (Please note: Both the survey and the address page are hosted on a secure web server.)

Responses to this survey will only be used for statistical purposes. The reports prepared for this study will summarize findings across schools and will not associate responses with a specific district, school or individual. We will not provide information that identifies you, your school or district to anyone outside the research team, except as required by law.

Thank you for your participation!

The U.S. Department of Education wants to protect the confidentiality of individuals who participate in surveys. We want to assure you that the results will never be presented in a way that will permit any responses to be associated with any individual, and only the researchers will have access to the data. This survey is authorized by law (1) Sections 171(b) and 173 of the Education Sciences Reform Act of 2002, Pub. L. 107-279 (2002); and (2) Section 9601 of the Elementary and Secondary Education Act (ESEA), as amended by the No Child Left Behind (NCLB) Act of 2001 (Pub. L. 107-110). Responses to this data collection will be used only for statistical purposes. The reports prepared for this study will summarize findings across the sample and will not associate responses with a specific district or individual. We will not provide information that identifies you or your district to anyone outside the study team, except as required by law.

According to the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number. The valid OMB control number for this information collection is 1850-0838. The time required to complete this information collection is estimated to average 25 minutes per respondent, including the time to review instructions, gather the data needed, and complete and review the information collected. If you have any comments concerning the accuracy of the time estimate(s) or suggestions for improving this form, please write to: U.S. Department of Education, Washington, DC 20202. If you have comments or concerns regarding the status of your individual submission of this form, write directly to: U.S. Department of Education, Institute of Education Sciences, 555 New Jersey Avenue, NW, Washington, DC 20208.

1) In which area is your school located?

2) At which school do you work? (If you work at more than one school, please select the school you work at the most. If your time is split equally between schools, please randomly select one school.)

The answers to the questions in this survey should reflect your experiences at the following school:

If your school is correct, please click "Next Page." If your school is incorrect, please click "Previous Page" below to change your response.

If your school is correct, please click "Next Page." If your school is incorrect, please click "Previous Page" below to change your response.

3) What is your position in this school?
❍ Classroom teacher
❍ Specialist teacher (ELL, Spec. Ed., Art, Music, Science, etc.)
❍ Educational or teaching assistant
❍ Office Staff
❍ Social Work, Psychologist
❍ Other (please specify)

If you selected other, please specify
_____

4) What percentage time is your position at this school?
❍ Less than .25 FTE
❍ .25 to .49 FTE
❍ .50 to .75 FTE
❍ More than .75 FTE

School Environment

This section relates to your school's environment. Please answer the questions based on your observations or opinions. If you feel that you are not in a position that enables you to answer a question, just leave it blank.

5) To what extent do you agree or disagree with the following statements about your school?

	Strongly Agree	Somewhat Agree	Neither Agree nor Disagree	Somewhat Disagree	Strongly Disagree
a. The primary mission of my school is that all students become proficient in core subjects.	❍	❍	❍	❍	❍
b. My school sets ambitious goals for student achievement.	❍	❍	❍	❍	❍
c. My school has an explicit statement of high expectations concerning student achievement.	❍	❍	❍	❍	❍
d. My school supports all teachers in their efforts to improve student achievement.	❍	❍	❍	❍	❍

6) To what extent do you agree or disagree with the following statements about your school?

	Strongly Agree	Somewhat Agree	Neither Agree nor Disagree	Somewhat Disagree	Strongly Disagree
a. Year-to-year changes in student achievement are monitored at the student level.	❍	❍	❍	❍	❍
b. School-level progress towards academic proficiency is communicated to all teachers at my school.	❍	❍	❍	❍	❍
c. Teachers in my school are provided with opportunities to collaboratively use assessment results to discuss student progress.	❍	❍	❍	❍	❍

7) To what extent do you agree or disagree with the following statements about your school?

	Strongly Agree	Somewhat Agree	Neither Agree nor Disagree	Somewhat Disagree	Strongly Disagree
a. Our staff values school improvement.	❍	❍	❍	❍	❍
b. All teachers in my school believe that students can reach standards and objectives.	❍	❍	❍	❍	❍
c. Our teachers assume responsibility for ensuring that all students learn.	❍	❍	❍	❍	❍
d. Teachers in my school emphasize that student performance can always be improved.	❍	❍	❍	❍	❍

8) To what extent do you agree or disagree with the following statements about your school?

	Strongly Agree	Somewhat Agree	Neither Agree nor Disagree	Somewhat Disagree	Strongly Disagree
a. My school has a specific parent involvement initiative that encourages parents to participate in decisions about school policies.	❍	❍	❍	❍	❍
b. School staff and teachers are open to suggestions from parents.	❍	❍	❍	❍	❍
c. My school pays specific attention to parents who are hard to reach.	❍	❍	❍	❍	❍

9) To what extent do you agree or disagree with the following statements about your school?

	Strongly Agree	Somewhat Agree	Neither Agree nor Disagree	Somewhat Disagree	Strongly Disagree
a. My school views strong parental support as an important condition for school effectiveness.	❍	❍	❍	❍	❍
b. Teachers frequently talk with parents/families about the best conditions to support student's learning at home.	❍	❍	❍	❍	❍
c. Teachers and staff are readily accessible to parents.	❍	❍	❍	❍	❍
d. Parents are offered various options for involvement (e.g., tutoring their children at home, helping in the classrooms, joining school council, etc.)	❍	❍	❍	❍	❍

10) To what extent do you agree or disagree with the following statements about your school?

	Strongly Agree	Somewhat Agree	Neither Agree nor Disagree	Somewhat Disagree	Strongly Disagree
a. There is a safe, orderly learning environment at my school.	❍	❍	❍	❍	❍
b. Rules are well understood by staff and students.	❍	❍	❍	❍	❍
c. Staff members uniformly apply sanctions to students who defy school policies.	❍	❍	❍	❍	❍
d. There are positive and open interactions between staff and students.	❍	❍	❍	❍	❍

11) To what extent do you agree or disagree with the following statements about your school?

	Strongly Agree	Somewhat Agree	Neither Agree nor Disagree	Somewhat Disagree	Strongly Disagree
a. Students in my school are acknowledged and rewarded for good behavior.	❍	❍	❍	❍	❍
b. Teachers work hard to create a safe, orderly climate in their classrooms.	❍	❍	❍	❍	❍
c. My school administrators strive to create a safe, orderly learning environment.	❍	❍	❍	❍	❍

Professional Community and Community Support

This section relates to the support available to staff at your school. Please answer the questions based on your observations or opinions. If you feel that you are not in a position that enables you to answer a question, just leave it blank.

12) To what extent do you agree or disagree with the following statements about your school?

	Strongly Agree	Somewhat Agree	Neither Agree nor Disagree	Somewhat Disagree	Strongly Disagree
a. Most teachers and staff members feel comfortable voicing their concerns in this school.	❍	❍	❍	❍	❍
b. Teachers and other staff members are recognized for a job well done.	❍	❍	❍	❍	❍
c. There is a great deal of cooperative effort among staff at this school.	❍	❍	❍	❍	❍

13) To what extent do you agree or disagree with the following statements about your school?

	Strongly Agree	Somewhat Agree	Neither Agree nor Disagree	Somewhat Disagree	Strongly Disagree
a. Teachers share responsibility for all students' learning at this school.	❍	❍	❍	❍	❍
b. Teachers at this school are continually learning.	❍	❍	❍	❍	❍
c. Teachers are involved in making important educational decisions at this school.	❍	❍	❍	❍	❍
d. Teachers have influence on the content/focus of professional development at this school.	❍	❍	❍	❍	❍
e. There is a formal support system for beginning teachers.	❍	❍	❍	❍	❍

14) To what extent do you agree or disagree with the following statements about your school?

	Strongly Agree	Somewhat Agree	Neither Agree nor Disagree	Somewhat Disagree	Strongly Disagree
a. Teachers at this school are able to get through to difficult students.	❍	❍	❍	❍	❍
b. Teachers here are confident they will be able to motivate their students.	❍	❍	❍	❍	❍
c. Teachers at this school really believe every child can learn.	❍	❍	❍	❍	❍
d. If a child doesn't want to learn, teachers at this school give up.	❍	❍	❍	❍	❍
e. Teachers at this school don't have the skills needed to produce meaningful student learning.	❍	❍	❍	❍	❍
f. Teachers in this school do not have the skills to deal with student disciplinary problems.	❍	❍	❍	❍	❍

15) To what extent do you agree or disagree with the following statements about your school?

	Strongly Agree	Somewhat Agree	Neither Agree nor Disagree	Somewhat Disagree	Strongly Disagree
a. Students at this school come ready to learn.	❍	❍	❍	❍	❍
b. Home life provides so many advantages the students at this school are bound to learn.	❍	❍	❍	❍	❍
c. Students at this school just aren't motivated to learn.	❍	❍	❍	❍	❍
d. The opportunities in this community help ensure that students at this school will learn.	❍	❍	❍	❍	❍

e. Learning is more difficult at this school because students are worried about their safety.	❍	❍	❍	❍	❍
f. Drug and alcohol abuse in the community make learning difficult for students at this school.	❍	❍	❍	❍	❍

Mission, Goals and School Improvement Efforts

The following section is about the mission, goals and the school improvement efforts at your school. Please answer the questions based on your observations or opinions. If you feel that you are not in a position that enables you to answer a question, just leave it blank.

16) To what extent do you agree or disagree with the following statements about your school?

	Strongly Agree	**Somewhat Agree**	**Neither Agree nor Disagree**	**Somewhat Disagree**	**Strongly Disagree**
a. Administrators, teachers, and parents share a common vision of school improvement.	❍	❍	❍	❍	❍
b. Teachers share the principal's beliefs and values about what the central mission of this school should be.	❍	❍	❍	❍	❍
c. In my school, we have a shared purpose about our work.	❍	❍	❍	❍	❍
d. Teachers are aware of what the leadership believes regarding teaching and learning.	❍	❍	❍	❍	❍

17) To what extent do you agree or disagree with the following statements about your school?

	Strongly Agree	Somewhat Agree	Neither Agree nor Disagree	Somewhat Disagree	Strongly Disagree
a. Specific goals for student achievement have been established for the students in my school.	❍	❍	❍	❍	❍
b. Our school-wide goals are understood by all teachers.	❍	❍	❍	❍	❍
c. Our school-wide goals are a prominent part of our day-to-day lives.	❍	❍	❍	❍	❍
d. The school mission provides a clear sense of direction for teachers.	❍	❍	❍	❍	❍

18) To what extent do you agree or disagree with the following statements about your school?

	Strongly Agree	Somewhat Agree	Neither Agree nor Disagree	Somewhat Disagree	Strongly Disagree
a. Leaders support risk-taking and innovation in teaching.	❍	❍	❍	❍	❍
b. Teachers in the school are continually learning and seeking new ideas.	❍	❍	❍	❍	❍
c. The principal is interested in innovation and new ideas.	❍	❍	❍	❍	❍
d. In my school, we systematically consider new and better ways of doing things.	❍	❍	❍	❍	❍
e. The principal is comfortable making changes in how things are done.	❍	❍	❍	❍	❍

19) To what extent do you agree or disagree with the following statements about your school?

	Strongly Agree	Somewhat Agree	Neither Agree nor Disagree	Somewhat Disagree	Strongly Disagree
a. Unless we make or continue to make changes in my school, student achievement is not going to improve.	❍	❍	❍	❍	❍
b. The school's efforts to improve have good results in the education students receive.	❍	❍	❍	❍	❍
c. My school's most pressing improvement needs are addressed in a timely manner.	❍	❍	❍	❍	❍
d. At my school, resources are prioritized in the budget to support improvement efforts.	❍	❍	❍	❍	❍
e. Improvement initiatives are specifically focused on student-related outcomes or goals.	❍	❍	❍	❍	❍

20) Do you work directly with students in an instructional capacity (includes classroom teachers, education assistants and specialists, such as Special Ed, ELL/ESL, Title I, Art, Music, Physical Education, etc.)? (*Please choose one*)

❍ Yes
❍ No

Your Teaching and Your Students

The next section relates to the activities of staff who work directly with students in an instructional capacity. If you feel that you are not in a position to answer a question, just leave it blank.

21) To what extent do you agree or disagree with the following statements about your teaching?

	Strongly Agree	Somewhat Agree	Neither Agree nor Disagree	Somewhat Disagree	Strongly Disagree
a. I frequently evaluate whether individual students are sufficiently progressing.	❍	❍	❍	❍	❍
b. I use academic materials specific to individual student skill levels.	❍	❍	❍	❍	❍
c. I make adjustments in my teaching based on student capabilities.	❍	❍	❍	❍	❍
d. I provide sustained assistance to individual students.	❍	❍	❍	❍	❍
e. I tutor or use others as tutors to meet individual learning needs.	❍	❍	❍	❍	❍
f. I seek information from others about my students' strengths and weaknesses.	❍	❍	❍	❍	❍
g. I make modifications in my teaching to improve students' success.	❍	❍	❍	❍	❍
h. I team up with parents to motivate my students.	❍	❍	❍	❍	❍
i. I frequently use time outside the classroom to help students learn.	❍	❍	❍	❍	❍

22) To what extent do you agree or disagree with the following statements about your school?

	Strongly Agree	Somewhat Agree	Neither Agree nor Disagree	Somewhat Disagree	Strongly Disagree
a. I frequently use various assessment data (e.g., end-of-chapter tests, homework, standardized tests, state tests, etc.) to adjust my teaching practices.	❍	❍	❍	❍	❍
b. I frequently give students individual feedback on their progress.	❍	❍	❍	❍	❍
c. I evaluate and return students' work at least once a week.	❍	❍	❍	❍	❍
d. I have access to my students' standardized test scores.	❍	❍	❍	❍	❍
e. I frequently use assessment results to monitor students' progress toward being proficient on academic standards.	❍	❍	❍	❍	❍

23) In the classroom, to what extent do your students…?

	Great Extent	Considerable Extent	Some Extent	Very Limited Extent	Not at All
a. Know their learning goals.	❍	❍	❍	❍	❍
b. Work on learning goals until they are achieved.	❍	❍	❍	❍	❍
c. Apply their knowledge to a variety of situations.	❍	❍	❍	❍	❍
d. Follow guidance (such as guidance on how to estimate, self-monitor, prepare a speech, etc.)	❍	❍	❍	❍	❍
e. Independently manage their classwork.	❍	❍	❍	❍	❍

f. Focus their discussions on lesson objectives.	❍	❍	❍	❍	❍
g. Receive written or verbal feedback on their progress.	❍	❍	❍	❍	❍
h. Receive tangible rewards for effort and persistence.	❍	❍	❍	❍	❍

Instructional Guidance and Professional Development

The next section relates to instructional guidance and professional development. Please answer the questions based on your observations or opinions.

24) To what extent do you agree or disagree with the following statements about your school?

	Strongly Agree	**Somewhat Agree**	**Neither Agree nor Disagree**	**Somewhat Disagree**	**Strongly Disagree**
a. The principal is directly involved in helping teachers design curricular activities for their classes.	❍	❍	❍	❍	❍
b. In my school, the principal provides guidance for the teachers in knowing what effective classroom practice is.	❍	❍	❍	❍	❍
c. The principal continually monitors the effectiveness of the instructional practices used in our school.	❍	❍	❍	❍	❍
d. Leaders in our school facilitate teachers working together.	❍	❍	❍	❍	❍

25) To what extent do you agree or disagree with the following statements about your school?

	Strongly Agree	Somewhat Agree	Neither Agree nor Disagree	Somewhat Disagree	Strongly Disagree
a. In my school, the instructional time of teachers is well-protected.	❍	❍	❍	❍	❍
b. In my school, the principal has been successful at ensuring that teachers have the necessary resources and professional opportunities to support high-quality instruction.	❍	❍	❍	❍	❍
c. Our principal believes it is important that teachers cover all of the materials in the prescribed curriculum.	❍	❍	❍	❍	❍
d. Our principal is well-prepared to assist teachers in the implementation of instruction that supports our content standards.	❍	❍	❍	❍	❍

26) To what extent do your state-, district-, or school-sponsored professional development activities during the past school year have the following characteristics? (Please do not include college or university courses)

	Great Extent	Considerable Extent	Some Extent	Very Limited Extent	Not at all	Not Applicable
a. The content was specific to the teaching of state or district academic content standards.	❍	❍	❍	❍	❍	❍
b. Addressed your knowledge and skills to help diverse learners.	❍	❍	❍	❍	❍	❍
c. Deepened your knowledge in a content area.	❍	❍	❍	❍	❍	❍

d. Provided adequate time for reflection on how to improve your teaching.	❍	❍	❍	❍	❍	❍
e. Occurred in professional development sessions that were connected and built on one another.	❍	❍	❍	❍	❍	❍
f. Were directly applicable to classroom practices.	❍	❍	❍	❍	❍	❍
g. Analyzed samples of student work.	❍	❍	❍	❍	❍	❍
h. Addressed student test results.	❍	❍	❍	❍	❍	❍

Planning Time and Teacher Collaboration

The next section relates to teacher collaboration and its effect on your teaching. Please answer the questions based on your opinions and observations.

27) During teachers' contracted time in school, how many hours per week do teachers have for planning?
❍ None
❍ Less than 1 hour
❍ 1-2 hours
❍ 2-4 hours
❍ 4 or more hours

28) During teachers' contracted time in school, how many hours per week do teachers have for common planning (i.e., time for two or more teachers to plan together)?
❍ None
❍ Less than 1 hour
❍ 1-2 hours
❍ 2-4 hours
❍ 4 or more hours

29) To what extent did the following activities during the past school year improve your teaching? If you did not engage in an activity, please check *Not Applicable*.

	Great Extent	Considerable Extent	Some Extent	Very Limited Extent	Not at all	Not Applicable
a. Meeting with other teachers on lesson planning or other collaborative work related to instruction.	❍	❍	❍	❍	❍	❍
b. Discussing with other teachers how to help specific students.	❍	❍	❍	❍	❍	❍
c. Working with others (e.g., principal, other teachers) to analyze and address student test results.	❍	❍	❍	❍	❍	❍
d. Working with others (e.g., principal, other teachers) to develop curriculum that is aligned with state standards.	❍	❍	❍	❍	❍	❍

30) To what extent did the following activities during the past school year improve your teaching? If you did not engage in an activity, please check *Not Applicable*.

	Great Extent	Considerable Extent	Some Extent	Very Limited Extent	Not at all	Not Applicable
a. Having other teachers observe your classroom teaching and provide feedback.	❍	❍	❍	❍	❍	❍
b. Reviewing feedback about your teaching with the principal or other administrator.	❍	❍	❍	❍	❍	❍
c. Engaging in mentoring with another teacher.	❍	❍	❍	❍	❍	❍

d. Working with a mathematics or language arts curriculum specialist.	❍	❍	❍	❍	❍	❍

Background and Experience

This section asks about your background and experience.

31) If your school is a treatment school, were you a leadership team member?
❍ Yes
❍ No
❍ N/A

32) What is your highest earned degree?
❍ Bachelor's Degree (BA, BS)
❍ Education Specialist's Degree
❍ Master's Degree (MA, MS)
❍ Doctorate (PhD, EdD)
❍ Other (please specify)

If you selected other, please specify

_____

33) Which of the following teacher certifications do you currently hold for the state in which you are teaching? *Please select one.*
❍ Provisional or Initial
❍ Professional
❍ Substitute
❍ Associate or Limited (highest degree held is Associate's degree)
❍ Conditional (hold Bachelor's and working towards teacher certification)
❍ Transitional or Temporary (hold valid out-of-state license)
❍ Professional-Technical (industry experience but do not need teaching license)
❍ Emergency
❍ Other (please specify)

If you selected other, please specify

_____

34) Please describe your primary role in this school.

❍ Regular classroom teacher
❍ Special education teacher
❍ Title I teacher
❍ Specialist (e.g., art, music, science)
❍ Other (please specify)

If you selected other, please specify

_____

35) Which grade level(s) do you currently teach? Please select all that apply.

❑ Pre-kindergarten
❑ Kindergarten
❑ 1st grade
❑ 2nd grade
❑ 3rd grade
❑ 4th grade
❑ 5th grade
❑ 6th grade
❑ All of the above

Appendix E. Rationale for cross-state data aggregation and *z*-score approach

Success in Sight is a systemic intervention designed to address schools' specific needs while building their capacities to plan, implement, and evaluate school improvement practices. A key aspect of the program is data-based decisionmaking. Collecting, analyzing, interpreting, and using state achievement data, in addition to other indicators of student and school performance, help inform decisions and establish and monitor school improvement goals. The intervention is ultimately intended to drive improvement in student performance on the state accountability tests. As a result, the key outcome measure in this study is school-level performance on state-administered achievement tests.

Using state-administered achievement tests as an outcome measure has a unique set of advantages and drawbacks (May et al. 2009). The main advantage of using state assessments is the fact that nearly every student is tested at state expense and grade- or school-level data are publicly available; these features serve to limit the cost of conducting a large-scale experimental study such as this one. The main disadvantage stems from concerns of comparability. Use of state-administered achievement tests can create complications when attempting to analyze and compare outcomes across grades, subjects, and states. To facilitate such comparisons in this analysis we followed the guidance of May et al. (2009), who prepared a recent IES report on this topic. Rather than comparing scale scores across states, we transformed all achievement data into *z*-scores. As part of our sensitivity analysis, we analyzed outcomes within states, calculated effect sizes, and then conducted a meta-analysis across states.

May et al. (2009) stress that researchers should address certain assumptions when combining impact estimates across grades and states using rescaled individual-level scores (that is, *z*-scores). These assumptions include consistency in the content assessed by state tests, homogeneity of the study sample across grades and states in representing the intervention's targeted sample, and similar underlying distributions of each state's test scores with the exception of differences in scale score means and standard deviations.

Cross-state content assessment comparisons

Based on the recommendation of May et al. (2009), researchers established criteria for identifying differences between the content of state assessments. Specifically, researchers defined the criteria for substantial differences in tested content between the two state assessments as follows: a set of items per any content strand represents greater than 40 percent of all items in the assessment, and between states a difference greater than 10 percent in the proportion of items per any strand. If both criteria were met, the set of items was considered a potentially inappropriate set of items on which to combine results across states. The content review indicated that the tests were comparable in the subject matter domains and in the format, length, mode, and timing of administration (see appendix F). The state assessments in both reading and mathematics demonstrate a broad sampling of content, and thus, the total scores in each domain reflect comparable measures of student achievement.

Cross-state sample characteristics

There were statistically significant differences between state study samples in 2008 student reading achievement for grades 3–5 as well as across all grades, with Minnesota sample schools performing lower than Missouri sample schools on their respective reading and mathematics state assessments, on average (table E1). There were statistically significant differences between state study samples in 2008 in the proportion of White, Hispanic, and Asian students, with the Minnesota sample having fewer White students and more Hispanic and Asian students than the Missouri sample (table E2). The Minnesota sample also had a statistically significantly higher percentage of students qualifying for free or reduced lunch and a higher number of Title I schools (see table E2). The Minnesota schools resided in one district in a city locale, which was statistically significantly different from the distribution of Missouri schools across city, suburb, town, and rural locales (see table E2).

Table E1. Baseline comparison of Minnesota and Missouri study sample schools on baseline scores and school demographics, 2007/08

	Minnesota study sample schools (n = 24)		Missouri study sample schools (n = 28)				
Characteristic	**Mean**	**Standard deviation**	**Mean**	**Standard deviation**	**Difference**	**Test statistic**	***p*-value**
Mean z-scores 2008 reading achievement[a]							
Grade 3	–0.67	0.99	–0.18	1.06	0.49	–4.48	<.01***
Grade 4	–0.65	1.14	–0.14	1.09	0.51	–4.77	<.01***
Grade 5	–0.64	0.94	–0.15	1.10	0.49	–5.19	<.01***
Total	–0.65	0.99	–0.16	1.14	0.49	–5.29	<.01***
Mean z-scores 2008 math achievement[a]							
Grade 3	–0.69	0.99	–0.16	1.01	0.53	–4.69	<.01***
Grade 4	–0.68	1.06	–0.13	1.03	0.55	–0.55	<.01***
Grade 5	–0.62	1.02	–0.15	1.12	0.47	–4.01	<.01***
Total	–0.66	1.00	–0.14	1.05	0.52	–5.04	<.01***
Number of students per school[b]	413.29	117.87	373.96	135.32	39.33	1.11	.27
Number of students per teacher[b]	14.15	1.76	14.95	2.86	–0.80	–1.23	.23
Students eligible for free or reduced lunch (percent)[b]	80.33	15.97	61.62	24.55	18.71	3.30	<.01***
Student population (percent)[b]							
White	17.11	14.26	60.09	38.87	–42.98	–5.44	<.01***
Black	35.70	20.54	31.28	41.76	4.42	0.50	.62
Hispanic	13.97	9.68	5.94	11.08	8.03	2.76	<.01***
Asian	30.58	17.92	1.42	2.16	21.17	7.92	<.01***
American Indian	2.64	6.86	1.28	1.76	1.36	1.01	.32

***Significant at p = .01.

a. Test statistics and p-values accounted for clustering of teachers within schools.

b. Test statistics and p-values were from t-tests between group means. Components may not sum to 100 percent because of rounding.

Source: Minnesota Department of Education 2008a; Missouri Department of Elementary and Secondary Education 2008a; U.S. Department of Education, National Center for Education Statistics 2008.

Table E2. Baseline comparison of Minnesota and Missouri study sample schools on school characteristics, 2007/08

Characteristic	Minnesota study sample schools (*n* = 24)		Missouri study sample schools (*n* = 28)		Test statistic	*p*-value
	Number	Percent	Number	Percent		
Schools receiving Title I funding (percent)[a]						
Title I–eligible school	22	91.67	20	71.42	–2.23	.14
Schoolwide Title I	21	87.50	15	53.57	–5.48	.02**
School urbanicity (percent)[a]						
City	24	100.00	10	35.71	–23.60	<.01***
Suburb	0	0.00	11	39.29		
Town	0	0.00	7[a]	25.00		
Rural	0	0.00				

**Significant at $p = .05$; ***significant at $p = .01$.
Note: Test statistics and *p*-values were from chi-square tests between percentages.
a. All categories were analyzed separately, but for the Missouri study sample schools the categories of town and rural were collapsed to preserve anonymity.
Source: U.S. Department of Education, National Center for Education Statistics 2008.

Cross-state sample distributions

To explore the degree to which the state samples exhibit similar underlying distributions by grade and content area, researchers created histograms using vertically scaled scores (May et al. 2009). Histograms depicted 2008 reading and math assessment *z*-scores across grades and states (figures E1–E4) and vertically scaled scores disaggregated by grade and subject area (figures E5–E16). Because the Minnesota and Missouri state assessments are scored using different scales, there are between-state differences regarding their ranges, means, and standard deviations.

Because these are normal distributions, they are unimodal and the means, medians, and modes fall in the middle of the distributions. Therefore, they exhibit zero skewness, and any score below the mean falls in the lower 50 percent of the distribution of scores, and any score above the mean falls in the upper 50 percent of the distribution of scores. The peakedness of the distributions reflected in the histograms show leptokurtic curves with more scores in the center of the distribution. One exception is the distribution for grade 3 reading scores in Minnesota, which reveals a mesokurtic curve consistent with a normal distribution. The shapes of these distributions indicate that the distributions are sufficiently similar across states to warrant aggregation.

Figure E1. Minnesota Comprehensive Assessment II reading scores grades 3–5, 2007/08

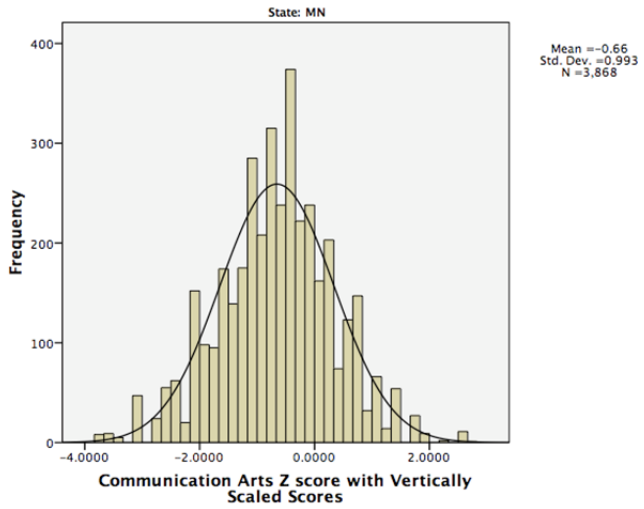


Figure E2. Missouri Assessment Program reading scores grades 3–5, 2007/08

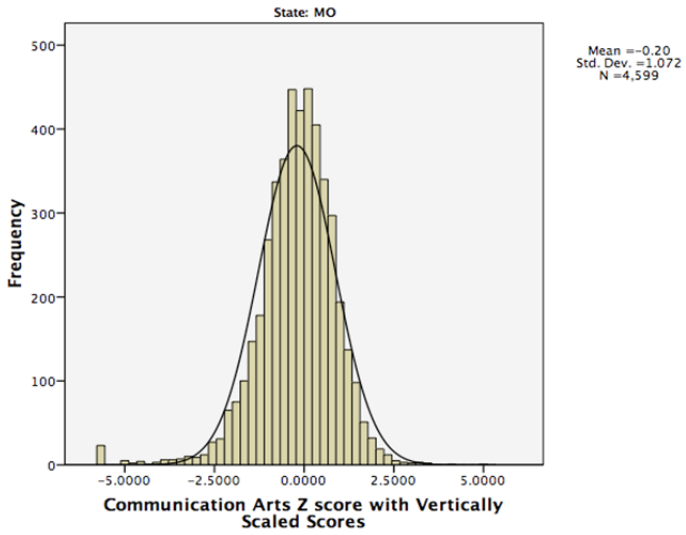


Figure E3. Minnesota Comprehensive Assessment II math scores grades 3–5, 2007/08

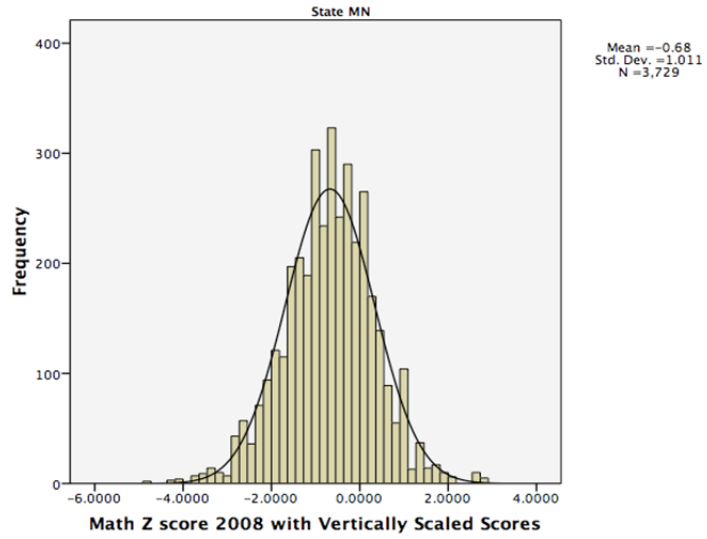


Figure E4. Missouri Assessment Program math scores grades 3–5, 2007/08

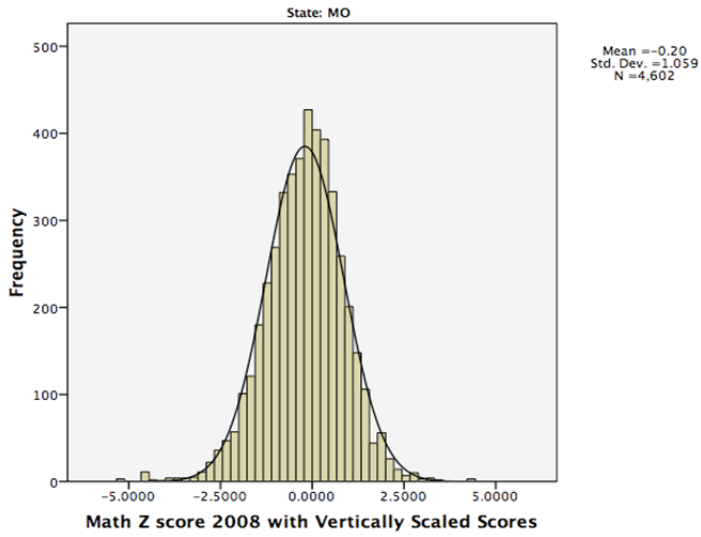


Figure E5. Grade 3 Minnesota Comprehensive Assessment II reading scores, 2007/08

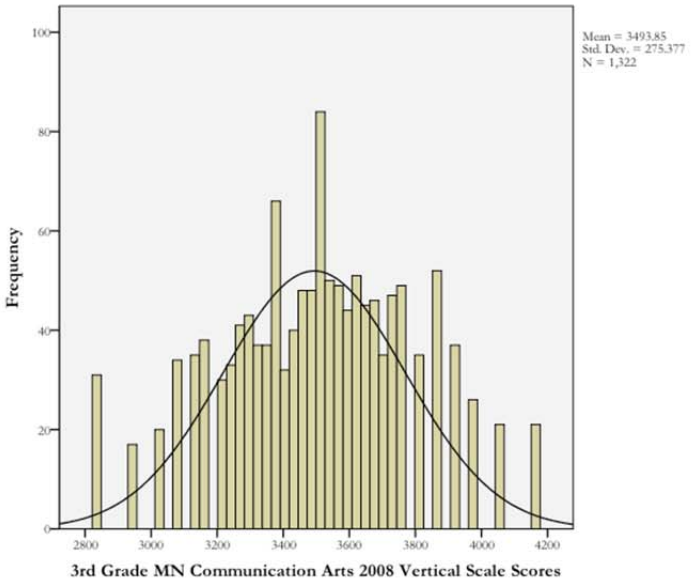


Figure E6. Grade 3 Missouri Assessment Program reading scores, 2007/08

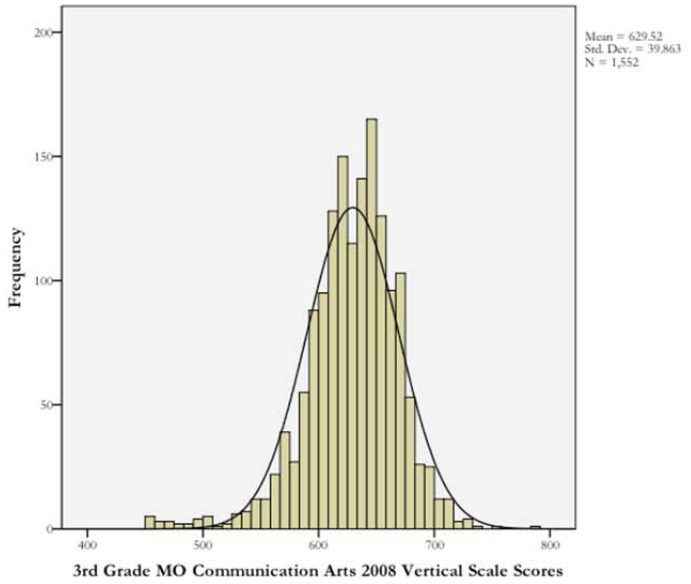


Figure E7. Grade 4 Minnesota Comprehensive Assessment II reading scores, 2007/08

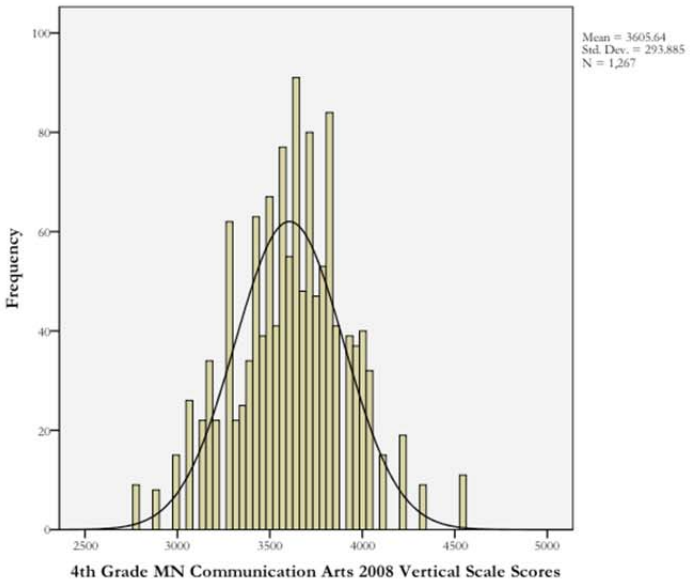


Figure E8. Grade 4 Missouri Assessment Program reading scores, 2007/08

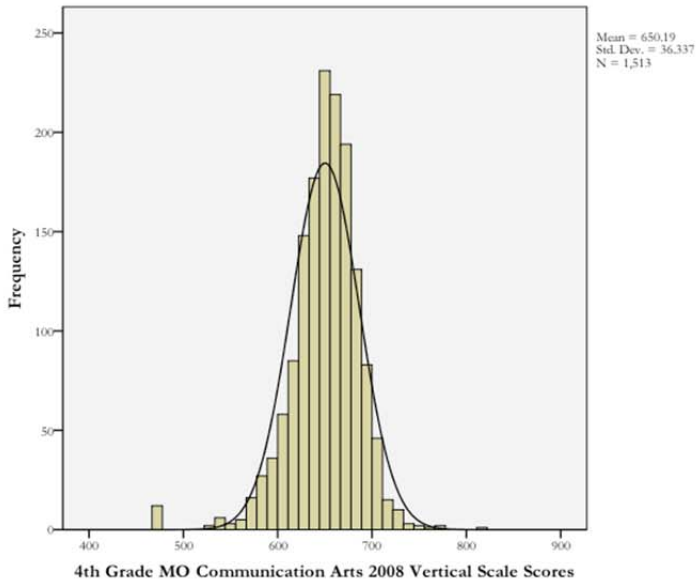


Figure E9. Grade 5 Minnesota Comprehensive Assessment II reading scores, 2007/08

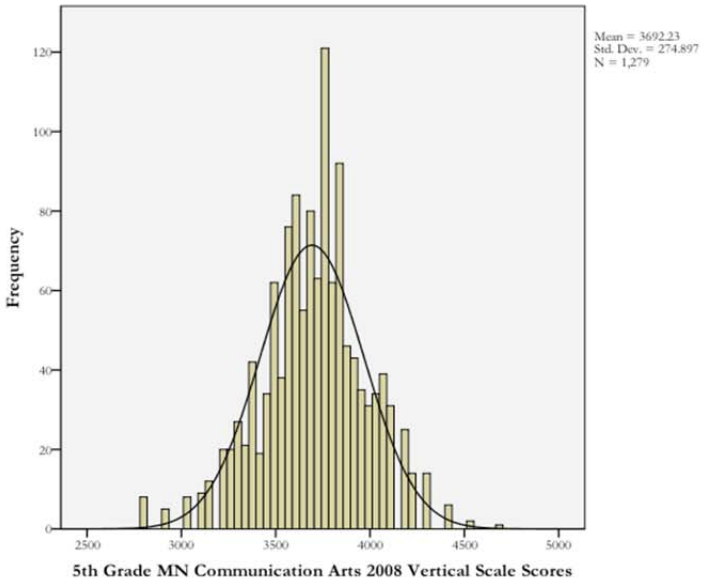


Figure E10. Grade 5 Missouri Assessment Program reading scores, 2007/08

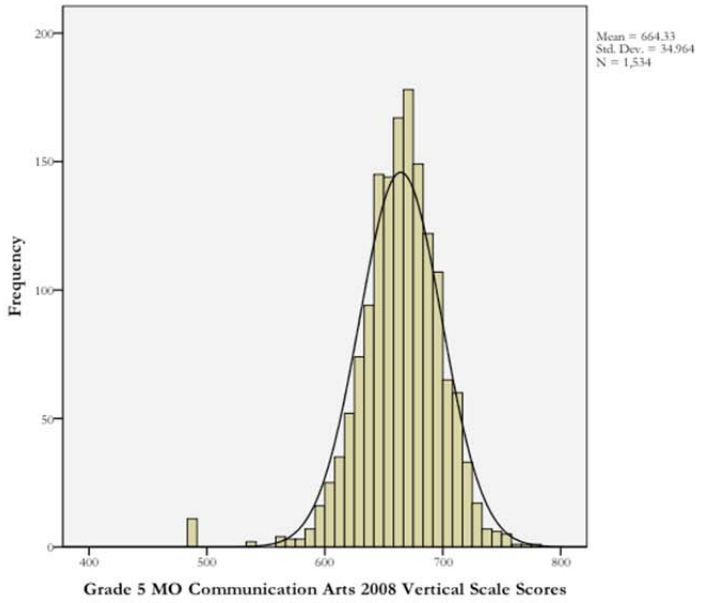


Figure E11. Grade 3 Minnesota Comprehensive Assessment II math scores, 2007/08

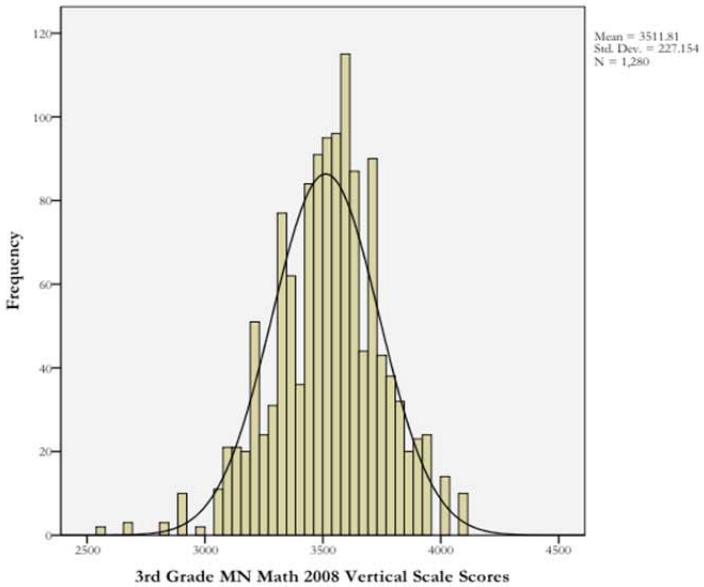


Figure E12. Grade 3 MAP math scores, 2007/08

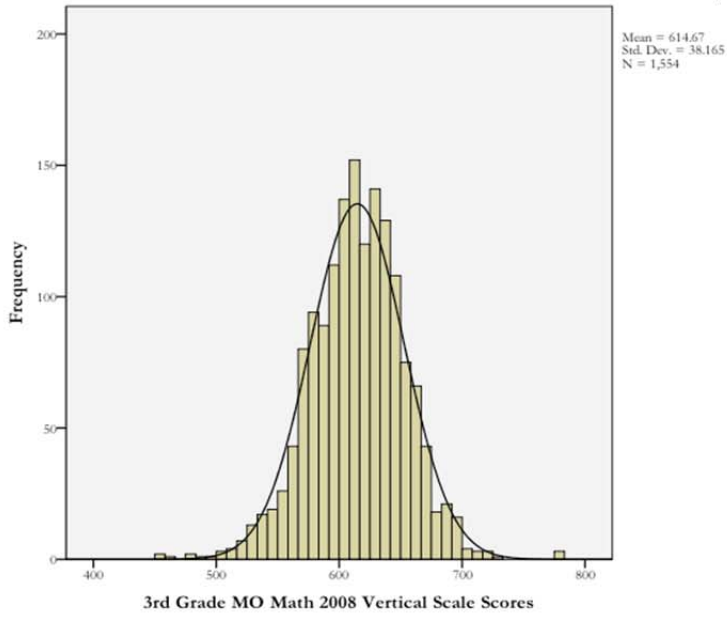


Figure E13. Grade 4 Minnesota Comprehensive Assessment II math scores, 2007/08

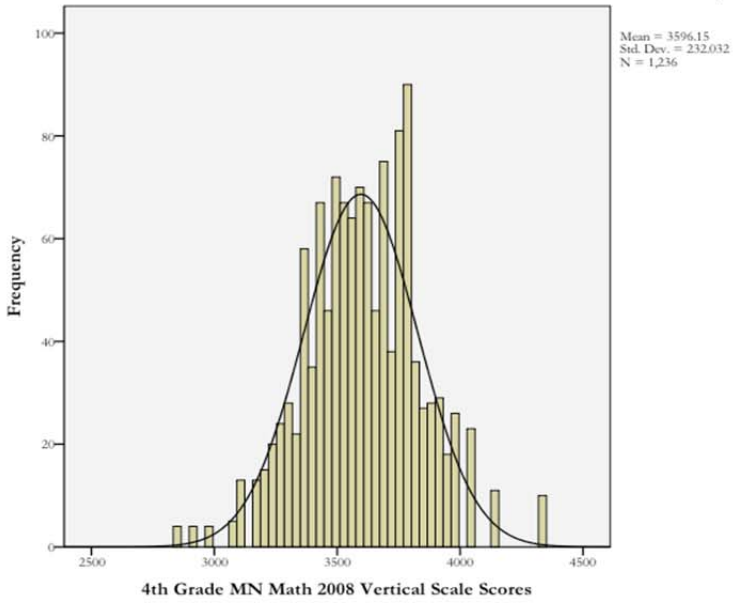


Figure E14. Grade 4 Missouri Assessment Program math scores, 2007/08

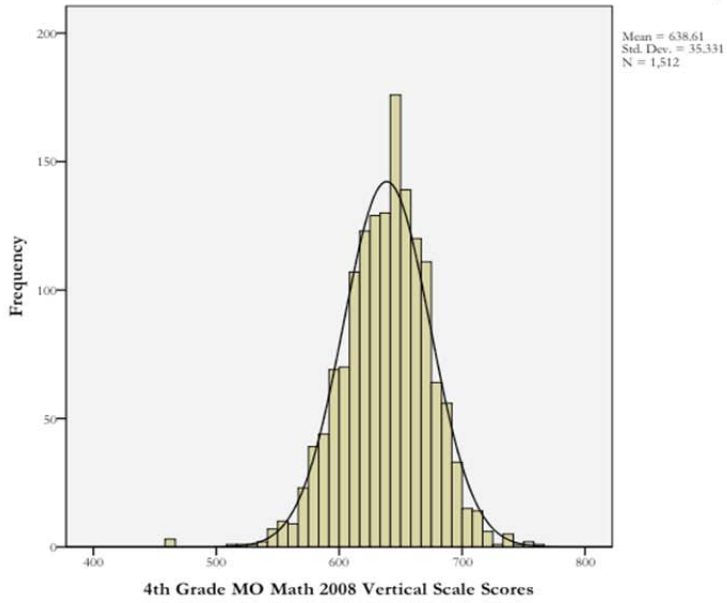


Figure E15. Grade 5 Minnesota Comprehensive Assessment II math scores, 2007/08

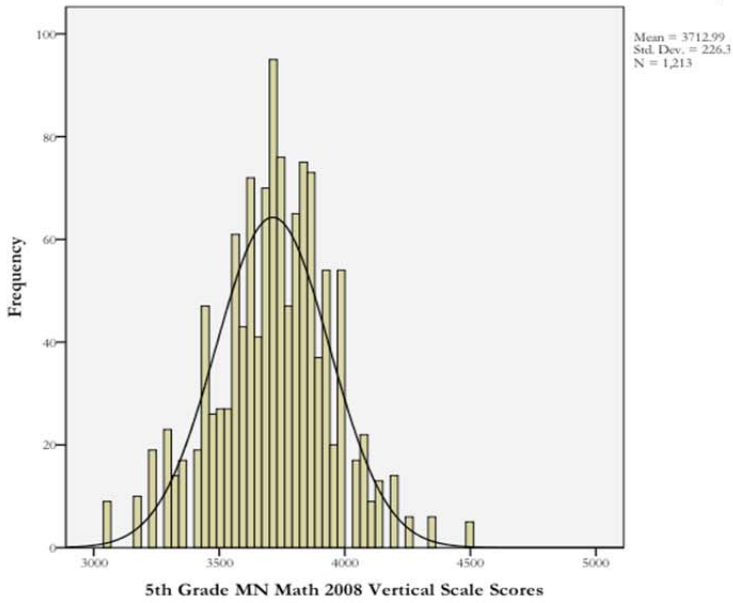
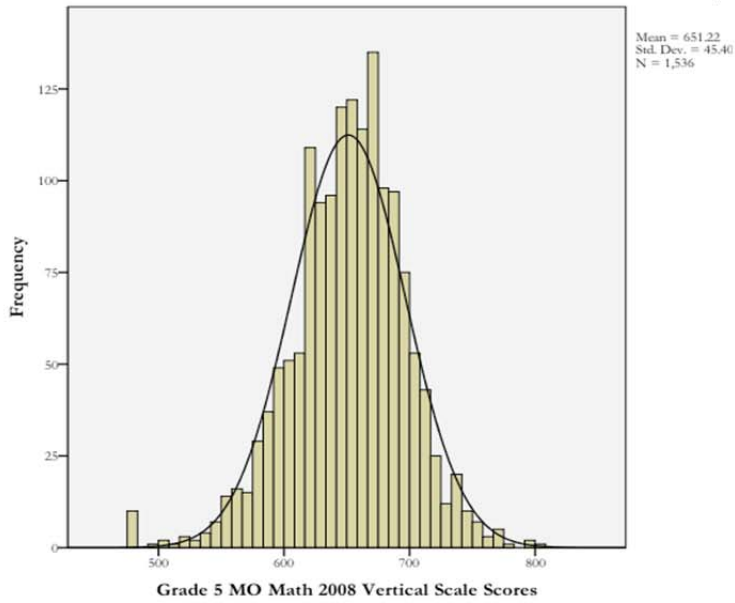


Figure E16. Grade 5 Missouri Assessment Program math scores, 2007/08



Appendix F. Content review of state assessments

This appendix compares the reliability and validity data as well as the content knowledge and skills of the Minnesota and Missouri state assessments in reading and mathematics. The purpose of this comparison is to provide a descriptive account of the assessments' similarities and differences.

Overview

Minnesota Comprehensive Assessments

The Minnesota Comprehensive Assessment II reading test is a pencil-and-paper test covering three substrands of reading. Students read poetry and expository narratives. Depending on grade level, students respond to 40–50 items. Reading assessment questions use a multiple-choice or constructed-response format. Schools can administer the test in four separate segments on different days.

The Minnesota Comprehensive Assessment II mathematics test is a pencil-and-paper test covering four different mathematics strands. Depending on grade level, students respond to 44–50 items. Questions are multiple-choice and constructed- and gridded-response (grades 5 and higher). Like the reading test, schools can administer the math section in four separate segments given on different days.

Missouri Assessment Program

The Missouri Assessment Program communication arts test is a paper-and-pencil test requiring three to five hours of test administration time over three to four sessions (depending on grade level). The assessment includes 66–69 multiple-choice and constructed-response format questions and a writing prompt. The writing prompt is an open-ended item that requires students to demonstrate their writing proficiency. Writing is scored holistically using a four-point scoring guide.

The Missouri Assessment Program mathematics test is a paper-and-pencil test requiring three to five hours of test administration time across three to four sessions (depending on grade level). The assessment includes 67–77 multiple-choice and constructed-response format questions.

Scale reliability and validity

Reliability

To test the consistency of their assessments, Missouri and Minnesota State Department of Education researchers used the following reliability measures; reliability coefficients, standard error of measurement (SEM), and inter-rater reliability. Both states examined the internal consistency of their measures using coefficient alphas. Coefficient alphas for the 2008 assessment administration ranged from .88 to .91 across all domain-specific assessments in reading and math on the Minnesota Comprehensive Assessment II (table F1) (Minnesota Department of Education, 2008b). Coefficient alphas ranged from .91 to .92 across all domain-specific assessments in reading and math on the Missouri Assessment Program (Missouri

Department of Elementary and Secondary Education, 2008b). Additionally, each state examined SEM to determine the projected range of students' scores. Both states also recruited trained raters and used inter-rater reliability to examine the percent of agreement between raters on constructed-responses, such as written essay items and open-ended responses to reading comprehension and math items. Average inter-rater reliability correlations ranged from .76 to .94 on the Minnesota Comprehensive Assessment II (see table F1). Average Cohen's Kappa agreement on the Missouri Assessment Program ranged from .77 to .96. Across both state assessments, at least 97% of ratings were in perfect or adjacent agreement.

Validity

State Department of Education researchers in both states used multiple forms of validity testing by following the *Standards for Educational and Psychological Testing* (American Educational Research Association, 1999). In addition to providing evidence of content and criterion validity through careful test construction, appropriate test administration and scoring, accurate score scaling, and standard settings, each state also addressed construct validity—the degree to which the assessment measures the characteristic of interest.

Researchers for each State Department of Education investigated construct validity using principal components analyses (PCA). Scree plots provided evidence that the subject area tests are unidimensional, such that the first components explained the greatest amount of variance for each area test (see table F1). In Missouri, the first component of PCA explained 17–19 percent of the variance in grade 3–5 reading and mathematics scores and the first eigenvalue of PCA was 5–7 times larger than the second eigenvalue for grade 3–5 reading and mathematics assessments (Missouri Department of Elementary and Secondary Education, 2008b). In Minnesota, the first eigenvalue of PCA was 8–10 times larger than the second eigenvalue for grades 3–5 reading and mathematics assessments (Minnesota Department of Elementary and Secondary Education, 2008b).

Additionally, Missouri provided information on divergent validity (the relationship between constructs that should not be related to each other), reporting that individual scores on mathematics and reading assessments were related and ranged from 0.74–0.76. The Missouri Department of Education noted that scores were highly related but not perfectly overlapping, indicating the presence of different constructs (Missouri Department of Elementary and Secondary Education, 2008b). Minnesota did not provide any information on divergent validity.

Table F1. State assessment reliability, validity, and scale type data for 2007/08

	Minnesota Comprehensive Assessment II		Missouri Assessment Program	
Characteristic	**Reading**	**Math**	**Reading**	**Math**
Internal consistency				
Grade 3	0.90	0.90	0.91	0.92
Grade 4	0.91	0.90	0.91	0.92
Grade 5	0.88	0.90	0.91	0.91
Inter-rater reliability				
Grade 3	0.83[a]	0.83[a]	0.84[b]	0.96[b]
Grade 4	0.76[a]	0.94[a]	0.82[b]	0.96[b]
Grade 5	0.88[a]	0.94[a]	0.77[b]	0.96[b]
Construct Validity	PCA revealed the unidimensional nature of each subject area assessment. The first eigenvalue was 8–10 times larger than the second eigenvalue for grades 3–5 reading and mathematics assessments.		PCA revealed the unidimensional nature of each subject area assessment. The first component explained 17–19% of the variance in grade 3–5 reading and mathematics assessments and the first eigenvalue was 5–7 times larger than the second eigenvalue for grade 3–5 reading and mathematics assessments.	
Divergent Validity (Correlation between reading and math assessments)			Grade 3: $r = 0.76$ Grade 4: $r = 0.74$ Grade 5: $r = 0.75$	
Type of scores	Vertically scaled using progress scores; scale, raw and achievement scores available for content areas and content subscores		Vertically scaled to match *TerraNova*; scale and cut for each content area, content subscores	

a. Average inter-rater reliability assessed using correlations between ratings.
b. Average inter-rater reliability assessed using Cohen's Kappa.
Source: Minnesota Department of Education, 2008b, 2008e; Missouri Department of Elementary and Secondary Education, 2008b.

Subject matter covered in assessment tests by grade and state

To determine comparability of the knowledge and skills assessed by the two state assessments, technical manuals, test blueprints, and released items were obtained for the 2008 assessments to serve as data sources. The proportion of items per strand in each domain were computed and recorded and then compared across states (tables F2 and F4). Strands common to each state were identified and used to create a combined matrix for each domain, record proportions of items, and calculate differences in the proportions of items in each state assessment (tables F3 and F5). Based on the recommendation of May et al., Gleason (2009) recommends not combining state assessment results from across states if there are substantial differences in the knowledge and skills assessed, criteria were established for substantial difference as follows: a set of items per any strand represents greater than 40 percent of all items in the assessment or between states a difference greater than 10 percent in the proportion of items per any strand. If both criteria were met, the set of items was considered a potentially inappropriate set of items on which to combine results across states. The content and comparability of each domain, first reading, followed by mathematics, are discussed below.

Reading assessments

In Minnesota, the emphasis is on reading comprehension in each grade, both fiction and nonfiction, literal and interpretive, with explicit attention also paid to vocabulary. In Missouri, the emphasis is also on reading comprehension (nonfiction and fiction), but unlike Minnesota, this assessment includes items on writing (see table F2). Two of the criteria for "substantial difference" were met. However, at each grade level in each state assessment, at least 65 percent of the items address the same general set of knowledge and skills associated with reading comprehension. Therefore, the differences between the assessments do not appear substantial enough to preclude combining the results in reading across states.

Table F2. Distribution of reading assessment items by grade and state, 2007/08

	Grade		
Assessment and item	**3**	**4**	**5**
Minnesota Comprehensive Assessment II—reading and literature			
Vocabulary (reading, listening, and speaking vocabularies) (percent)	18	13	12
Comprehension (literal, interpretive, and evaluative comprehension) (percent)	36	41	47
Literature (understanding, analyzing, interpreting fiction, poetry and nonfiction) (percent)	46	46	41
Missouri Assessment Program—communication arts			
Speaking/writing standard English (grammar, punctuation, spelling) (percent)	24	16	21
Reading—fiction/poetry drama (percent)	38	65	30
Reading—nonfiction (percent)	28	16	46
Writing formally and informally (percent)	10	3	3

Source: Minnesota Department of Education, 2008c; Missouri Department of Elementary and Secondary Education, 2008b.

Table F3. Proportions of items by strand in state reading assessments, 2007/08

	Grade								
	3			4			5		
Reading	Minnesota	Missouri	Difference	Minnesota	Missouri	Difference	Minnesota	Missouri	Difference
Vocabulary (percent)	18	0	18	13	0	13	12	0	12
Reading comprehension (percent)	82	66	16	87	81	6	88	76	12
Writing (percent)	0	34	–34	0	19	–19	0	24	–24

Note: Gray cells indicate differences of greater than 10 percent between state assessments on reading test strands.

Math assessments

Minnesota places greater emphasis on number operations and number sense than does Missouri, which places greater emphasis on geometric spatial relationships and measurement (see table F4). The greatest difference is for the strand of grade 4 number sense (see table F5); however, in neither state assessment does the set of items in question exceed 40 percent of the items in the assessment and thus does not meet the criteria for a substantial difference. Therefore, the differences between the assessments do not appear substantial enough to preclude combining the results in mathematics across states.

Table F4. Distribution of mathematics assessment items by grade and state, 2007/08

	Grade		
Assessment and item	3	4	5
Minnesota Comprehensive Assessment II—mathematics			
Number sense (percent)	37	39	35
Patterns, function, and algebra (percent)	15	14	19
Data, statistics, and probability (percent)	19	18	24
Spatial sense, geometry, and measurement (percent)	29	29	22
Missouri Assessment Program—mathematics			
Number and operations (percent)	36	27	23
Algebraic relations (percent)	19	19	19
Geometric and spatial relations (percent)	19	17	19
Measurement (percent)	15	19	20
Data and probability (percent)	10	17	19

Note: Numbers may not sum to 100 percent because of rounding.
Source: Minnesota Department of Education 2008c; Missouri Department of Elementary and Secondary Education 2008b.

Table F5. Differences in proportions of items by strand in state mathematics assessments, 2007/08

	Grade								
	3			4			5		
Mathematics	**Minnesota**	**Missouri**	**Difference**	**Minnesota**	**Missouri**	**Difference**	**Minnesota**	**Missouri**	**Difference**
Number sense (percent)	37	36	1	39	27	12	35	23	12
Patterns, function, and algebra (percent)	15	19	–4	14	19	–5	19	19	0
Data, statistics, and probability (percent)	19	10	9	18	17	1	24	19	5
Spatial sense, geometry, and measurement (percent)	29	34	–5	29	36	–7	22	39	–17

Note: Numbers may not sum to 100 percent because of rounding. Gray cells indicate differences of greater than 10 percent between state assessments on reading test strands.

Assessment administration periods

Schools administer the Minnesota Comprehensive Assessment II and Missouri Assessment Program during the spring semester of each school year. District websites provided the following periods for administration (table F6).

Table F6. Periods for assessment administration by school year and state, 2007/08 and 2009/10

School year	Minnesota Comprehensive Assessment II	Missouri Assessment Program
2007/08	April 14–May 2, 2008	March 31–April 25, 2008
2009/10	April 12–April 30, 2010	March 29–April 23, 2010

Source: Minnesota Department of Education 2008b, 2010e; Missouri Department of Elementary and Secondary Education 2008b, 2010c.

Conclusion

The representation of reading and mathematics strands on state assessments varied by state and by grade level. In Minnesota, 82–88 percent of the reading assessment covered comprehension and 12–18 percent covered vocabulary. In Missouri, 66–81 percent of the reading assessment covered comprehension and 19–34 percent covered writing. Despite these differences, 66–81 percent of items across states address the same set of skills related to reading comprehension.

In Minnesota, the number sense strand accounts for 35–39 percent of test items. In contrast, in Missouri, the spatial sense, geometry and measurement strand represents 34–39 percent of math test items. Despite these differences, in both states, no strand makes up more than 40 percent of the math assessment.

The study's impact estimates of primary outcomes involve estimating an overall mean treatment and control group difference based on comparisons of multiple pairs of schools taking the same state assessments. Within each state, schools are held accountable for the content on each state assessment, and the contrast is the same across each of the school pairs (that is, comparing a treatment versus control school on the test it was accountable for). Therefore, there is no confound in which any treatment and control schools are directly contrasted across differing assessments and content. Based on this and the recommendation of May et al. (2009), researchers determined that it was appropriate to aggregate scores across states based on the similarities in knowledge and skills assessed across states.

Appendix G. Development and description of the teacher survey measuring teacher capacity for school improvement practices

This effectiveness study of Success in Sight measured three intermediate outcomes as part of teacher capacity for school improvement practices. The three teacher outcomes—data-based decisionmaking, purposeful community, and shared leadership—were measured using self-report surveys administered to both treatment and control teachers. This appendix begins with a description of the development of the measure for each outcome and then presents results of psychometric analyses on the measures of each outcome.

Instrument Development

The measures of teacher capacity for school improvement practices were derived from two existing teacher surveys: the Teacher Survey of Policies and Practices (Mid-continent Research for Education and Learning 2005) and the 12-item Goddard Collective Efficacy Scale (Goddard 2002).

The Teacher Survey of Policies and Practices was developed to study the organization of successful, high-poverty elementary schools (Apthorp et al. 2005). The researchers postulated a model representing four components of a school's organization, including leadership, school environment, professional community, and instruction. They based the model and the corresponding survey of teacher perceptions on a review of the effective schools research and research on successful high-poverty schools in particular. Based on this review, they conceptualized and defined four scales, each measured with three or four subscales. Coefficient alphas for these scales and subscales ranged from .77 to .95.

Although the Teacher Survey of Policies and Practices and the Success in Sight intervention share the same developer, they have different theoretical foundations and are not overaligned. The survey research team developed scales and subscales several years prior to the beginning of the current effectiveness study of Success in Sight. These scales and subscales were based on a review of successful, high-need schools. The Success in Sight development team based the intervention on a review of effective schools research (Marzano 2003) and the education change literature (for example, Fullan 2001, 2002).

The Goddard Collective Efficacy Scale measures school faculty perceptions of positive influence on student learning (Goddard 2002). Respondents use a 6-point Likert scale ranging from strongly disagree (1) to strongly agree (6).[50] Example items include "Teachers here are confident they will be able to motivate their students," "Home life provides so many advantages the students here are bound to learn," and "Teachers in this school do not have the skills to deal with student disciplinary problems." Research has demonstrated the construct validity and reliability of both the 21-item scale and the 12-item scale (Goddard, Hoy, and Hoy 2000). Goddard (2002) reported a coefficient alpha of .94 for the 12-item Collective Efficacy scale.

The survey used in this study to assess teacher capacity for school improvement included seven of the Teacher Survey of Policies and Practices subscales (Mid-continent Research for Education

[50] Likert scale items 2–5 do not have labels. Scores are presented on a continuum.

and Learning 2005) and the 12-item Goddard Collective Efficacy Scale (Goddard 2002). Table G1 shows each intermediate outcome and its coefficient alpha, each respective subscale and its coefficient alpha, and the number of items of each subscale. Coefficient alphas reported are based on the combined Minnesota and Missouri data used in the impact analysis for the study. Coefficient alphas were .76 for data based decisionmaking, .89 for purposeful community, and .96 for shared leadership. These alphas exceed the What Works Clearinghouse standards for reliable outcome measures (What Works Clearinghouse 2008).

Data-based decisionmaking was measured with the 8-item assessment and monitoring subscale from the Teacher Survey of Policies and Practices. The assessment and monitoring subscale (8 items) measures the degree to which school staff use various types of assessments to monitor student progress, provide feedback, and inform instructional decisions.

Purposeful Community was measured with the professional development and the collaboration subscales of the Teacher Survey of Policies and Practices and the Goddard Collective Efficacy Scale. The professional development subscale (8 items) measures the extent to which teachers report that their state-, district-, or school-sponsored professional development activities focused on academic content standards, content knowledge, and improving classroom practices. The collaboration subscale (8 items) measures the extent to which teachers work together on lesson planning, analyzing student test results, mentoring, and providing feedback to each other. The Goddard Collective Efficacy scale (12 items) assesses the degree to which school faculty believe that they have the joint capacity to positively influence student achievement.

Shared leadership was measured using the support for teacher influence subscale (eight items), the shared mission and goals subscale (six items), the instructional guidance subscale (six items), and the organizational change subscale (10 items) from the Teacher Survey of Policies and Practices. The support for teacher influence subscale measures teachers' perceptions of their involvement in important decisions and comfort in being able to voice concerns. The shared mission and goals subscale measures teachers' perception for a common vision of school improvement and shared beliefs and values about their school's mission. The instructional guidance subscale measures the degree to which school leadership provides guidance to teachers regarding effective classroom practice, ensures that teachers have the resources necessary for high-quality instruction, and monitors the effectiveness of classroom instructional practices. The organizational change subscale measures teachers' perceptions regarding seeking new and innovative ideas for teaching and making changes to improve student achievement.

Table G1. Subscales, alpha coefficients, and number of items used to measure intermediate outcomes of teacher capacity for school improvement practices

Intermediate teacher outcome	Contributing subscales	Number of items
Data-based decisionmaking (.76)	Teacher Survey of Policies and Practices assessment and monitoring subscale (.76)	8
Purposeful community (.89)	Teacher Survey of Policies and Practices professional development subscale (.93)	8
	Teacher Survey of Policies and Practices collaboration subscale (.83)	8
	Goddard Collective Efficacy Scale (.85)	12
Shared leadership (.96)	Teacher Survey of Policies and Practices support for teacher influence subscale (.88)	8
	Shared mission and goals (.93)	6
	Instructional guidance (.89)	6
	Organizational change (.85)	10

Note: numbers in parentheses are coefficients alphas, which are based on the combined Minnesota and Missouri data used in the impact analysis for the study.
Source: 2008 teacher survey.

All items on the teacher survey used a 5-point scale (1 = strongly disagree, 2 = somewhat disagree, 3 = neither agree nor disagree, 4 = somewhat agree, and 5 = strongly agree).[51] Researchers calculated total scores for each of the three outcomes (data-based decisionmaking, purposeful community, and shared leadership) by averaging the ratings from the items from the corresponding subscales. Ratings were scored carefully, taking into account positively or negatively worded questions.

Confirmatory Factor Analysis

Researchers ran a confirmatory factor analysis to examine the psychometric properties of the items and subscales used to measure the three teacher outcomes. Researchers ran the confirmatory factor analysis with each teacher outcome as a latent variable and its corresponding subscales as indicators (figure G1). Data-based decisionmaking had only one subscale, which was split into two indicators to accommodate the requirements of the confirmatory factor analysis model. The first indicator included three items, and the second indicator included five items. Researchers split the scale this way because the items for the first indicator were presented

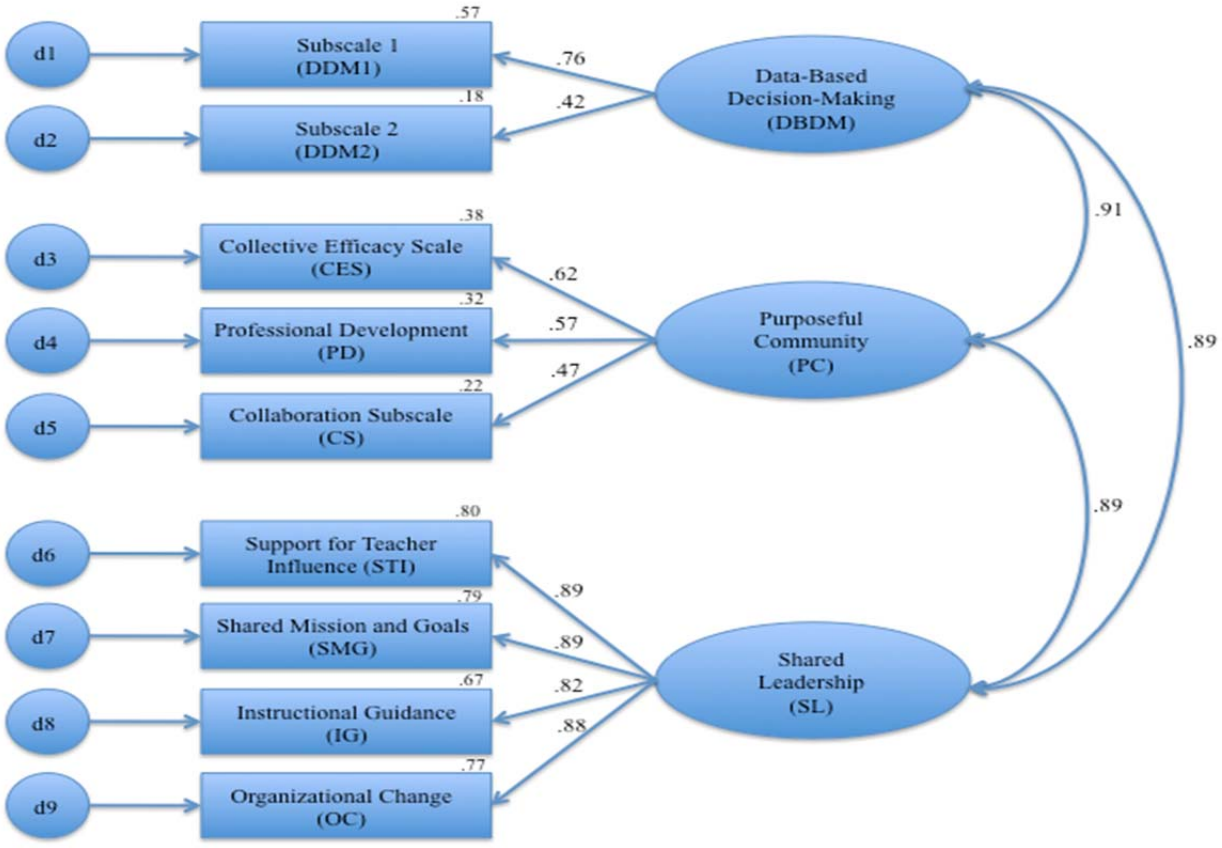
[51] Six items within the Collective Efficacy Scale subscale of the purposeful community scale were reverse coded to adjust for negatively valenced statements (1 = strongly agree, 2 = agree, 3 = neither agree nor disagree, 4 = somewhat disagree, 5 = strongly disagree).

together in one matrix on the survey, and the items for the second indicator were presented in a second matrix.

Results of the confirmatory factor analysis were as follows. The χ^2 test yielded a value of 266.63, which, evaluated with 24 degrees of freedom, has a corresponding *p*-value of < .01. This *p*-value is less than .05 and rejects the null hypothesis of a good fit (Loehlin 2004). A statistically significant χ^2 test, however, is common when the sample size is large. The root mean square of error approximation was .08, which researchers suggest represents reasonably good model fit (Brown and Cudeck 1993; Steiger 1989). Additional tests of model fit included the goodness of fit index (.96), the adjusted goodness of fit index (.92), and the comparative fit index (.96)—all with values higher than .90 that suggest a good model fit (Coursey 2008).

As can be seen in figure G1, the correlations between the three latent variables representing the three teacher outcomes were high. The correlation between purposeful community and data-based decisionmaking was .91. The correlation between shared leadership and purposeful community was .89, and the correlation between shared leadership and data-based decisionmaking was .89.

Figure G1. Results of the teacher survey confirmatory factor analysis, 2008



Source: 2008 teacher survey.

Researchers examined the results of the confirmatory factor analysis in terms of standardized regression weights and the squared multiple correlations for each indicator in the model (figure G1 and table G2). The standardized regression weights represent the correlation between the indicator and the latent variable. Two indicators had standardized loadings less than acceptable level of .50 as defined by Albright (2008): data-based decisionmaking (.42) and collaboration (.47). The remaining standardized loadings were all higher than .50.

The squared multiple correlations (R^2) represent the proportion of variance of the indicator accounted for by the latent variable. The R^2 values presented in table G2 show that the proportion of variance in four of the indicators was less than 50%. These were data-based decisionmaking 2 (.18), collaboration (.22), professional development (.32), and Collective Efficacy Subscale (.38). These results suggest that these subscales are weak indicators of their respective latent variables. Five of the R^2 values presented in table G2 were greater than 50% and suggest that these subscales were strong indicators of their respective latent variables. These subscales and their respective R^2 values presented in table G2 were: data-based decisionmaking 1 (.57), instructional guidance (.67), organizational change (.77), shared mission and goals (.79), and support for teacher influence (.80).

Table G2. Confirmatory factor analysis observed variable loadings on latent variables, 2008

Latent variable and indicator	Standardized loadings	R^2
Data-based decisionmaking		
Data-based decisionmaking 1	0.77	0.57
Data-based decisionmaking 2	0.42	0.18
Purposeful community		
Collective Efficacy Scale	0.62	0.38
Professional development	0.57	0.32
Collaboration	0.47	0.22
Shared leadership		
Support for teacher influence	0.89	0.80
Shared mission and goals	0.89	0.79
Instructional guidance	0.82	0.67
Organizational change	0.88	0.77

Source: 2008 teacher survey.

Researchers also examined the extent to which the indicators for each latent variable "converged," or shared variance; that is, the extent to which there was convergent validity. To evaluate the convergent validity, researchers examined mean variance extracted and the construct reliability loading for each latent variable from the confirmatory factor analysis. Mean variance extracted was calculated using the following equation:

$$AVE = \frac{\sum_{i=1}^{n} \lambda_i^2}{n}$$

Construct reliability was calculated using the following equation:

$$CR = \frac{(\sum_{i=1}^{n} \lambda_i)^2}{(\sum_{i=1}^{n} \lambda_i)^2 + (\sum_{i=1}^{n} \delta_i)}$$

Loehlin (2004) suggests that a mean variance extracted of .50 or higher indicates adequate convergent validity, while a mean variance extracted of less than .50 indicates that, on average, there is more error remaining in the items than there is variance explained by the latent factor structure imposed on the measure. For construct reliability, a measure of the internal consistency of the observed indicator variables, Loehlin (2004) suggests that a value of .70 or higher indicates good reliability and suggests the measures are consistently representing the observed indicator.

Results shown in table G3 suggest that one latent variable had adequate convergent validity (.76 for shared leadership), and two latent variables with less than adequate convergent validity: .37 for data-based decisionmaking and .31 for purposeful community. The latent variable for shared leadership had acceptable construct reliability, .93. Two latent variables had less than adequate construct reliability: .52 for data-based decisionmaking and .57 for purposeful community.

Table G3. Success in Sight latent variables construct mean variance extracted and construct reliability

Success in Sight latent variable	Mean variance extracted[a]	Construct reliability[b]
Data-based decisionmaking	0.37	0.52
Purposeful community	0.31	0.57
Shared leadership	0.76	0.93

a. Variance extracted calculated by taking the sum of the squared loadings for each factor and dividing by the number of loadings.
b. Construct reliability is computed from the sum of factor loadings (λi), squared for each construct, and the sum of the error variance terms for a construct (δi).
Source: 2008 teacher survey.

Conclusion

The coefficient alphas for the scales used to measure each intermediate teacher outcome as part of teacher capacity for school improvement practices were high and above acceptable levels. Results from the confirmatory factor analysis showed moderately good fit. The three latent variables in the confirmatory factor analysis represented that the three intermediate teacher outcomes were highly correlated. These high correlations suggest that the three intermediate outcomes represent highly related constructs of teacher capacity for school improvement practices. The confirmatory factor analysis results also suggest that shared leadership had adequate convergent validity and construct reliability. Additional results from the confirmatory factor analysis suggest that two of the latent variables—data-based decisionmaking and purposeful community—had less than adequate convergent validity and construct reliability.

Appendix H. Calculation of effect sizes

To describe the magnitude of the impact estimates for differences between treatment and control groups, researchers calculated effect sizes based on Glass's *d* approach (Glass, McGaw, and Smith 1981). For each effect size, the numerator was the difference between the adjusted treatment group and control group means, and the denominator was the control group standard deviation, calculated by taking the square root of the sum of the level 1 and level 2 variance components from the multilevel model (Spybrook et al. 2009). Using this approach yielded estimates of effect sizes expressed in standard deviation units of the control group, rather than pooled treatment and control group standard deviation units. This was important because schools most interested in participating in Success in Sight might be more similar to the control group schools than to treatment group schools that had participated in Success in Sight over the two-year study. For student achievement outcomes, the effect sizes are the same as the impact estimates because the impact estimates were already in standard deviation metrics (*z*-scores).

Appendix I. Procedures for handling missing data

This appendix describes the approaches used to handle missing data, which included listwise deletion, multiple imputation, and using available data. The type of approach implemented depended on the extent of missing data, available data, and the analytic models.

Impact analysis sample for primary outcomes

The benchmark analytic models for primary outcomes used posttest student achievement scores in reading and mathematics as outcome variables and baseline student achievement scores to calculate cluster-level covariates for mean school baseline student achievement in reading and mathematics. The impact analysis sample for primary outcomes included all students from participating schools with available reading or mathematics scores from the 2010 administration of the Minnesota and Missouri state assessments. Researchers examined the degree to which there was missing baseline and outcome data for students in the treatment and control group (table I1).

Table I1. Available and missing student achievement data, 2007/08 and 2009/10

	Treatment group				Control group			
Data category	Eligible students	Available scores	Missing scores	Response rate (percent)	Eligible students	Available scores	Missing scores	Response rate (percent)
Reading								
Baseline	4,705	4,665	40	99.15	3,904	3,802	102	97.39
Posttest	4,473	4,403	70	98.44	3,867	3,779	88	97.72
Math								
Baseline	4,557	4,519	38	99.17	3,881	3,812	69	98.22
Posttest	4,468	4,413	55	98.77	3,861	3,800	61	98.42

Source: Minnesota Department of Education 2008a, 2010b; Missouri Department of Elementary and Secondary Education 2008a, 2010b.

The amount of missing data at any specific data point was less than 3 percent (see table I1). Research suggests that listwise deletion will not contribute to consequential bias (that is, bias greater than 0.05 standard deviation of the outcome measure) and loss of power when missing data is less than approximately 5 percent (for example, see Graham, Cumsille, and Elek-Fisk 2003; Graham 2009; Puma et al. 2009). In addition, listwise deletion is appropriate for a variety of analyses and does not require specialized software. Therefore, researchers used listwise deletion to handle missing student achievement data. In other words, students who were missing test scores were excluded from analyses.

Impact analysis sample for secondary outcomes

The benchmark analytic models for secondary outcomes used posttest capacity for school improvement practice scores (from the teacher survey) as outcome variables and baseline school improvement practice scores (from the teacher survey) to calculate cluster-level covariates for baseline mean school improvement practices. For the impact analysis sample for secondary outcomes, there were two types of missing data—when teachers did not respond to specific items

(item-level nonresponse) and when eligible teachers did not complete either the baseline or posttest survey (wave nonresponse) (table I2).

Table I2. Available and missing teacher survey data, 2008 and 2010

Data category	Eligible teachers	Complete responses	Incomplete responses	Percent with incomplete responses	Wave nonresponses	Percent with wave nonresponses
Treatment						
Baseline survey	819	429	321	39.19	69	8.42
Posttest survey	825	583	232	28.12	10	1.21
Control						
Baseline survey	755	342	282	37.35	131	17.35
Posttest survey	737	514	187	25.37	36	4.88

Source: 2008 and 2010 teacher survey.

Item-level missing data

Item-level nonresponse led to missing data for 39.19 percent of treatment group cases and 37.35 percent of control group cases at baseline and 28.12 percent of treatment group cases and 25.37 percent of control group cases at posttest (see table I2). Because more than 5 percent of the baseline and posttest teacher survey cases had missing item-level data and because there were appropriate data available to include in the imputer's model, researchers determined that it would be beneficial to impute missing item-level data. Specifically, researchers used multiple imputation with the multiple imputation with chained equations procedure (Van Buuren and Groothuis-Oudshoorn forthcoming) because it offered flexibility in handling data with different levels of measurement.

The multiple imputation with chained equations procedure uses three steps to implement multiple imputation: creates multiple versions of the imputed data sets by using existing values to predict missing variables, performs repeated statistical analysis to incorporate missing data uncertainty on each of the imputed data sets, and combines the results of the analyses (mean) to produce one set of results (Van Buuren and Oudshoorn 1999). Multiple imputation maintains overall variability in the missing data by creating imputed values based on variables correlated with the missing data. Uncertainty is accounted for by creating different versions of the missing data and observing the variability between imputed data sets (Rubin 1987, 1996).

To simplify matters and avoid any potential confound, researchers separated the items by subscale and ran the multiple imputation with chained equations procedure by subscale.[52] Researchers ran the multiple imputation with chained equations procedure for each of the 13 subscales (12 for the Teacher Survey of Policies and Practices and 1 for the Collective Efficacy Scale) by state and by treatment group.[53] These multiple imputation procedures produced five

[52] The confound can occur if a spurious relationship exists among items and the use of chained equations exacerbates that spurious relationship by imputing values that conform more to the spurious relationship than to the internally consistent relationship among the subscale items.

[53] Researchers ran these separately by state to protect against potential confounds that could have been introduced by state.

complete datasets for each subscale. To determine whether the internal consistency of the subscales was congruent between the two state samples and across the five multiply imputed datasets, researchers calculated coefficient alpha for each subscale, within each state, and within each dataset (tables I3 and I4). The findings indicated high internal consistency within all subscales. The findings indicated congruence between datasets across states and across imputations. Based on these analyses and findings that the rate of missing information (Rubin 1987; McKnight et al. 2007) revealed almost no information was lost due to the imputation ($\gamma <$ .0001), researchers determined it was appropriate to select one imputed dataset at random for use as the final, complete dataset for all subsequent analyses.

Table I3. 2008 Coefficient alphas by scale, subscale, state, and imputed data set

	MN 1	MN 2	MN 3	MN 4	MN 5	MO 1	MO 2	MO 3	MO 4	MO 5
Assessment and Monitoring	.74	.74	.73	.74	.74	.75	.74	.74	.74	.74
Purposeful Community	.79	.79	.79	.79	.79	.84	.84	.84	.84	.84
Professional Development	.91	.91	.91	.91	.91	.94	.94	.94	.94	.94
Collaboration	.81	.81	.81	.81	.80	.83	.83	.83	.83	.83
Collective Efficacy	.83	.83	.83	.83	.83	.85	.85	.85	.85	.85
Shared Leadership	.95	.95	.95	.96	.96	.96	.96	.96	.96	.96
Teacher Influence	.87	.87	.87	.87	.87	.88	.88	.88	.88	.88
Shared Mission and Goals	.93	.93	.93	.93	.93	.91	.91	.91	.91	.91
Instructional Guidance	.87	.87	.87	.87	.87	.89	.88	.89	.88	.88
Organizational Change	.84	.84	.84	.84	.84	.85	.85	.85	.85	.85

Source: 2008 teacher survey.

Table I4. 2010 Coefficient alphas by scale, subscale, state, and imputed data set

	MN 1	MN 2	MN 3	MN 4	MN 5	MO 1	MO 2	MO 3	MO 4	MO 5
Assessment and Monitoring	.73	.72	.73	.72	.72	.76	.76	.76	.76	.76
Purposeful Community	.82	.83	.82	.82	.83	.85	.85	.85	.85	.85
Professional Development	.93	.93	.93	.93	.93	.94	.94	.94	.94	.94
Collaboration	.83	.83	.83	.83	.83	.83	.83	.83	.83	.83
Collective Efficacy	.84	.84	.84	.84	.84	.85	.85	.85	.85	.85
Shared Leadership	.95	.95	.95	.95	.95	.96	.96	.96	.96	.96
Teacher Influence	.86	.86	.86	.86	.86	.89	.89	.89	.89	.89
Shared Mission and Goals	.90	.91	.91	.91	.91	.93	.93	.93	.93	.93
Instructional Guidance	.86	.86	.86	.86	.86	.88	.88	.88	.88	.88
Organizational Change	.84	.84	.84	.84	.84	.85	.85	.85	.85	.85

Source: 2010 teacher survey.

Wave-level missing data

Teacher nonresponse led to missing data for 8.42 percent of treatment group cases and 17.35 percent of control group cases at baseline and 1.21 percent of treatment group cases and 4.88 percent of control group cases at posttest (see table I2). Because the amount of missing posttest data was less than 5 percent, researchers used listwise deletion to address missing posttest wave-level teacher survey data for the outcome variable. Because the amount of missing baseline wave-level was more than 5 percent, researchers considered data- and model-based procedures, such as multiple imputation and the dummy variable method, to address missing data (Puma et al. 2009). However, researchers determined that multiple imputation was not appropriate because this study did not link teachers' baseline and posttest responses, making it impossible to use teachers' available responses to impute data for missing responses. Furthermore, because the analytic model included a cluster-level covariate calculated from wave-level data rather than an individual-level covariate, the dummy variable method was inappropriate for addressing the wave-level missing data. Therefore, the analytic models for secondary outcomes included cluster-level covariates (one for each model) calculated from available data.

Appendix J. Procedures to control for multiple comparisons

For this study, researchers applied the Benjamini-Hochberg correction for multiple comparisons for statistically significant findings regarding impact analyses on primary and secondary outcomes. Specifically, this correction was applied to multiple comparisons within the two achievement domains of reading and mathematics and to multiple comparisons within the three teacher capacity for school improvement domains of data-based decision making, purposeful community, and shared leadership. This correction was applied as follows:

1) Researchers determined the number of statistically significant findings within each domain for impact analysis of primary and secondary outcomes. Within each domain, the number of statistically significant findings was denoted by *m*.
2) Researchers rank ordered each of the *m* statistically significant findings based on their corresponding *p*-values, so that $p_1 \leq p_2 \leq p_3 \leq ...p_m$.
3) For each of the *m* statistically significant findings, researchers computed *pi'* using the following formula:

   $pi' = i\alpha/m$

   in which *i* represents the rank for each statistically significant *p*-value, α represents the study's target level determining statistical significance (.05), and *m* represents the number of statistically significant findings within the domain.
4) Researchers identified the largest *p*-value rank (*i*) for which the original *p*-value was greater than or equal to *pi'* to establish the cut-point for statistical significance based on the Benjamini-Hochberg correction. Findings with *p*-values less than or equal to this cut point were considered statistically significant after applying the correction, and findings with *p*-values greater than this cut point were not considered statistically significant after applying the correction.

Appendix K. Meta-analytic methods for combining state-specific impact estimates

To test the robustness of the benchmark impact analyses of primary outcomes to the methods used to combine estimates across state samples, researchers conducted sensitivity analyses using student achievement scale scores (instead of *z*-scores) to estimate separate models for each state, and combined the results meta-analytically. Specifically, after running each model, researchers used comprehensive meta-analysis software (Borenstein and Rothstein 1999) to compute the overall, weighted mean effect using the treatment and control means, standard deviations, and sample sizes (table K1). The procedure involved standardizing each impact estimate (by calculating separate effect size estimates for each state), weighting the separate effects to retain the characteristics of each state's assessment in terms of the variability and sample size, and combining the weighted effects by computing the weighted mean effect using the new standard error following the procedures outlined in Shadish and Haddock (1994). The uncertainty of the estimate of effect in terms of its standard error was quantified by computing a confidence interval with the usual formula:

$$\bar{T}_{\bullet} \pm z_{\alpha/2}\sqrt{v_{\bullet}}$$

This allowed researchers to report upper and lower limits around the overall, weighted mean effect. To accept or reject the null hypothesis, the overall weighted mean effect was compared with the two-tailed critical *z*-value of the standard normal distribution and an alpha level of 0.05.

Table K1. Means, sample sizes, and standard deviations used for meta-analytic calculation of weighted effect size, 2009/10

Data	Treatment group mean	Treatment group sample size	Control group mean	Control group sample size	Standard deviation
Minnesota regression-adjusted posttest reading *z*-score	3,593.04	12	3,595.76	12	289.76
Missouri regression-adjusted posttest reading *z*-score	650.97	14	651.12	14	40.78
Minnesota regression-adjusted posttest mathematics *z*-score	3,588.65	12	3,622.99	12	253.17
Missouri regression-adjusted posttest mathematics *z*-score	639.96	14	640.12	14	43.53

Note: The means and standard deviations were from multilevel models that accounted for the nesting of students in schools. The means were also regression adjusted, and the standard deviations were from control group null models.
Source: Minnesota Department of Education 2008a, 2010b; Missouri Department of Elementary and Secondary Education 2008a, 2010b.

Appendix L. Comparisons of the local context for treatment and control schools

The tables in this appendix include descriptive information relevant to the local context of treatment and control schools. Tables L1–L3 are based on publicly available data from the Minnesota and Missouri state departments of education. Tables L4–L6 are based on qualitative data from interviews with 155 participants in treatment and control schools. Interviewees included principals, leadership team members, and classroom teachers (see chapter 2 for more information on data collection). Because the purpose is to make comparisons at the school level in tables L4–L6, researchers present the numbers of schools in each table rather than the number of interviewees. A school was counted one time in each category if at least one interviewee reported the school had experienced change in a particular area (tables L4 and L5) or had participated in a school improvement initiative (table L6). Because Tables L4–L6 are based on qualitative interview data intended to provide descriptive information about the local contexts of schools, it is not appropriate to run tests of statistical significance with these data. Some data are suppressed to preserve school anonymity.

Table L1. Percentage of schools making adequate yearly progress in prior three years (2005/06, 2006/07, 2007/08), by condition

Characteristic	Treatment (*n* = 26) Prior adequate yearly progress status (percent)	Control (*n* = 26) Prior adequate yearly progress status (percent)	Test statistic	*p*-value
At-risk for failing to make adequate yearly progress, but made adequate yearly progress in all adequate yearly progress criteria for all three years (2005/06, 2006/07, 2007/08)	8	23	11.10	0.01
Failed to make adequate yearly progress one of three years (2005/06, 2006/07, 2007/08) in any adequate yearly progress criterion	42	15		
Failed to make adequate yearly progress in two of the three years (2005/06, 2006/07, 2007/08) in any adequate yearly progress criterion	19	50		
Failed to make adequate yearly progress in all three years (2005/06, 2006/07, 2007/08) in any adequate yearly progress criterion	31	12		

Note: Analyses were 4 by 2 chi-square tests between the prior adequate yearly progress status frequency for total treatment groups compared to control groups. Adequate yearly progress status is based on all students tested within schools and with regard to state adequate yearly progress criteria.
Source: Minnesota Department of Education 2008d; Missouri Department of Elementary and Secondary Education 2010d.

Table L2. Number of schools making adequate yearly progress by state and experimental group, 2007/08 and 2009/10

Characteristic	Treatment			Control			Test statistic	*p*-value
	Minnesota (*n* = 12) Spring 2008	Missouri (*n* = 14) Spring 2008	Total (*n* = 26) Spring 2008	Minnesota (*n* = 12) Spring 2008	Missouri (*n* = 14) Spring 2008	Total (*n* = 26) Spring 2008		
Number of schools not making adequate yearly progress in reading								
All students	[a]	[a]	6	[a]	[a]	11	1.65	0.20
Number of schools making adequate yearly progress in mathematics								
All students	5	6	11	3	11	14	0.32	0.57
Number of schools making adequate yearly progress in both reading and mathematics								
All students	7	4	11	5	9	14	1.22	0.54

	Treatment			Control			Test statistic	*p*-value
	Minnesota (*n* = 12) Spring 2010	Missouri (*n* = 14) Spring 2010	Total (*n* = 26) Spring 2010	Minnesota (*n* = 12) Spring 2010	Missouri (*n* = 14) Spring 2010	Total (*n* = 26) Spring 2010		
Number of schools not making adequate yearly progress in reading								
All students	5	3	8	6	6	12	0.85	0.36
Number of schools making adequate yearly progress in mathematics								
All students	7	6	13	10	6	16	0.28	0.60
Number of schools making adequate yearly progress in both reading and mathematics								
All students	[a]	[a]	7	5	3	8	2.20	0.33

Note: Analyses were 2 by 2 chi-square tests between the adequate yearly progress frequency for total treatment compared to control groups. Adequate yearly progress status is based on all students tested within schools.
a. Value suppressed to preserve anonymity.
Source: Minnesota Department of Education 2008d, 2010d; Missouri Department of Elementary and Secondary Education 2010d.

Table L3. Comparison of treatment and control sample demographics, 2007/08 and 2009/10

	Treatment						Control					
	Minnesota		Missouri		All		Minnesota		Missouri		All	
Characteristic	Spring 2008	Spring 2010	Spring 2008	Spring 2010	Spring 2008	Spring 2010	Spring 2008	Spring 2010	Spring 2008	Spring 2010	Spring 2008	Spring 2010
Total enrollment [a] (grades 3–5)	2,100	1,992	2,575	2,432	4,675	4,424	1,809	1,796	2,033	2,013	3,842	3,809
Students eligible free or reduced-price lunch (percent)	82.48	84.04	65.05	69.70	72.88	76.15	82.53	82.02	69.70	72.28	75.77	76.87
Student population[b] (percent)												
White	14.05	12.90	54.21	54.98	36.28	36.03	16.75	19.15	51.16	53.85	34.96	37.49
Black	37.14	35.84	7.18	7.44	20.59	20.23	32.95	33.74	3.49	2.93	17.36	17.46
Hispanic	11.33	11.80	34.87	33.51	24.24	23.73	15.87	13.42	43.53	41.63	30.50	28.33
Asian	34.38	35.49	0.00	0.00	15.43	15.98	33.00	32.24	0.00	0.00	15.54	15.20
American Indian	3.10	3.97	2.60	3.33	2.80	3.62	1.44	1.39	0.79	1.09	1.09	1.23
Other	na	na	0.89	0.74	0.49	0.41	na	na	0.74	0.50	0.39	0.26

na is not applicable

Note: Percentages were calculated using total enrollment for the denominator.

a. Includes students enrolled in treatment and control schools in grades 3–5 at the time of the reading or mathematics state assessments.

b. Components may not sum to 100 because of rounding and because states did not provide information for 27 students.

Source: Minnesota Department of Education 2008a, 2010b; Missouri Department of Elementary and Secondary Education 2008a, 2010b; authors' compilation.

Table L4. Number of treatment and control schools reporting school improvement influenced by other student and budget changes by state, 2008–2010

	Treatment			Control		
Change area	Minnesota (*n* = 12)	Missouri (*n* = 14)	Total (*n* = 26)	Minnesota (*n* = 12)	Missouri (*n* = 14)	Total (*n* = 26)
Student demographics	[a]	[a]	3	[a]	[a]	5
Student enrollment	[a]	[a]	4	[a]	[a]	4
Budget cuts	3	3	6	[a]	[a]	8

a. Value suppressed to preserve anonymity.

Source: Principal, leadership team, and staff interviews, spring 2010.

Table L5. Number of treatment and control schools reporting changes in local education policies and practices by state, 2008–2010

	Treatment			Control		
Change area	Minnesota (*n* = 12)	Missouri (*n* = 14)	Total (*n* = 26)	Minnesota (*n* = 12)	Missouri (*n* = 14)	Total (*n* = 26)
Grade-level configuration	[a]	[a]	[a]	[a]	[a]	3
Curriculum	12	14	26	12	14	26
Instruction	[a]	[a]	4	[a]	[a]	3
Assessment	[a]	[a]	4	[a]	[a]	4
Start time	0	4	4	0	3	3

a. Value suppressed to preserve anonymity.
Source: Principal, leadership team, and staff interviews, spring 2010.

Table L6. Number of treatment and control schools reporting school improvement initiatives by state, 2008–2010

	Treatment			Control		
Initiatives	Minnesota (*n* = 12)	Missouri (*n* = 14)	Total (*n* = 26)	Minnesota (*n* = 12)	Missouri (*n* = 14)	Total (*n* = 26)
Reading First	—	14	14	—	14	14
Mondo	12	0	12	12	0	12
Phonological Awareness Literacy Screening	12	—	12	12	—	12
Regional Professional Development Centers School Improvement Services	na	8	8	na	3	3
Leadership academies	[a]	[a]	3	6	0	6
Professional learning communities	12	7	19	10	8	18
Response to intervention	[a]	[a]	6	[a]	—	[a]

— is not reported during interview.
na is not applicable to state.
a. Value suppressed to preserve anonymity.
Source: Principal, leadership team, and staff interviews, spring 2010.

Table L7. Curriculum developers' estimations of professional development and implementation time spent on school improvement initiatives, 2008-2010.

Initiatives	Professional development components	Estimated professional development time per school (hours)	Implementation components	Estimated implementation time per school (hours)
Success in Sight (26 treatment schools)	Six professional development 15-hour sessions over two years (estimated 90 hours) 10 six-hour site visits over two years (estimated 60 hours) Eight additional two-hour meetings without Success in Sight facilitators over two years (estimated 16 hours)	166 (over two years)	Weekly implementation of fractal experiences over two years (estimated 80 hours) Three hours of distance support each month over two years through phone or email (estimated 72 hours)	152 (over two years)
Reading First (14 treatment schools, 14 control schools)	Mean of six reading professional development workshops totaling 31 hours	31 (over one year)	Mean of 103 minutes spent on daily reading activities for forty instructional weeks (estimated 343 hours)	343 (over one year)
Mondo (12 treatment schools, 12 control schools)	Six two-day workshops for principals and literacy coaches (estimated 96 hours) Five weeks of site visits by specialists (unknown amount of time)	96 (over one year)	Weekly meetings with other classroom teachers for literacy planning and preparation (estimated 40 hours) 90-minute daily reading block for 40 instructional weeks (estimated 300 hours)	340 (over one year)

Initiatives	Professional development components	Estimated professional development time per school (hours)	Implementation components	Estimated implementation time per school (hours)
Phonological Awareness Literacy Screening (12 treatment schools, 12 control schools)	Phonological Awareness Literacy Screening district coordinator provides teachers with in-service sessions and site observations (unknown amount of time)	na	All K–3 students tested once in the fall (10–25 minutes per student; estimated 67–167 hours for 400 students, mean 117 hours) Students scoring below grade level are tested throughout the year and in the spring (10–25 minutes per student, estimated 17–42 hours for 100 students, mean 30 hours) Students scoring below grade level in the fall receive 2.5 hours each week of intervention instruction (estimated 100 hours)	247 (over one year)
Regional Professional Development Centers School Improvement Services (8 treatment schools, 3 control schools)	Time spent varies by course Four- to eight-hour workshops meeting one to four times over a month period on topics encompassing a wide variety of areas (such as school improvement, assessment)	4–32 (per workshop)	Professional development can be implemented in a variety of ways (unknown amount of time)	na

Initiatives	Professional development components	Estimated professional development time per school (hours)	Implementation components	Estimated implementation time per school (hours)
Leadership academies (8 treatment and control schools)	15 full day professional development workshops for principals (estimated eight-hour days at 120 hours total) Two-day trip for principals to visit schools in another state (estimated eight-hour days at 16 hours total) Four, two-hour Saturday seminars for principals (estimated eight hours)	168 (over one year)	Commitment by principal to mentor a future program participant (unknown amount of time) Monthly meetings (principal) with a mentor who has completed the Academy and a business mentor (unknown amount of time)	na
Professional learning communities (Minnesota) (12 treatment schools, 10 control schools)	Workshops on a wide variety of topics (such as data use, improving staff meetings) Workshops range from four four-hour sessions (estimated 16 hours) to three six-hour sessions (estimated 18 hours)	16–18 (per workshop)	Professional development experiences from workshops vary and can be implemented in a variety of ways (unknown amount of time)	na
Professional learning communities (Missouri) (7 treatment schools, 8 control schools)	Leadership teams attend a three-day summer workshop (estimated 24 hours) and participate in seven eight-hour trainings in year one, five eight-hour trainings in year two, and three eight-hour trainings in year three (estimated 120 hours) Leadership team members participate in a two-day conference each year (estimated 48 hours)	192 (over three years)	Over the three years, teams receive onsite assistance and participate in observations (unknown amount of time)	na

Initiatives	Professional development components	Estimated professional development time per school (hours)	Implementation components	Estimated implementation time per school (hours)
Response to intervention (6 treatment schools)	School districts provide access to response to intervention webinars and seminars varying in length from 20-90 minutes	.33–1.5 (per webinar)	Schools implement RTI (unknown amount of time)	na

na is not available.

Note: Time estimations are based upon data from curriculum developers.

Source: Success in Sight program records; U.S. Department of Education 2008b; school district website, identity protected; Phonological Awareness Literacy Screening 2007d; Missouri Department of Elementary and Secondary Education 2010e; Center for School Change 2010**;** Education Minnesota 2010; Missouri Department of Elementary and Secondary Education 2009a; Missouri Department of Elementary and Secondary Education 2011.

Appendix M. Raw means and standard deviations

In this appendix, vertically-scaled scores were used to calculate raw means and raw standard deviations for student scores by grade and state, and survey scores were used to calculate raw means and raw standard deviations for teacher characteristics. Reported raw means and standard deviations were not adjusted for covariates or clustering of students and teachers within schools.

Table M1. Raw means and standard deviations for Minnesota student achievement scores, 2007/08 and 2009/10

	Treatment (schools = 12)			Control (schools = 12)		
Measure	***n*** **(students)**	**Mean**	**Standard deviation**	***n*** **(students)**	**Mean**	**Standard deviation**
Reading Baseline						
Grade 3	702	3501.03	271.79	620	3485.71	279.39
Grade 4	680	3626.70	301.79	587	3581.25	282.76
Grade 5	715	3694.28	275.13	564	3689.62	274.82
Total	2097	3607.67	294.08	1771	3582.31	291.08
Reading Posttest						
Grade 3	663	3479.65	258.68	588	3478.32	267.03
Grade 4	681	3614.08	286.52	611	3580.75	263.04
Grade 5	634	3710.99	268.24	568	3725.46	285.29
Total	1978	3600.08	287.33	1767	3593.18	289.49
Mathematics Baseline						
Grade 3	655	3516.60	228.98	625	3506.80	225.30
Grade 4	641	3615.34	228.91	595	3575.47	233.78
Grade 5	651	3713.20	228.79	562	3712.74	223.72
Total	1947	3614.84	242.54	1782	3594.68	242.92

Measure	Treatment (schools = 12)			Control (schools = 12)		
	n (students)	Mean	Standard deviation	*n* (students)	Mean	Standard deviation
Mathematics Posttest						
Grade 3	664	3518.54	219.25	601	3531.80	231.28
Grade 4	685	3589.90	232.47	617	3601.68	229.47
Grade 5	637	3706.89	232.32	575	3741.66	250.84
Total	1986	3603.56	240.58	1793	3623.15	252.25

Note: Raw means and standard deviations are from descriptive statistics of measure by grade by condition.
Source: Minnesota Department of Education 2008a, 2010b.

Table M2. Raw means and standard deviations for Missouri student achievement scores, 2007/08 and 2009/10

Measure	Treatment (schools = 14)			Control (schools = 14)		
	n (students)	Mean	Standard deviation	*n* (students)	Mean	Standard deviation
Reading Baseline						
Grade 3	854	631.89	37.17	698	626.61	42.78
Grade 4	844	652.14	35.49	669	647.74	37.26
Grade 5	870	665.37	31.86	664	662.98	38.64
Total	2568	649.89	37.51	2031	645.46	42.39
Reading Posttest						
Grade 3	833	631.05	35.91	690	630.17	35.20

Measure	Treatment (schools = 14)			Control (schools = 14)		
	n (students)	Mean	Standard deviation	*n* (students)	Mean	Standard deviation
Grade 4	803	657.07	39.36	657	649.62	41.41
Grade 5	789	666.55	39.70	665	666.55	38.57
Total	2425	651.22	41.18	2012	648.55	41.21
Mathematics Baseline						
Grade 3	856	616.59	37.36	698	612.32	39.03
Grade 4	843	638.59	36.15	669	638.64	34.30
Grade 5	873	653.63	43.79	663	648.03	47.30
Total	2572	636.37	42.14	2030	632.66	43.26
Mathematics Posttest						
Grade 3	832	616.69	38.36	689	616.31	40.11
Grade 4	803	643.16	37.35	655	638.85	35.91
Grade 5	792	659.58	46.99	663	656.67	45.41
Total	2427	639.44	44.72	2007	637.00	43.92

Note: Raw means and standard deviations are from descriptive statistics of measure by grade by condition.
Source: Missouri Department of Elementary and Secondary Education 2008a, 2010b.

Table M3. Raw means and standard deviations for teacher outcomes, 2008 and 2010

	Treatment (schools = 26)			Control (schools = 26)		
Measure	***n*** **(teachers)**	**Mean (survey score)**	**Standard deviation**	***n*** **(teachers)**	**Mean (survey score)**	**Standard deviation**
Data-based decisionmaking						
Baseline	750	4.41	0.53	624	4.43	0.53
Posttest	815	4.48	0.48	701	4.49	0.50
Purposeful community						
Baseline	750	3.30	0.64	624	3.32	0.61
Posttest	815	3.43	0.66	701	3.44	0.69
Shared leadership						
Baseline	750	3.78	0.83	624	3.87	0.80
Posttest	815	3.98	0.73	701	3.88	0.83

Note: Raw means and standard deviations were from descriptive statistics of measure by condition.
Source: 2008 and 2010 teacher survey.

Appendix N. Variance components estimates and intraclass correlations

This appendix presents the estimates for variance components and intraclass correlations from the following null models: null models with full sample run on student outcomes, null models with student stayer sample run on student outcomes, null models run on student outcomes (scale scores) separately by state, and null models run on capacity for school improvement practice outcomes. Null models were multilevel models run on outcome variables of interest that did not include any level-1 or level-2 predictors as covariates. Each null model equated to one-way analyses of variance with random effects (Raudenbush and Bryk 2002). Running these null models yielded estimates of the level 1 variance (σ^2), which is the variance that occurs within groups, and the level 2 variance (τ_{00}), which is the variance that occurs between groups. This also enabled researchers to calculate an intraclass correlation coefficient for each model, which is a ratio of the between-group variance to the total variance. Calculating intraclass correlation coefficients allowed researchers to verify that multilevel modeling was an appropriate analytic approach for the impact estimates

Table N1. Variance components and intraclass correlation coefficients from null model for student outcomes with full sample, 2009/10

Measure	Variance within groups (σ^2)	Variance between groups (τ_{00})	Total variance ($\sigma^2 + \tau_{00}$)	Intraclass correlation coefficient (τ_{00})/ ($\sigma^2 + \tau_{00}$)
Reading	0.89	0.18	1.07	.17
Mathematics	1.57	0.25	1.82	.14

Note: These models included all students with available baseline or posttest achievement data.
Source: Minnesota Department of Education 2010b; Missouri Department of Elementary and Secondary Education 2010b.

Table N2. Variance components and intraclass correlation coefficients from null model for student outcomes with stayer sample, 2009/10

Measure	Variance within groups (σ^2)	Variance between groups (τ_{00})	Total variance ($\sigma^2 + \tau_{00}$)	Intraclass correlation coefficient (τ_{00})/ ($\sigma^2 + \tau_{00}$)
Reading	0.84	0.18	1.02	.18
Mathematics	0.85	0.26	1.11	.23

Note: These models included all students with available baseline or posttest achievement data.
Source: Minnesota Department of Education 2010b; Missouri Department of Elementary and Secondary Education 2010b.

Table N3. Variance components and intraclass correlation coefficients from null models for sensitivity analysis with separate models for Minnesota and Missouri, 2009/10

Measure	Variance within groups (σ^2)	Variance between groups (τ_{00})	Total variance ($\sigma^2 + \tau_{00}$)	Intraclass correlation coefficient (τ_{00})/ ($\sigma^2 + \tau_{00}$)
Minnesota scale score				
Reading	75,736.32	7,936.41	83,672.73	.09
Mathematics	55,279.94	5,799.64	61,079.58	.10
Missouri scale score				
Reading	1,482.96	189.20	1,672.16	.11
Mathematics	1,654.74	290.08	1,944.82	.15

Source: Minnesota Department of Education 2010b; Missouri Department of Elementary and Secondary Education 2010b

.

Table N4. Variance components and intraclass correlation coefficients from null model for school improvement practices outcomes, 2010

Measure	Variance within groups (σ^2)	Variance between groups (τ_{00})	Total variance ($\sigma^2 + \tau_{00}$)	Intraclass correlation coefficient (τ_{00})/ ($\sigma^2 + \tau_{00}$)
Data-based decisionmaking	0.22	0.03	0.25	.12
Purposeful community	0.29	0.07	0.36	.19
Shared leadership	0.43	0.20	0.63	.32

Source: 2010 teacher survey.

Appendix O. Supporting tables for impact analyses of primary outcomes

This appendix provides supporting tables for the impact analyses of primary outcomes.

Baseline means, standard errors, and effect sizes for impact analyses of primary outcomes

Researchers conducted multilevel modeling to examine the difference between baseline treatment and control group mean student achievement. The results reveal no statistically significant differences between treatment and control groups on their baseline student achievement means in reading or mathematics (table O1).

Table O1. Baseline means, standard errors, and effect sizes for student achievement outcomes

	Treatment			Control			Estimated difference				
Baseline measure (*z*-score)	Mean	Standard deviation	Sample size	Mean	Standard deviation	Sample size	Value	Standard error	95 percent confidence interval	*p*-Value	Effect size[a]
Reading	–0.36	1.04	4,665	–0.41	1.10	3,802	0.05	0.12	–0.19–0.29	.70	0.05
Math	–0.37	1.06	4,519	–0.39	1.07	3,812	0.02	0.13	–0.24–0.28	.87	0.02

a. Calculated by dividing the estimated difference in means by the control group standard deviation.

Note: Results are from multilevel models that account for the nesting of students in schools. Analyses included 26 schools in the treatment group and 26 schools in the control group. Differences between group means may not equal the estimated differences because of rounding.

Source: Minnesota Department of Education 2008a; Missouri Department of Elementary and Secondary Education 2008a.

Results from multilevel models for benchmark impact analyses of primary outcomes

Table O2. Multilevel results for benchmark analysis: impact of Success in Sight on student reading outcome, 2009/10

Parameter	Estimate	Standard error	*t*-ratio	Degrees of freedom	*p*-value
Intercept	–0.42	0.02	–21.13	23	< .01***
Treatment	–0.01	0.03	–0.32	23	.75
Grade 4	–0.01	0.03	–0.26	8,151	.80
Grade 5	0.03	0.03	0.85	8,151	.40
Size	–0.01	0.01	–1.59	23	.13
Block 2	0.01	0.07	0.07	23	.95
Block 3	–0.04	0.14	–0.32	23	.75
Block 4	–0.10	0.13	–0.77	23	.45
Block 5	–0.18	0.09	–1.86	23	.08
Block 6	–0.15	0.08	–1.76	23	.09
Block 7	0.05	0.11	0.45	23	.66
Block 8	0.02	0.08	0.19	23	.85
Block 9	0.06	0.19	0.32	23	.75
Block 10	–0.06	0.14	–0.41	23	.68
Block 11	–0.09	0.16	–0.58	23	.57
Block 12	–0.05	0.10	–0.52	23	.61
Block 13	0.21	0.15	1.42	23	.17
Block 14	–0.06	0.15	–0.41	23	.69
Block 15	–0.01	0.13	–0.06	23	.95
Block 16	0.21	0.19	1.12	23	.27
Block 17	–0.09	0.14	–0.62	23	.54
Block 18	0.03	0.14	0.18	23	.86
Block 19	0.22	0.13	1.67	23	.11
Block 20	0.18	0.15	1.15	23	.26
Block 21	0.10	0.15	0.64	23	.53
Block 22	0.03	0.20	0.15	23	.89
Block 23	–0.08	0.08	–1.03	23	.31
Block 24	–0.12	0.09	–1.36	23	.19
Block 25	0.09	0.10	0.91	23	.37
Block 26	0.07	0.15	0.48	23	.63
School mean baseline	0.81	0.13	6.02	23	< .01***
Random effects	**Standard deviation**	**Variance component**	**Chi-square**	**Degrees of freedom**	***p*-value**
Random error for student *i* in school *j*	0.94	0.88			
Random error term for school *j*	0.13	0.02	85.60	23	<.01***

***Significant at $p = .01$.

Source: Minnesota Department of Education 2010b; Missouri Department of Elementary and Secondary Education 2010b.

Table O3. Multilevel results for benchmark analysis: impact of Success in Sight on student mathematics outcome, 2009/10

Parameter	Estimate	Standard error	*t*-ratio	Degrees of freedom	*p*-value
Intercept	–0.42	0.02	–16.99	23	< .01***
Treatment	–0.06	0.04	–1.74	23	.10
Grade 4	–0.03	0.04	–0.83	8,182	.41
Grade 5	0.01	0.05	0.20	8,182	.84
Size	–0.01	0.01	–1.74	23	.10
Block 2	0.41	0.08	5.49	23	< .01***
Block 3	0.10	0.07	1.39	23	.18
Block 4	0.26	0.09	3.05	23	< .01***
Block 5	0.28	0.09	3.03	23	< .01***
Block 6	–0.13	0.07	–1.86	23	.08
Block 7	0.31	0.07	4.30	23	< .01***
Block 8	0.33	0.11	2.92	23	< .01***
Block 9	0.36	0.07	5.14	23	< .01***
Block 10	0.08	0.11	0.74	23	.47
Block 11	0.17	0.19	0.89	23	.38
Block 12	0.13	0.08	1.58	23	.13
Block 13	0.47	0.17	2.74	23	.01***
Block 14	0.17	0.22	0.78	23	.44
Block 15	0.24	0.13	1.89	23	.07
Block 16	0.50	0.18	2.77	23	.01***
Block 17	0.33	0.19	1.74	23	.10
Block 18	0.31	0.15	2.04	23	.05**
Block 19	0.50	0.12	4.09	23	< .01***
Block 20	0.55	0.18	3.11	23	< .01***
Block 21	0.40	0.14	2.77	23	.01***
Block 22	0.43	0.18	2.40	23	.03**
Block 23	0.21	0.04	4.74	23	< .01***
Block 24	0.08	0.06	1.44	23	.16
Block 25	0.35	0.13	2.77	23	.01***
Block 26	0.64	0.12	5.13	23	< .01***
School mean baseline	0.81	0.15	5.53	23	< .01***
Random effects	**Standard deviation**	**Variance component**	**Chi-square**	**Degrees of freedom**	***p*-value**
Random error for student *i* in school *j*	0.97	0.95			
Random error term for school *j*	0.16	0.03	113.33	23	<.01***

**Significant at $p = .05$; ***significant at $p = .01$.

Source: Minnesota Department of Education 2010b; Missouri Department of Elementary and Secondary Education 2010b.

Appendix P. Analytic models for sensitivity analyses of primary outcomes

This appendix presents the analytic models for this study's sensitivity analyses testing the robustness of the primary impact estimates.

Sensitivity test for impact analysis of primary outcomes: use of baseline achievement covariate

The analyses using the following model included all students with available baseline or posttest achievement data.

Level 1:

$$Y_{ij} = \beta_{0j} + r_{ij}$$

where Y_{ij} is the posttest reading or mathematics performance of student i in a particular school j, β_{0j} is the mean posttest performance of students in school j, and r_{ij} is the random error for student i in school j.

Level 2:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(TREATMENT)_j + \gamma_{02}(SIZE)_j + \gamma_{03}(BLOCK\ 2)_j \ldots + \gamma_{029}(BLOCK\ 26)_j + u_{0j}$$

where γ_{00} is the adjusted mean posttest performance for average-size schools in the control group, while controlling for assignment block; γ_{01} is the effect of being in the treatment or control group, which represents the treatment–control difference in adjusted mean school performance; γ_{02} is the regression coefficient for school size; γ_{03}–γ_{029} are the regression coefficients for the random assignment blocks; and u_{0j} is the random error term for school j.

Sensitivity test for impact analysis of primary outcomes: Student sample

The analyses using the following model included only students who remained in the same school throughout the study period. This sample of students is referred to as "stayers."

Level 1:

$$Y_{ij} = \beta_{0j} + r_{ij}$$

where Y_{ij} is the posttest reading or mathematics performance of student i in a particular school j, β_{0j} is the mean posttest performance of students in school j, and r_{ij} is the random error for student i in school j.

Level 2:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(TREATMENT)_j + \gamma_{02}(SIZE)_j + \gamma_{03}(BLOCK\ 2)_j \ldots + \gamma_{029}(BLOCK\ 26)_j + \gamma_{030}(PREACHIEVE)_j + u_{0j},$$

where γ_{00} is the adjusted mean posttest performance for average-size, average-performing schools in the control group, while controlling for assignment block; γ_{01} is the effect of being in the treatment or control group, which represents the treatment–control difference in adjusted mean school performance; γ_{02} is the regression coefficient for school size; γ_{03}–γ_{029} are the regression coefficients for the random assignment blocks; γ_{030} is the regression coefficient for school mean baseline achievement; and u_{0j} is the random error term for school j.

Sensitivity test for impact analysis of primary outcomes: separate models for each state

The analyses using the following model included all students with available baseline or posttest achievement data. Researchers ran separate models for each state. Level 1 was consistent between each state, but level 2 varied slightly because of different random assignment blocks used within each state.

Level 1:

$$Y_{ij} = \beta_{0j} + \beta_{1j}(GRADE\ 4)_{ij} + \beta_{2j}(GRADE\ 5)_{ij} + \underline{r_{ij}}$$

where Y_{ij} is the posttest reading or mathematics performance of student i in a particular school j, β_{0j} is the mean posttest performance of students in school j, β_{1j} is the coefficient for the fixed level 1 covariate for grade 4, β_{2j} is the coefficient for the fixed level 1 covariate for grade 5, and r_{ij} is the random error for student i in school j.

Level 2 equation for Minnesota:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(TREATMENT)_j + \gamma_{02}(SIZE)_j + \gamma_{03}(BLOCK\ 2)_j \ldots + \gamma_{013}(BLOCK\ 12)_j + \gamma_{014}(PREACHIEVE)_j + u_{0j},$$

$$\beta_{1j} = \gamma_{10},$$

$$\beta_{2j} = \gamma_{20}$$

where γ_{00} is the adjusted mean posttest performance for average-size, average-performing schools in the control group, while controlling for assignment block; γ_{01} is the effect of being in the treatment or control group, which represents the treatment–control difference in adjusted mean school performance; γ_{02} is the regression coefficient for school size; γ_{03}–γ_{013} are the regression coefficients for the random assignment blocks; γ_{014} is the regression coefficient for school mean baseline achievement; and u_{0j} is the random error term for school j.

Level 2 equation for Missouri:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(TREATMENT)_j + \gamma_{02}(SIZE)_j + \gamma_{03}(BLOCK\ 2)_j \ldots + \gamma_{015}(BLOCK\ 14)_j + \gamma_{016}(PREACHIEVE)_j + u_{0j},$$

$$\beta_{1j} = \gamma_{10},$$

$$\beta_{2j} = \gamma_{20}$$

where γ_{00} is the adjusted mean posttest performance for average-size, average-performing schools in the control group, while controlling for assignment block; γ_{01} is the effect of being in the treatment or control group, which represents the treatment-control difference in adjusted mean school performance; γ_{02} is the regression coefficient for school size; γ_{03}–γ_{015} are the regression coefficients for the random assignment blocks; γ_{016} is the regression coefficient for school mean baseline achievement; and u_{0j} is the random error term for school j.

Appendix Q. Supporting tables for sensitivity analyses for primary outcomes

This appendix provides supporting tables for the sensitivity analyses for primary outcomes.

Results from sensitivity test with no baseline achievement covariate

Table Q1. Results from sensitivity analysis estimating treatment and control group differences in mean student achievement outcomes, unadjusted for baseline, 2009/10

	Treatment			Control			Estimated difference				
Unadjusted posttest (*z*-score) measure	Mean	Standard deviation	Sample size	Mean	Standard deviation	Sample size	Value	Standard error	95 percent confidence interval	*p*-value	Effect size[a]
Reading	–0.42	1.03	4,403	–0.44	1.02	3,779	0.03	0.04	–0.05–0.11	.52	0.03
Math	–0.48	1.10	4,413	–0.43	1.09	3,800	–0.05	0.05	–0.15–0.05	.29	–0.05

Note: Results are from multilevel models that account for the nesting of students in schools. Analyses included all 26 schools in the treatment group and all 26 schools in the control group. Differences between group means may not equal the estimated differences because of rounding.

a. Calculated by dividing the estimated difference in means by the control group standard deviation.

Source: Minnesota Department of Education 2010b; Missouri Department of Elementary and Secondary Education 2010b.

Results from multilevel models for sensitivity analyses for primary outcome of student achievement

Table Q2. Multilevel results for sensitivity analysis with no baseline achievement covariate: impact of Success in Sight on student reading outcome, 2009/10

Parameter	Estimate	Standard error	*t*-ratio	Degrees of freedom	*p*-value
Intercept	–0.44	0.03	–15.69	24	< .01***
Treatment	0.03	0.04	0.66	24	.52
Grade 4	–0.01	0.03	–0.27	8,152	.79
Grade 5	0.03	0.03	0.84	8,152	.40
Size	–0.01	0.01	–1.26	24	.22
Block 2	0.02	0.08	0.19	24	.85
Block 3	0.10	0.11	0.93	24	.36
Block 4	0.32	0.08	3.79	24	< .01***
Block 5	–0.14	0.10	–1.41	24	.17
Block 6	–0.21	0.10	–2.09	24	.047**
Block 7	0.32	0.17	1.87	24	.07
Block 8	0.04	0.09	0.49	24	.63
Block 9	0.66	0.24	2.74	24	.01***
Block 10	–0.31	0.11	–2.90	24	< .01***
Block 11	0.30	0.31	0.98	24	.34
Block 12	0.24	0.14	1.71	24	.10
Block 13	0.98	0.10	10.16	24	< .01***
Block 14	0.65	0.10	6.48	24	< .01***
Block 15	0.59	0.09	6.49	24	< .01***
Block 16	1.23	0.10	12.01	24	< .01***
Block 17	0.57	0.10	5.52	24	< .01***
Block 18	0.71	0.10	7.12	24	< .01***
Block 19	0.84	0.08	10.17	24	< .01***
Block 20	0.77	0.10	7.83	24	< .01***
Block 21	0.89	0.11	8.38	24	< .01***
Block 22	0.11	0.18	0.59	24	.56
Block 23	0.01	0.09	0.15	24	.88
Block 24	0.15	0.09	1.75	24	.09
Block 25	0.30	0.12	2.59	24	.02**
Block 26	0.77	0.10	7.38	24	< .01***
Random effects	**Standard deviation**	**Variance component**	**Chi-square**	**Degrees of freedom**	***p*-value**
Random error for student *i* in school *j*	0.94	0.88			
Random error term for school *j*	0.19	0.04	173.40	24	< .01***

**Significant at $p = .05$; ***significant at $p = .01$.

Source: Minnesota Department of Education 2010b; Missouri Department of Elementary and Secondary Education 2010b.

Table Q3. Multilevel results for sensitivity analysis with no baseline achievement covariate: impact of Success in Sight on student mathematics outcome, 2009/10

Parameter	Estimate	Standard error	*t*-ratio	Degrees of freedom	*p*-value
Intercept	–0.43	0.03	–13.61	24	< .01***
Treatment	–0.05	0.05	–1.08	24	.29
Grade 4	–0.03	0.04	–0.82	8,183	.41
Grade 5	0.01	0.05	0.21	8,183	.83
Size	–0.01	0.01	–1.01	24	.32
Block 2	0.33	0.12	2.72	24	.01***
Block 3	0.42	0.14	2.96	24	< .01***
Block 4	0.66	0.14	4.67	24	< .01***
Block 5	0.27	0.15	1.82	24	.08
Block 6	–0.08	0.14	–0.56	24	.58
Block 7	0.59	0.12	4.88	24	< .01***
Block 8	0.19	0.12	1.58	24	.13
Block 9	0.63	0.14	4.38	24	< .01***
Block 10	–0.38	0.13	–2.90	24	< .01***
Block 11	0.48	0.34	1.40	24	.17
Block 12	0.41	0.16	2.58	24	.02**
Block 13	1.31	0.16	8.29	24	< .01***
Block 14	0.96	0.19	5.21	24	< .01***
Block 15	0.88	0.13	6.95	24	< .01***
Block 16	1.48	0.14	10.94	24	< .01***
Block 17	0.96	0.15	6.43	24	< .01***
Block 18	0.95	0.18	5.22	24	< .01***
Block 19	1.11	0.12	9.10	24	< .01***
Block 20	1.15	0.21	5.61	24	< .01***
Block 21	1.13	0.15	7.40	24	<.01***
Block 22	0.34	0.16	2.13	24	.04**
Block 23	0.14	0.11	1.25	24	.22
Block 24	0.31	0.13	2.47	24	.02**
Block 25	0.52	0.21	2.49	24	.02**
Block 26	1.26	0.16	7.80	24	< .01***

Random effects	Standard deviation	Variance component	Chi-square	Degrees of freedom	*p*-value
Random error for student *i* in school *j*	0.97	0.95			
Random error term for school *j*	0.22	0.05	209.06	24	< .01***

**Significant at $p = .05$; ***significant at $p = .01$.

Source: Minnesota Department of Education 2010b; Missouri Department of Elementary and Secondary Education 2010b.

Table Q4. Multilevel results for sensitivity analysis with student stayer sample: Impact of Success in Sight on student reading outcome, 2009/10

Parameter	Estimate	Standard error	*t*-ratio	Degrees of freedom	*p*-Value
Intercept	–0.30	0.02	–12.95	23	< .01***
Treatment	–0.06	0.03	–1.72	23	.10
Size	–0.01	0.01	–0.21	23	.84
Block 2	0.11	0.11	0.98	23	.34
Block 3	–0.02	0.07	–0.27	23	.79
Block 4	–0.03	0.09	–0.31	23	.76
Block 5	–0.13	0.10	–1.38	23	.18
Block 6	–0.21	0.12	–1.72	23	.10
Block 7	0.17	0.14	1.14	23	.27
Block 8	–0.18	0.08	–2.20	23	.04**
Block 9	0.10	0.14	0.71	23	.48
Block 10	–0.19	0.14	–1.34	23	.19
Block 11	–0.02	0.12	–0.17	23	.87
Block 12	–0.19	0.09	–2.01	23	.06
Block 13	0.41	0.16	2.54	23	.02**
Block 14	–0.02	0.12	–0.17	23	.87
Block 15	0.15	0.10	1.61	23	.12
Block 16	0.37	0.12	3.00	23	< .01***
Block 17	0.07	0.14	0.47	23	.64
Block 18	0.13	0.16	0.86	23	.40
Block 19	0.13	0.11	1.22	23	.23
Block 20	0.19	0.10	1.87	23	.07
Block 21	0.27	0.12	2.19	23	.04**
Block 22	0.02	0.19	0.13	23	.90
Block 23	–0.01	0.10	–0.11	23	.91
Block 24	0.09	0.17	0.51	23	.62
Block 25	–0.02	0.10	–0.18	23	.86
Block 26	0.26	0.10	2.61	23	.02**
School mean baseline	0.64	0.08	8.23	23	< .01***

Random effects	Standard deviation	Variance component	Chi-square	Degrees of freedom	*p*-value
Random error for student *i* in school *j*	0.91	0.84			
Random error term for school *j*	0.03	0.01	22.41	23	> .50

**Significant at $p = .05$; ***significant at $p = .01$.

Source: Minnesota Department of Education 2010b; Missouri Department of Elementary and Secondary Education 2010b.

Table Q5. Multilevel results for sensitivity analysis with student stayer sample: Impact of Success in Sight on student mathematics outcome, 2007/08 and 2009/10

Parameter	Estimate	Standard error	*t*-ratio	Degrees of freedom	*p*-value
Intercept	–0.29	0.03	–10.44	23	< .01***
Treatment	–0.11	0.04	–2.60	23	.02**
Size	–0.01	0.01	–1.35	23	.19
Block 2	0.67	0.14	4.91	23	< .01***
Block 3	–0.03	0.05	–0.63	23	.53
Block 4	0.47	0.07	6.60	23	< .01***
Block 5	0.43	0.11	3.99	23	< .01***
Block 6	0.01	0.06	0.17	23	.87
Block 7	0.18	0.22	0.82	23	.42
Block 8	–0.06	0.07	–0.84	23	.41
Block 9	0.13	0.07	1.85	23	.08
Block 10	–0.45	0.15	-3.02	23	< .01***
Block 11	0.25	0.17	1.52	23	.14
Block 12	0.03	0.10	0.34	23	.74
Block 13	0.51	0.15	3.39	23	< .01***
Block 14	0.24	0.25	0.98	23	.34
Block 15	0.25	0.08	3.07	23	< .01***
Block 16	0.50	0.13	3.81	23	< .01***
Block 17	0.19	0.10	1.86	23	.08
Block 18	0.12	0.11	1.11	23	.28
Block 19	0.20	0.13	1.54	23	.14
Block 20	0.37	0.07	5.71	23	< .01***
Block 21	0.58	0.20	2.92	23	< .01***
Block 22	0.43	0.09	4.77	23	< .01***
Block 23	–0.16	0.10	–1.54	23	.14
Block 24	0.11	0.07	1.50	23	.15
Block 25	0.18	0.04	4.84	23	< .01***
Block 26	0.46	0.09	5.09	23	< .01***
School mean baseline	0.70	0.11	6.38	23	< .01***
Random effects	**Standard deviation**	**Variance component**	**Chi-square**	**Degrees of freedom**	***p*-value**
Random error for student *i* in school *j*	0.92	0.85			
Random error term for school *j*	0.13	0.02	35.65	23	.045**

**Significant at $p = .05$; ***significant at $p = .01$.

Source: Minnesota Department of Education 2008a, 2010b; Missouri Department of Elementary and Secondary Education 2008a, 2010b.

Table Q6. Benjamini-Hochberg correction for multiple comparisons for sensitivity analysis with student stayer sample: impact of Success in Sight on student mathematics outcome, 2009/10

Outcome	Clustering corrected *p*-value	*p*-value rank	Benjamini-Hochberg correction calculation *p*-value[a]	Is the clustering corrected *p*-value less than or equal to the Benjamini-Hochberg corrected *p*-value?	Statistical significance after Benjamini-Hochberg correction
Mathematics achievement	.02	1	.03	Yes	Significant

a. The Benjamini-Hochberg correction calculation was calculated by multiplying the rank for the significant *p*-value (one) by the alpha level (.05) and dividing the result by the number of findings in the domain (two).

Table Q7. Multilevel results for the sensitivity analysis examining the impact of Success in Sight on student reading achievement in Minnesota, 2009/10

Parameter	Estimate	Standard error	*t*-ratio	Degrees of freedom	*p*-value
Intercept	3,595.76	15.65	229.84	9	< .01***
Treatment	–2.72	22.52	–0.12	9	.91
Grade 4	121.54	10.23	11.89	3,728	< .01***
Grade 5	240.62	10.43	23.07	3,728	< .01***
Size	–0.09	0.14	–0.63	9	.54
Block 2	2.07	52.78	0.04	9	.97
Block 3	–18.88	57.87	–0.33	9	.75
Block 4	–46.30	67.18	–0.69	9	.51
Block 5	–54.81	61.25	–0.90	9	.39
Block 6	–39.49	59.38	–0.67	9	.52
Block 7	–6.28	62.65	–0.10	9	.92
Block 8	6.12	55.73	0.11	9	.92
Block 9	–8.60	82.17	–0.11	9	.92
Block 10	–3.35	72.66	–0.05	9	.97
Block 11	–40.50	69.23	–0.59	9	.57
Block 12	–28.31	67.97	–0.42	9	.69
School mean baseline	0.95	0.26	3.63	9	< .01***
Random effects	**Standard deviation**	**Variance component**	**Chi-square**	**Degrees of freedom**	***p*-value**
Random error for student *i* in school *j*	257.49	66,298.73			
Random error term for school *j*	48.68	2,369.38	55.58	9	<.01***

***Significant at $p = .01$.

Note: Estimates for previous tables in Appendix Q were calculated from state achievement z-scores calculated across states. Estimates from Table Q7 were calculated by using scale scores from the Minnesota Comprehensive Assessment II.

Source: Minnesota Department of Education 2010b.

Table Q8. Multilevel results for the sensitivity analysis examining the impact of Success in Sight on student reading achievement in Missouri, 2009/10

Parameter	Estimate	Standard error	*t*-ratio	Degrees of freedom	*p*-value
Intercept	651.12	1.40	466.42	11	< .01***
Treatment	–0.05	2.00	–0.03	11	.98
Grade 4	22.25	1.31	16.99	4,418	< .01***
Grade 5	35.06	1.31	26.75	4,418	< .01***
Size	–0.01	0.01	–0.67	11	.52
Block 14	–11.16	5.29	–2.11	11	.06
Block 15	–12.37	5.37	–2.31	11	.04**
Block 16	5.24	5.98	0.88	11	.40
Block 17	–13.12	5.43	–2.42	11	.03**
Block 18	–9.07	5.20	–1.74	11	.11
Block 19	–3.48	5.77	–0.60	11	.56
Block 20	–5.64	6.12	–0.92	11	.38
Block 21	–3.67	5.12	–0.72	11	.49
Block 22	–21.51	10.24	–2.10	11	.06
Block 23	–25.89	9.92	–2.61	11	.03**
Block 24	–23.11	7.90	–2.93	11	.01***
Block 25	–16.38	8.93	–1.84	11	.09
Block 26	–6.89	5.11	–1.35	11	.21
School mean baseline	0.37	0.29	1.27	11	.23
Random effects	**Standard deviation**	**Variance component**	**Chi-square**	**Degrees of freedom**	***p*-value**
Random error for student *i* in school *j*	35.67	1,272.13			
Random error term for school *j*	3.92	15.39	31.37	11	< .01

**Significant at $p = .05$; ***significant at $p = .01$.

Note: Estimates for previous Tables Q1-Q6 were calculated from state achievement z-scores calculated across states. Estimates from Table Q8 were calculated by using scale scores from the Missouri Assessment Program.

Source: Missouri Department of Elementary and Secondary Education 2010b.

Table Q9. Multilevel results for the sensitivity analysis examining the impact of Success in Sight on student mathematics achievement in Minnesota, 2009/10

Parameter	Estimate	Standard error	*t*-ratio	Degrees of freedom	*p*-value
Intercept	3,622.99	11.12	325.76	9	< .01***
Treatment	–34.34	15.99	–2.15	9	.06
Grade 4	72.48	8.72	8.31	3,762	< .01***
Grade 5	198.13	8.90	22.27	3,762	< .01***
Size	0.01	0.10	0.11	9	.92
Block 2	100.04	37.85	2.64	9	.03**
Block 3	35.09	44.39	0.79	9	.45
Block 4	63.51	47.22	1.35	9	.21
Block 5	82.69	43.86	1.89	9	.09
Block 6	–11.32	42.22	–0.27	9	.80
Block 7	71.39	44.64	1.60	9	.14
Block 8	90.93	41.44	2.20	9	.06
Block 9	100.32	47.34	2.12	9	.06
Block 10	58.42	59.79	0.98	9	.35
Block 11	53.52	46.74	1.15	9	.28
Block 12	44.62	48.17	0.93	9	.38
School mean baseline	0.90	0.24	3.75	9	< .01***
Random effects	**Standard deviation**	**Variance component**	**Chi-square**	**Degrees of freedom**	***p*-value**
Random error for student *i* in school *j*	220.72	48,718.54			
Random error term for school *j*	33.08	1094.02	41.15	9	< .01***

**Significant at $p = .05$; ***significant at $p = .01$.
Source: Minnesota Department of Education 2010b.

Table Q10. Multilevel results for the sensitivity analysis examining the impact of Success in Sight on student mathematics achievement in Missouri, 2009/10

Parameter	Estimate	Standard error	*t*-ratio	Degrees of freedom	*p*-value
Intercept	640.12	2.11	303.87	11	<.01***
Treatment	–0.15	3.02	–0.05	11	.96
Grade 4	23.80	1.36	17.49	4,415	<.01***
Grade 5	40.92	1.36	30.06	4,415	<.01***
Size	–0.02	0.02	–1.21	11	.25
Block 14	–12.06	7.95	–1.52	11	.16
Block 15	–10.62	8.31	–1.28	11	.23
Block 16	4.14	8.23	0.50	11	.63
Block 17	-6.93	8.89	–0.78	11	.45
Block 18	–7.86	8.37	–0.94	11	.37
Block 19	2.29	9.56	0.24	11	.82
Block 20	4.45	10.00	0.45	11	.67
Block 21	–1.76	8.14	–0.22	11	.83
Block 22	–5.52	18.59	–0.30	11	.77
Block 23	–13.87	18.07	–0.77	11	.46
Block 24	–17.73	13.30	–1.33	11	.21
Block 25	–6.24	14.82	–0.42	11	.68
Block 26	5.48	8.62	0.64	11	.54
School mean baseline	0.68	0.38	1.80	11	.10
Random effects	**Standard deviation**	**Variance component**	**Chi-square**	**Degrees of freedom**	***p*-value**
Random error for student *i* in school *j*	37.03	1,371.38			
Random error term for school *j*	7.00	49.07	68.98	11	<.01***

**Significant at $p = .05$; ***significant at $p = .01$.
Source: Missouri Department of Elementary and Secondary Education 2010b.

Table Q11. Weighted mean effect of the impact of Success in Sight on student achievement in reading and mathematics, 2009/10

Outcome measure	Minnesota effect size[a]	Missouri effect size[a]	Weighted mean effect[b]	95 percent confidence interval	*p*-value
Posttest reading scale score	–0.01	–0.01	–0.01	–0.55–0.54	.98
Posttest mathematics scale score	–0.14	–0.01	–0.07	–0.61–0.48	.82

a. Calculated by dividing the estimated difference in means by the control group standard deviation.
b. Calculated using CMA software by calculating separate effect size estimates for each state, weighting the separate effects, and combining the weighted effects by computing the weighted mean effect using the new standard error.
Source: Minnesota Department of Education 2010b; Missouri Department of Elementary and Secondary Education 2010b.

Appendix R. Supporting tables for impact analyses of secondary outcomes

Baseline means, standard errors, and effect sizes for impact analyses of secondary outcomes

Researchers conducted multilevel modeling to examine the difference between baseline treatment and control group means. The results reveal no statistically significant differences between treatment and control groups on their baseline mean capacity for school improvement scores (tables R1–R5).

Table R1. Baseline means, standard errors, and effect sizes for treatment and control group capacity for school improvement outcomes, 2008

	Treatment			Control			Estimated difference				
Baseline measure	Mean	Standard deviation	Sample size	Mean	Standard deviation	Sample size	Value	Standard error	95 percent confidence interval	*p*-value	Effect size
Data-based decisionmaking	4.43	0.53	815	4.45	0.54	701	–0.02	0.06	–0.14–0.10	.78	–0.04
Purposeful community	3.32	0.65	815	3.34	0.62	701	–0.02	0.07	–0.15–0.12	.75	–0.03
Shared leadership	3.81	0.83	815	3.90	0.83	701	–0.09	0.14	–0.36–0.18	.52	–0.11

Note: Results are from multilevel models that account for the nesting of teachers in schools. Analyses included 26 schools in the treatment group and 26 schools in the control group. Differences between group means may not equal the estimated differences because of rounding.
a. Calculated by dividing the estimated difference in means by the control group standard deviation.
Source: 2008 teacher survey.

Results from multilevel models for impact analyses of secondary outcomes

Table R2. Multilevel results for the impact of Success in Sight on capacity for school improvement in data-based decisionmaking, 2010

Parameter	Estimate	Standard error	*t*-ratio	Degrees of freedom	*p*-value
Intercept	4.49	0.02	294.36	23	<.01***
Treatment	0.03	0.02	1.56	23	.13
Size	–0.0003	0.0001	–3.43	23	<.01***
Block 2	0.17	0.03	6.19	23	<.01***
Block 3	0.02	0.07	0.27	23	.79
Block 4	0.14	0.07	2.02	23	.06
Block 5	–0.05	0.11	–0.45	23	.66
Block 6	0.19	0.06	3.08	23	.01***
Block 7	0.24	0.04	6.53	23	<.01***
Block 8	0.18	0.03	5.65	23	<.01***
Block 9	0.04	0.07	0.55	23	.59
Block 10	–0.09	0.06	–1.65	23	.11
Block 11	–0.21	0.09	–2.25	23	.03**
Block 12	–0.006	0.06	–0.11	23	.91
Block 13	0.35	0.05	6.58	23	<.01***
Block 14	0.28	0.09	3.00	23	.01***
Block 15	0.20	0.06	3.51	23	<.01***
Block 16	0.18	0.05	3.75	23	<.01***
Block 17	0.23	0.07	3.14	23	.01***
Block 18	0.18	0.11	1.67	23	.11
Block 19	0.26	0.03	8.01	23	<.01***
Block 20	0.36	0.05	6.90	23	<.01***
Block 21	0.18	0.05	3.83	23	<.01***
Block 22	0.26	0.03	8.14	23	<.01***
Block 23	0.17	0.02	7.13	23	<.01***
Block 24	0.12	0.04	2.84	23	.01***
Block 25	0.24	0.04	6.45	23	<.01***
Block 26	0.22	0.05	4.08	23	<.01***
School mean baseline for data-based decisionmaking	0.22	0.10	2.14	23	.04**
Random effects	**Standard deviation**	**Variance component**	**Chi-square**	**Degrees of freedom**	***p*-value**
Random error for teacher *i* in school *j*	0.47	0.22			
Random error term for school *j*	0.06	0.004	33.42	23	.07

**Significant at $p = .05$; ***significant at $p = .01$.

Source: 2010 teacher survey.

Table R3. Multilevel results for the impact of Success in Sight on capacity for school improvement in purposeful community, 2010

Parameter	Estimate	Standard error	*t*-ratio	Degrees of freedom	*p*-value
Intercept	3.45	0.03	113.57	23	<.01***
Treatment	0.03	0.04	0.70	23	.49
Size	–0.01	0.01	–1.08	23	.29
Block 2	0.31	0.04	7.89	23	<.01***
Block 3	0.30	0.13	2.30	23	<.05**
Block 4	0.33	0.04	8.08	23	<.01***
Block 5	–0.01	0.16	–0.01	23	.99
Block 6	0.23	0.11	2.18	23	.04**
Block 7	0.44	0.15	2.97	23	<.01***
Block 8	0.32	0.03	10.70	23	<.01***
Block 9	0.50	0.21	2.45	23	<.05**
Block 10	–0.02	0.11	–0.20	23	.84
Block 11	0.16	0.06	2.66	23	<.05**
Block 12	0.05	0.05	0.93	23	.37
Block 13	0.62	0.11	5.69	23	<.01***
Block 14	0.89	0.15	5.94	23	<.01***
Block 15	0.67	0.09	7.09	23	<.01***
Block 16	0.72	0.14	5.21	23	<.01***
Block 17	0.51	0.13	3.93	23	<.01***
Block 18	0.57	0.20	2.86	23	<.01***
Block 19	0.70	0.07	9.93	23	<.01***
Block 20	0.70	0.07	10.56	23	<.01***
Block 21	0.38	0.06	6.21	23	<.01***
Block 22	0.38	0.11	3.43	23	<.01***
Block 23	0.56	0.03	18.43	23	<.01***
Block 24	0.22	0.05	4.59	23	<.01***
Block 25	0.49	0.22	2.21	23	<.05**
Block 26	0.40	0.20	1.96	23	.06
School mean baseline for purposeful community	0.40	0.17	2.35	23	<.05**

Random effects	Standard deviation	Variance component	Chi-square	Degrees of freedom	*p*-value
Random error for teacher *i* in school *j*	0.60	0.36			
Random error term for school *j*	0.18	0.03	81.93	23	<.01***

**Significant at $p = .05$; ***significant at $p = .01$.
Source: 2010 teacher survey.

Table R4. Multilevel results for the impact of Success in Sight on capacity for school improvement in shared leadership, 2010

Parameter	Estimate	Standard error	*t*-ratio	Degrees of freedom	*p*-value
Intercept	3.90	0.05	81.99	23	<.01***
Treatment	0.16	0.07	2.49	23	.02**
Size	–0.001	0.0003	–3.07	23	.01***
Block 2	0.53	0.12	4.30	23	<.01***
Block 3	0.24	0.19	1.25	23	.23
Block 4	0.38	0.25	1.52	23	.14
Block 5	0.16	0.46	0.34	23	.74
Block 6	0.38	0.14	2.66	23	.02**
Block 7	0.38	0.24	1.62	23	.12
Block 8	0.61	0.17	3.55	23	<.01***
Block 9	0.36	0.26	1.39	23	.18
Block 10	0.10	0.29	0.35	23	.73
Block 11	–0.16	0.27	–0.61	23	.55
Block 12	–0.14	0.15	–0.97	23	.34
Block 13	0.64	0.23	2.84	23	.01***
Block 14	0.74	0.19	3.82	23	<.01***
Block 15	0.71	0.18	3.94	23	<.01***
Block 16	0.70	0.19	3.57	23	<.01***
Block 17	0.62	0.20	3.14	23	<.01***
Block 18	0.42	0.23	1.84	23	.08
Block 19	0.75	0.15	4.91	23	<.01***
Block 20	0.89	0.17	5.14	23	<.01***
Block 21	0.20	0.18	1.09	23	.29
Block 22	0.42	0.14	2.96	23	.01***
Block 23	0.52	0.17	3.02	23	.01***
Block 24	0.41	0.16	2.58	23	.02**
Block 25	0.54	0.27	2.03	23	.05**
Block 26	0.24	0.35	0.70	23	.49
School mean baseline for shared leadership	0.31	0.11	2.78	23	.01***
Random effects	**Standard deviation**	**Variance component**	**Chi-square**	**Degrees of freedom**	***p*-value**
Random error for teacher *i* in school *j*	0.66	0.43			
Random error term for school *j*	0.33	0.11	182.33	23	<.01***

**Significant at $p = .05$; ***significant at $p = .01$.

Source: 2010 teacher survey.

Table R5. Benjamini-Hochberg correction for multiple comparisons for benchmark analyses of secondary outcomes: Impact of Success in Sight on shared leadership, 2009/10

Outcome	Clustering corrected *p*-value	*p*-value rank	Benjamini-Hochberg correction calculation *p*-value[a]	Is the clustering corrected *p*-value less than or equal to the Benjamini-Hochberg corrected *p*-value?	Statistical significance after Benjamini-Hochberg correction
Shared leadership	.02	1	<.02	No	Not significant

a. The Benjamini-Hochberg correction calculation was calculated by multiplying the rank for the significant *p*-value by the alpha level (.05) and dividing the result by the number of findings in the domain.

Appendix S. Analytic model for sensitivity analyses for secondary outcomes

This appendix presents the analytic model for this study's sensitivity analyses testing the robustness of the impact estimates of secondary outcomes.

Sensitivity test for impact analysis of secondary outcomes: use of baseline capacity for school improvement practice covariate

The analyses using the following model included all school staff participants who completed the 2008 or 2010 teacher survey.

Level 1:

$$Y_{ij} = \beta_{0j} + r_{ij}$$

where Y_{ij} is the posttest data-based decisionmaking, purposeful community, or shared leadership score of teacher i in a particular school j, β_{0j} is the mean posttest data-based decisionmaking, purposeful community, or shared leadership score of teachers in school j, r_{ij} is the random error for teacher i in school j.

Level 2:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(TREATMENT)_j + \gamma_{02}(SIZE)_j + \gamma_{03}(BLOCK\ 2)_j \ldots + \gamma_{029}(BLOCK\ 26)_j + u_{0j}$$

where γ_{00} is the adjusted mean posttest teacher-reported capacity for school improvement score for average-size control schools, while controlling for assignment block; γ_{01} is the effect of being in the treatment or control group and represents the treatment–control difference in adjusted mean teacher-reported capacity for school improvement (in data-based decisionmaking, purposeful community, or shared leadership); γ_{02} is the regression coefficient for school size; γ_{03}–γ_{029} are the regression coefficients for the random assignment blocks; and u_{0j} is the random error term for school j.

Appendix T. Supporting tables for sensitivity analyses for impact analyses of secondary outcomes

Results from sensitivity test with no baseline capacity for school improvement covariate

Table T1. Results from sensitivity analysis estimating treatment and control group differences in mean capacity for school improvement outcomes, unadjusted for baseline, 2010

	Treatment			Control			Estimated difference				
Unadjusted posttest measure	Mean	Standard deviation	Sample size	Mean	Standard deviation	Sample Size	Value	Standard error	95 percent confidence interval	*p*-value	Effect Size
Data-based decisionmaking	4.51	0.48	815	4.50	0.51	701	0.02	0.02	–0.02–0.06	.27	0.04
Purposeful community	3.47	0.66	815	3.46	0.62	701	0.02	0.04	–0.06–0.10	.63	0.03
Shared leadership	4.03	0.73	815	3.94	0.86	701	0.14	0.07	0.003–0.28	.05**	0.16

**Significant at $p = .05$.

Note: Results are from multilevel models that account for the nesting of teachers in schools. Analyses included 26 schools in the treatment group and 26 schools in the control group. Differences between group means may not equal the estimated differences because of rounding.

a. Calculated by dividing the estimated difference in means by the control group standard deviation.

Source: 2010 teacher survey.

Table T2. Benjamini-Hochberg correction for multiple comparisons for sensitivity analyses of secondary outcomes: Impact of Success in Sight on shared leadership, 2009/10

Outcome	Clustering corrected *p*-value	*p*-value rank	Benjamini-Hochberg correction calculation *p*-value[a]	Is the clustering corrected *p*-value less than or equal to the Benjamini-Hochberg corrected *p*-value?	Statistical significance after Benjamini-Hochberg correction
Shared leadership	.05	1	<.02	No	Not significant

a. The Benjamini-Hochberg correction calculation was calculated by multiplying the rank for the significant *p*-value by the alpha level (.05) and dividing the result by the number of findings in the domain.

Appendix U. Analytic model for exploratory analysis

Level 1:

$$Y_{ij} = \beta_{0j} + \beta_{1j}(GRADE\ 4)_{ij} + \beta_{2j}(GRADE\ 5)_{ij} + r_{ij}$$

where Y_{ij} is the posttest performance of student i in a particular school j, β_{0j} is the mean posttest performance of students in school j, β_{1j} is the coefficient for the fixed level 1 covariate for grade 4, β_{2j} is the coefficient for the fixed level 1 covariate for grade 5, and r_{ij} is the random error for student i in school j.

Level 2:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(PREACHIEVE)_j + \gamma_{02}(POSTDATA)_j + \gamma_{03}(POSTCOMMUNITY)_j + \gamma_{04}(POSTLEADERSHIP)_j + \gamma_{05}(SIZE)_j + \gamma_{06}(BLOCK\ 2)_j \ldots + \gamma_{032}(BLOCK\ 26)_j + u_{0j},$$

$$\beta_{1j} = \gamma_{10},$$

$$\beta_{2j} = \gamma_{20}$$

where γ_{00} is the estimated mean posttest student achievement (in reading or mathematics) when all other predictors are zero, γ_{01} is the regression coefficient for baseline school mean achievement (in reading or mathematics), γ_{02} is the regression coefficient for the posttest school mean score for data-based decisionmaking, γ_{03} is the regression coefficient for the posttest school mean score for purposeful community, γ_{04} is the regression coefficient for the posttest school mean score for shared leadership, γ_{05} is the regression coefficient for school size, γ_{06}–γ_{032} are the regression coefficients for the random assignment blocks, and u_{0j} is the random error term for school j.

Appendix V. Supporting tables for exploratory analysis

This appendix provides supporting tables for the exploratory analyses.

Table V1. Multilevel results for exploratory analysis examining the relationship of capacity for school improvement practices and student reading achievement, 2009/10

Parameter	Estimate	Standard error	*t*-Ratio	Degrees of freedom	*p*-value
Intercept	–0.40	0.61	–0.66	21	.52
Data-based decisionmaking	0.10	0.19	0.53	21	.60
Purposeful community	0.04	0.15	0.30	21	.77
Shared leadership	–0.16	0.07	–2.37	21	.03**
Grade 4	–0.01	0.03	–0.26	8149	.79
Grade 5	0.03	0.03	0.85	8149	.40
Size	–0.01	0.01	–2.47	21	.02**
Block 2	0.06	0.05	1.10	21	.29
Block 3	–0.03	0.13	–0.21	21	.84
Block 4	–0.07	0.12	–0.59	21	.56
Block 5	–0.17	0.05	–3.20	21	< .01***
Block 6	–0.12	0.05	–2.43	21	.02**
Block 7	0.05	0.10	0.52	21	.61
Block 8	0.08	0.05	1.54	21	.14
Block 9	0.10	0.17	0.58	21	.57
Block 10	–0.06	0.10	–0.56	21	.58
Block 11	–0.10	0.12	–0.82	21	.42
Block 12	–0.06	0.06	–0.94	21	.36
Block 13	0.30	0.13	2.27	21	.03**
Block 14	0.03	0.14	0.24	21	.82
Block 15	0.10	0.11	0.91	21	.37
Block 16	0.33	0.16	2.06	21	.05**
Block 17	0.01	0.12	0.07	21	.95
Block 18	0.09	0.11	0.77	21	.45
Block 19	0.33	0.11	2.92	21	< .01***
Block 20	0.28	0.14	2.01	21	.06
Block 21	0.13	0.13	1.05	21	.31
Block 22	0.06	0.19	0.33	21	.75
Block 23	–0.05	0.06	–0.90	21	.38
Block 24	–0.06	0.07	–0.97	21	.35
Block 25	0.14	0.10	1.35	21	.19
Block 26	0.10	0.12	0.88	21	.39
School mean baseline	0.77	0.13	5.83	21	< .01***
Random Effects	**Standard deviation**	**Variance component**	**Chi-square**	**Degrees of freedom**	***p*-value**
Random error for student *i* in school *j*	0.94	0.88			
Random error term for school *j*	0.13	0.02	78.19	21	<.01***

**Significant at $p = .05$; ***significant at $p = .01$.

Source: Minnesota Department of Education 2010b; Missouri Department of Elementary and Secondary Education 2010b; 2010 teacher survey.

Table V2. Multilevel results for exploratory analysis examining the relationship of capacity for school improvement practices and student mathematics achievement, 2009/10

Parameter	Estimate	Standard error	*t*-Ratio	Degrees of freedom	*p*-value
Intercept	1.68	1.06	1.58	21	.13
Data-based decisionmaking	–0.63	0.28	–2.21	21	.04**
Purposeful community	0.53	0.15	3.53	21	< .01***
Shared leadership	–0.28	0.09	–3.06	21	< .01***
Grade 4	–0.03	0.04	–0.83	8180	.41
Grade 5	0.01	0.05	0.21	8180	.84
Size	–0.01	0.01	–4.41	21	< .01***
Block 2	0.44	0.10	4.57	21	< .01***
Block 3	0.08	0.07	1.21	21	.24
Block 4	0.37	0.09	3.96	21	< .01***
Block 5	0.23	0.05	5.01	21	< .01***
Block 6	–0.04	0.07	–0.56	21	.58
Block 7	0.43	0.08	5.38	21	< .01***
Block 8	0.39	0.12	3.34	21	< .01***
Block 9	0.27	0.09	3.05	21	< .01***
Block 10	–0.19	0.09	–1.97	21	.06
Block 11	–0.03	0.15	–0.18	21	.86
Block 12	0.17	0.11	1.56	21	.13
Block 13	0.88	0.16	5.35	21	< .01***
Block 14	0.40	0.22	1.80	21	.09
Block 15	0.49	0.13	3.65	21	< .01***
Block 16	0.79	0.16	4.99	21	< .01***
Block 17	0.64	0.15	4.29	21	< .01***
Block 18	0.52	0.14	3.62	21	< .01***
Block 19	0.77	0.11	6.71	21	< .01***
Block 20	0.95	0.16	5.82	21	< .01***
Block 21	0.67	0.13	5.26	21	< .01***
Block 22	0.50	0.19	2.66	21	.02**
Block 23	0.14	0.07	2.07	21	.05**
Block 24	0.25	0.06	3.89	21	< .01***
Block 25	0.46	0.11	4.30	21	< .01***
Block 26	0.86	0.13	6.64	21	< .01***
School mean baseline	0.49	0.11	4.31	21	< .01***

Random Effects	Standard deviation	Variance component	Chi-square	Degrees of freedom	*p*-value
Random error for student *i* in school *j*	0.97	0.95			
Random error term for school *j*	0.15	0.02	91.75	21	< .01***

**Significant at $p = .05$; ***significant at $p = .01$.

Source: Minnesota Department of Education 2010b; Missouri Department of Elementary and Secondary Education 2010b; 2010 teacher survey.

References

Abbott, M.L. and Joireman, J. (2001). *The relationships among achievement, low income, and ethnicity across six groups of Washington state students*. Lynwood, WA: Washington School Research Center.

Adams, J. D. (1997). Creating critical mass to support change. In D. C. Van Eynde, J. C. Hoy, and D. C. Van Eynde (Eds.), *Organization development classics: the practice and theory of change—the best of the OD Practitioner* [Kindle version]. San Francisco, CA: Jossey-Bass Publishers.

Adelman, H. S., and Taylor, L. (2007). Systemic change for school improvement. *Journal of Educational and Psychological Consultation*, *17*(1), 55–77.

Albright, J. (2008). *Confirmatory factor analysis I*. The Trustees of Indiana University. Retrieved from http://www.indiana.edu/~statmath

American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education. Joint Technical Committee. (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

American Institutes for Research. (2005). *National AYP and identification database.* Retrieved from http://www.air.org/focus-area/education/index.cfm?fa=viewContent&content_id=860

Apthorp, H., Barley, Z., Englert, K., Lauer, P., and Van Buhler, B. (2005). *Final report: high needs schools—what does it take to beat the odds?* Aurora, CO: Mid-continent Research for Education and Learning.

Argyris, C. (1976). Leadership, learning, and changing the status quo. *Organizational Dynamics*, *4*(3), 29–43.

Argyris, C. and Schon, D. A. (1996). *Organizational learning II: theory, method, and practice.* Reading, MA: Addison-Wesley Publishing Company.

Bandura, A. (1997). *Self-efficacy: the exercise of control.* New York: W. H. Freeman.

Beckhard, R. (1969). *Organization development: strategies and models*. Reading, MA: Addison-Wesley Publishing Company.

Beckhard, R. and Pritchard, W. (1992). *Changing the essence: the art of creating and leading fundamental change in organizations* [Kindle version]. San Francisco, CA: Jossey-Bass Publishers.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a new and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, Series B, 57, 1289–1300.

Bernhardt, V. (2003). *Using data to improve student learning in elementary schools.* Larchmont, NY: Eye on Education, Inc.

Berry, B., Johnson, D., and Montgomery, D. (2005). The power of teacher leadership. *Educational Leadership, 62*(5), 56–60.

Bloom, H. S., Bos, J. M., and Lee, S.-W. (1999). Using cluster random assignment to measure program impacts: statistical implications for the evaluation of education programs. *Evaluation Review, 23*(4), 445–489.

Bloom, H. S., Richburg-Hayes, L., and Black, A. R. (2005). *Using covariates to improve precision: empirical guidance for studies that randomize schools to measure the impacts of educational interventions* (working paper). (ERIC Document Reproduction Service Number: ED486654). New York: Manpower Demonstration Research Corporation.

Borenstein, M., and Rothstein, H. (1999). *Comprehensive meta-analysis: a computer program for research synthesis.* Englewood, NJ: Biostat, Inc.

Borman, G., Hewes, G., Overman, L., and Brown, S. (2003). Comprehensive school reform and achievement: a meta-analysis. *Review of Educational Research, 73*(2), 125–230.

Brown, M. W., and Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen and J. S. Long (Eds.), *Testing structural equation models* (pp.136–162). Thousand Oaks, CA: Sage.

Bryk, A., S., Sebring, P. B., Allensworth, E., Luppescu, S., and Easton, J. Q. (2010). *Organizing schools for improvement: lessons from Chicago.* Chicago: University of Chicago Press.

Camburn, E., Rowan, B., and Taylor, J. (2003). Distributed leadership in schools: the case of elementary schools adopting comprehensive school reform models. *Educational Evaluation and Policy Analysis, 25*(4), 347–373.

Campbell, M. K., Elbourne, D. R., and Altman, D. G. (2004). CONSORT statement: extension to cluster randomised trials. *British Medical Journal, 328*, 702–708.

Carlson, D., Borman, G. D., and Robinson, M. (2010). *A multi-state district-level cluster randomized trial of the impact of data driven reform on reading and mathematics achievement.* Unpublished manuscript. University of Wisconsin: Madison.

Center for School Change (2010). *Leadership Academy for Minnesota Charter and Alternative Schools*. Retrieved from http://www.centerforschoolchange.org/projects/leadershipacademy/

Cicchinelli, L., Dean, C., Galvin, M., Goodwin, B., and Parsley, D. (2006). *Success in Sight: a comprehensive approach to school improvement.* Retrieved from http://www.mcrel.org/product/253/

Clune, W. (1998). *Toward a theory of systemic reform: the case of nine NSF statewide systemic initiatives.* Research Monograph No. 16. Madison, WI: National Institute for Science Education.

Collins, D. (1998). *Organizational change: Sociological perspectives*. London: Routledge.

Coursey, D. (2008). CFA: A practical intro. In K. Yang and G. Miller, *Handbook of research methods in public administration* (p. 569). Boca Raton, FL: Taylor and Francis.

Creemers, B. P. M. (1994). *The effective classroom.* New York: Cassell.

Datnow, A., Lasky, S.G., and Stringfield, S.C. (2005). Systemic integration for educational reform in racially and linguistically diverse contexts: a summary of the evidence. *Journal of Education for Students Placed at Risk, 10,* 445–453.

Dewey, J. (1938). *Experience and education*. New York: Simon & Schuster.

DiBella, A. J., and Nevis, E. C. (1998). *How organizations learn: an integrated strategy for building learning capability*. San Francisco, CA: Jossey-Bass, Inc.

DuFour, R. (2004). What is a "professional learning community"? *Educational Leadership, 61*(8), 6–11.

Duncan, A. (2009, August) *Robust data gives us the roadmap to reform.* Address to the Fourth Annual Institute of Education Sciences Research Conference, National Harbor, MD. Retrieved from http://www2.ed.gov/news/speeches/2009/06/06082009.pdf

Education Minnesota. (2010). *Professional learning communities*. Retrieved from http://www.educationminnesota.org/en/professional-development/tall/plcs.aspx

Ehlert, M., Podgursky, M., Koedel, C., and Kim, E.Y. (2009). Longitudinal Analysis of MAP Achievement Growth: Initial Findings. P-20 Educational Research and Data Center, University of Missouri, Columbia. Retrieved from http://p20.missouri.edu/documents/MAP_Achievement_Growth_Jan-09.pdf

Ellsworth, J. (2000). *Surviving change: a survey of educational change models*. (ERIC Document Reproduction Service No. ED443417). Syracuse, NY: ERIC Clearinghouse on Information and Technology, Syracuse University.

Freire, P. (1998). *Pedagogy of freedom: ethics, democracy, and civic courage*. Lanham, MD: Rowman & Littlefield Publishers, Inc.

Fullan, M. (1999). *Change forces: the sequel.* Philadelphia: Falmer Press.

Fullan, M. (2001). *Leading in a culture of change*. San Francisco: Jossey-Bass.

Fullan, M. (2002). Leadership and sustainability. *Principal Leadership, 3*(4), 14–17.

Fullan, M. (2007). *The new meaning of educational change* (4th ed.). New York: Teachers College Press.

Fullan, M. (2010). *Motion leadership: the skinny on becoming change savvy*. Thousand Oaks, CA: Corwin.

Fullan, M., and Stiegelbauer, S. (1991). *The new meaning of educational change*. New York, NY: Teachers College Press.

Glass, G. V., McGaw, B., and Smith, M. L. (1981). *Meta-analysis in social research.* Beverly Hills, CA: Sage.

Goddard, R. (2002). A theoretical and empirical analysis of the measurement of collective efficacy: the development of a short form. *Educational and Psychological Measurement*, *62*(1), 97–110.

Goddard, R. D., Hoy, W. K., and Hoy, A. (2000). Collective teacher efficacy: its meaning, measure, and effect on student achievement. *American Education Research Journal*, *37*(2), 479–507.

Goertz, M.E., Floden, R.E., O'Day, J. (1996). *Studies of Education: Systemic Reform.* Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement.

Goertz, M. E., Olah, L. N., and Riggan, M. (2009). *From testing to teaching: The use of interim assessments in classroom instruction.* Research Report 65. Philadelphia, Pa.: University of Pennsylvania, Consortium for Policy Research in Education.

Graham, J. W. (2009). Missing data analysis: making it work in the real world. *Annual Review of Psychology, 60,* 549–576.

Graham, J. W., Cumsille, P. E., and Elek-Fisk, E. (2003). Methods for handling missing data. In J. A. Schinka and W. F. Velicer (Eds.), *Research Methods in Psychology* (pp. 87–114). Volume 2 of the *Handbook of Psychology* (I. B. Weiner, Editor-in-Chief). New York: John Wiley & Sons.

Hall, G., and Hord, S. (1987). *Change in schools: facilitating the process*. Albany, NY: State University of New York Press.

Hallinger, P., and Heck, R. (2010). Collaborative leadership and school improvement: understanding the impact on school capacity and student learning. *School Leadership and Management, 30*(2), 95–110.

Hargreaves, A., Halász, G., and Pont, B. (2007). School leadership for systemic improvement in Finland: A case study report for the OECD activity improving school leadership. Retrieved from www.oecd.org/dataoecd/43/17/39928629.pdf

Havelock, R., and Zlotolow, S. (1995). *The change agent's guide* (2nd ed.). Englewood Cliffs, NJ: Educational Technology Publications.

Herman, R., Dawson, P., Dee, T., Greene, J., Maynard, R., Redding, S., and Darwin, M. (2008). *Turning around chronically low-performing schools: a practice guide* (NCEE #2008-4020). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://ies.ed.gov/ncee/wwc/pdf/practice_guides/Turnaround_pg_04181.pdf

HLM (Version 6.08) [Computer software]. Lincolnwood, IL: Scientific Software International, Inc.

Hollins, E. R., McIntyre, L. R., DeBose, C., Hollins, K. S., and Towner, A. (2004). Promoting a self-sustaining learning community: investigating an internal model for teacher development. *International Journal of Qualitative Studies in Education, 17*(2), 247–264.

Hoy, W.K., Smith, P.A., and Sweetland, S.R. (2002). A test of a model of school achievement in rural schools: the significance of collective efficacy. In W.K. Hoy and C. Miskel (Eds), *Theory and research in education administration: Vol. 1* (pp. 185–202). Greenwich, CT: Information Age Publishing.

Hulpia, H., Devos, G., and Rosseel, Y. (2009). The relationship between the perception of distributed leadership in secondary schools and teachers' and teacher leaders' job satisfaction and organizational commitment. *School Effectiveness and School Improvement*, *20*(3), 291–317.

Kim, J.J., and Crasco, L.M. (2003). *Best policies and practices in urban educational reform: a summary of empirical analysis focusing on student achievements and equity*. Paper presented at the Annual Meeting of the American Educational Research Association in Chicago, IL.

Kouzes, J. M., and Posner, B. Z. (1997). Leadership is everyone's business and other lessons from over a dozen years of leadership research. In D. F. Van Eynde, J. C. Hoy, and D. C. Van Eynde (Eds.), *Organization development classics: the practice and theory of change—the best of the OD Practitioner* [Kindle version]. San Francisco, CA: Jossey-Bass Publishers.

Krathwohl, D. R. (1993). *Methods of educational and social science research*. New York, NY: Longman.

Kutash, J., Nico, E., Gorin, E., Rahmatullah, S., and Tallant, K. (2010). School turnaround field guide. San Francisco, CA: FSG Social Impact Consultants. Retrieved from

http://www.fsg.org/tabid/191/ArticleId/88/Default.aspx?srpush=true

Leithwood, K., Louis, K., Anderson, S., and Wahlstrom, K. (2004). *Review of research. How leadership influences student learning*. New York: The Wallace Foundation.

Lippitt, G. L., and Lippitt, R. (1986). *The consulting process in action* (2nd ed.). San Francisco: Jossey-Bass/Pfeiffer.

Lippitt, R., Watson, J., and Westley, B. (1958). *The dynamics of planned change: a comparative study of principles and techniques*. New York: Harcourt, Brace and Company.

Loehlin, J. (2004). *Latent variables models.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Liu, X., Spybrook, J., Congdon, R., Martinez, A., and Raudenbush, S. (2006). *Optimal Design for Multi-Level and Longitudinal Research (Version 1.77)* [Computer software]. University of Michigan, Ann Arbor: Survey Research Center.

Louis, K. S., and Marks, H. M. (1998). Does professional community affect the classroom? Teachers' work and student experiences in restructuring schools. *American Journal of Education,* 106, 532–575.

Marzano, R. J. (2000). *A quantitative synthesis of research on school-level, teacher-level, and student-level variables related to academic achievement* [REL Deliverable 2000-05]. Aurora, CO: Mid-continent Research for Education and Learning.

Marzano, R. J. (2003). *What works in schools: translating research into action*. Alexandria, VA: Association for Supervision and Curriculum Development.

Marzano, R. J., Waters, T., and McNulty, B. A. (2005). *School leadership that works: from research to results*. Alexandria, VA: Association for Supervision and Curriculum Development.

Massell, D., Kirst, M., and Hoppe, M. (1997). *Persistence and change: standards-based reform in nine states* (CPRE Research Reports Series #37). New Brunswick, NJ: Consortium for Policy Research in Education.

May, H., Perez-Johnson, I., Haimson, J., Sattar, S., and Gleason, P. (2009). *Using state tests in education experiments: a discussion of the issues.* Princeton, NJ: Mathematica Policy Research, Inc.

Mclntire, T. (2005). *Data: maximize your mining*. Retrieved from http://www.techlearning.com/article/data-maximize-your-mining-part-one/45194

McKnight, P. E., McKnight, K. M., Sidani, S., and Figueredo, A. J. (2007). *Missing data: a gentle introduction.* New York: Guilford.

Mid-continent Research for Education and Learning. (2005). *Final report: high-needs schools—what does it take to beat the odds?* Denver, CO: Mid-continent Research for Education and Learning.

Mid-continent Research for Education and Learning. (2008). *Success in Sight: facilitator manual.* Denver, CO: Mid-continent Research for Education and Learning.

Mid-continent Research for Education and Learning. (n.d.). *Success in Sight: evidence of success.* Unpublished manuscript, Institutional Development, Denver, CO: Mid-continent Research for Education and Learning.

Minnesota Department of Education. (2008a). *MCA-II student reading and mathematics assessment scores and demographics* [Data file].

Minnesota Department of Education. (2008b). Procedures manual for Minnesota assessments 2007–2008 [Technical manual]. Roseville, MN: The Minnesota Department of Education.

Minnesota Department of Education. (2008c). State means and standard deviations for Reading and Mathematics: 2008 [Data file]. Retrieved from http://education.state.mn.us/MDE/Accountability_Programs/Assessment_and_Testing/Assessments/MCA/TechReports/index.html

Minnesota Department of Education. (2008d). *2007–2008 Accountability Data*. [Data file]. Retrieved from http://www.education.state.mn.us/mde/Data/Data_Downloads/Accountability_Data/NCLB_AYP/index.html

Minnesota Department of Education. (2008e). *2007–2008 MCA-II Yearbook* [Data file].

Minnesota Department of Education. (2008f). *2007–2008 Testing Schedule.* [Calendar file]. Retrieved from http://education.state.mn.us/mdeprod/groups/Assessment/documents/Schedule/008335.pdf

Minnesota Department of Education. (2010a). *MCA-II excel files* [Data files]. Retrieved from http://education.state.mn.us/MDE/Data/Data_Downloads/Accountability_Data/Assessment_MCA_II/MCA_II_Excel_files/index.html

Minnesota Department of Education. (2010b). *MCA-II student reading and mathematics assessment scores and demographics* [Data file].

Minnesota Department of Education. (2010c). *2006–2008 Not Making AYP* [Data files]. http://education.state.mn.us/WebsiteContent/ContentArchive.jsp?siteId=7&siteSection=Data%2FData+Downloads%2FAccountability+Data%2FAYP%3B

Minnesota Department of Education. (2010d). *2009–2010 Accountability Data.* [Data file]. Retrieved from http://www.education.state.mn.us/mde/Data/Data_Downloads/Accountability_Data/NCLB_AYP/index.html

Minnesota Department of Education. (2010e). *2009–2010 Testing Schedule*. [Calendar file]. Retrieved from http://education.state.mn.us/mdeprod/groups/Assessment/documents/Schedule/034354.pdf

Missouri Department of Elementary and Secondary Education Activity Catalog. (2010). *My Learning Plan*. Retrieved from https://www.mylearningplan.com/webreg/catalog.asp?D=11005&H=1&P=66725

Missouri Department of Elementary and Secondary Education. (2008a). *MAP student reading and mathematics assessment scores and demographics* [Data file].

Missouri Department of Elementary and Secondary Education. (2008b). *MAP Technical Report 2008.* [Technical manual]. Retrieved from http://dese.mo.gov/divimprove/assess/tech/

Missouri Department of Elementary and Secondary Education. (2009a). *2006-2009 AYP Summary* [Data file].

Missouri Department of Elementary and Secondary Education. (2009b). "State Officials Pleased with Steady Gains on MAP Tests: End-of-Course Tests Debut; Fewer Schools Meet AYP Targets." News Release, 43(59), August 12. Retrieved from http://dese.mo.gov/news/2009/mapeoc.htm#TitleI

Missouri Department of Elementary and Secondary Education. (2010a). *Annual Reporting of School District Data: MAP and End-of-Course (EOC)* [data files]. Retrieved from http://dese.mo.gov/schooldata/ftpdata.html

Missouri Department of Elementary and Secondary Education. (2010b). *MAP student reading and mathematics assessment scores and demographics* [Data file].

Missouri Department of Elementary and Secondary Education. (2010c). *MAP Technical Report 2010*. [Technical report]. Retrieved from http://dese.mo.gov/divimprove/assess/tech/

Missouri Department of Elementary and Secondary Education. (2010d). *School Data.* [Data file]. Retrieved from http://dese.mo.gov/schooldata/school_data.html

Missouri Department of Elementary and Secondary Education Professional Learning Communities. (2009). School Improvement Initiatives: Professional Learning Communities. Retrieved from http://dese.mo.gov/divteachqual/sii/prolearning/

Missouri Department of Elementary and Secondary Education Three-Tiered Models of Intervention. (2011). Three-Tiered Models of Intervention and Evidence Based Practice(RtI). Retrieved from http://dese.mo.gov/3tieredmodels/

Mourshed, M., Chijioke, C., and Barber, M. (2010). How the world's most improved school systems keep getting better. New York: McKinsey & Company. Retrieved from http://www.mckinsey.com/clientservice/Social_Sector/our_practices/Education/Knowledge_Highlights/How%20School%20Systems%20Get%20Better.aspx

National Association of State Directors of Special Education. (2005). *Response to intervention: policy considerations and implementations.* Alexandria, VA: National Association of State Directors of Special Education.

PALS (2007). Phonological Awareness Literacy Screening. Retrieved from http://pals.virginia.edu/

Phillips, J. (2003). Powerful learning: creating learning communities in urban school reform. *Journal of Curriculum and Supervision, 18*(3), 240–258.

Printy, S. M., and Marks, H. M. (2006). Shared leadership for teacher and student learning. *Theory into Practice*, *45*(2), 125–132.

Protheroe, N. (2001). Improving teaching and learning with data-based decisions: asking the right questions and acting on answers. *ERS Spectrum, 19*(3), 4–9.

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

Raudenbush, S. W., and Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods.* Thousand Oaks, CA: Sage Publications, Inc.

Reynolds, D., Teddlie, C., Hopkins, D., and Stringfield, S. (2000). Linking school effectiveness and school improvement. In C. Teddlie and D. Reynolds (Eds). *The international handbook of school effectiveness research* (pp. 206–231). New York: Falmer Press.

Rhoton, J. (2001). School science reform: An overview and implications for the secondary school principal. *NASSP Bulletin, 85*, 10–23.

Rothwell, W. J., Stavros, J. M., and Sullivan, R. L. (2010). Organization development and change. In W. J. Rothwell, J. M. Stavros, R. L. Sullivan, and A. Sullivan (Eds.), *Practicing organization development: a guide for leading change* [Kindle version] (3rd ed.). San Francisco, CA: Pfeiffer.

Rubin, D. B. (1987). *Multiple imputation for non-response in surveys*. New York, NY: J. Wiley & Sons.

Rubin, D. (1996). Multiple imputation after 18+ years (with discussion). *Journal of the American Statistical Association*, *91*, 473–518.

Saphier, J., and King, M. (1985). Good seeds grow in strong cultures. *Educational Leadership, 42(6)*, 67–73.

Sashkin, M., and Egermeier, J. (1993). *School change models and processes: a review and synthesis of the research and practice*. Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement, Programs for the Improvement of Practice.

Schochet, P. Z. (2005). *Statistical power for random assignment evaluations of education programs* (No. 6046-310). (ERIC Document Reproduction Service No. ED489855). Princeton, NJ: Mathematica Policy Research Inc.

Schochet, P. Z. (2008a). *Technical methods report: guidelines for multiple testing in impact evaluations* (NCEE 2008-4018). Washington, D.C.: U.S. Department of Education, National Center for Education and Regional Assistance.

Schochet, P. Z. (2008b). Statistical power for random assignment evaluations of education programs. *Journal of Educational and Behavioral Statistics, 33,* 62–87.

Senge, P. (1990). *The fifth discipline: the art & practice of the learning organization*. New York, NY: Doubleday.

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston, MA: Houghton Mifflin Company.

Shadish, W. R., and Haddock, C. K. (1994). Combining estimates of effect size. In H. Cooper and L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 261–281). New York: Russell Sage Foundation.

Spybrook, J., Raudenbush, S. W., Congdon, R., and Martinez, A. (2009). *Optimal Design for longitudinal and multilevel research: documentation for the Optimal Design software V.2.0*. Retrieved from www.wtgrantfoundation.org

Steiger, J. H. (1989). *EzPath: casual modeling*. Evanston, IL: SYSTAT Inc.

Strahan, D. (2003). Promoting a collaborative professional culture in three elementary schools that have beaten the odds. *Elementary School Journal, 104*(2), 127–146.

Sun, H., Creemers, B. P. M., and de Jong, R. (2007). Contextual factors and effective school improvement. *School Effectiveness and School Improvement, 18*(1), 93–122.

Supovitz, J.A. (2002). Developing communities of instructional practice. *Teachers College Record, 104*(8), 1591–1626.

Supovitz, J.A., and Christman, J.B. (2003, November). Developing communities of instructional practice: Lessons for Cincinnati and Philadelphia. *CPRE Policy Briefs*. University of Pennsylvania, 1–9.

Supovitz, J.A., and Taylor, B.S. (2005). Systemic education evaluation: evaluating the impact of systemwide reform education. *American Journal of Evaluation, 26*, 204–230.

Teddlie, C., and Reynolds, D. (2000). *The international handbook of school effectiveness research.* New York: Falmer Press.

U.S. Department of Education. (2008a). Mapping America's Educational Progress 2008. Retrieved from http://www2.ed.gov/nclb/accountability/results/progress/nation.html

U.S. Department of Education. (2008b). *Reading First Implementation Evaluation Final Report*. Retrieved from http://www2.ed.gov/rschstat/eval/other/readingfirst-final/index.html

U.S. Department of Education, National Center for Education Statistics. (2008, 2010). *Common Core of Data*. Retrieved from http://nces.ed.gov/ccd/bat/

Van Buuren, S., and Groothuis-Oudshoorn, K. (Forthcoming). MICE: Multivariate Imputation by Chained Equations. *Journal of Statistical Software*.

Van Buuren, S., and Oudshoorn, G. (1999). *Flexible multivariate imputation by MICE* (TNO report: PG/VGZ/99.054). Retrieved from http://web.inter.nl.net/users/S.van.Buuren/mi/docs/rapport99054.pdf

Warrick, D. D. (2005). Launch: assessment and action planning. In W. J. Rothwell and R. L. Sullivan (Eds.), *Practicing organization development: a guide for consultants* (2nd ed.). San Francisco, CA: John Wiley & Sons, Inc.

Waters, T., and Cameron, G. (2007). *The balanced leadership framework: connecting vision with action.* Denver, CO: Mid-continent Research for Education and Learning.

Wayman, J., Stringfield, S., and Yakimowski, M. (2004). *Software enabling school improvement through analysis of student data.* Baltimore, MD: Center for Research on the Education of Students Placed At Risk. Retrieved from http://www.csos.jhu.edu/crespar/techReports/Report67.pdf

What Works Clearinghouse. (2008). *What Works Clearinghouse procedures and standards handbook* (Version 2.0). Retrieved from http://ies.ed.gov/ncee/wwc/references/idocviewer/doc.aspx?docid=19&tocid=1

Williams, V. S. L., Jones, L. V., and Tukey, J. W. (1999). Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement. *Journal of Educational and Behavioral Statistics, 24*(1), 42–69.

Witte, J., and Walsh, D. (1990). A systematic test of the effective schools model. *Educational Evaluation and Policy Analysis*, *(12)*2, 188–212.

Wolf, M.A. (2007). Systemic school reform: a guiding hand. *T.H.E. Journal, 34*, 12–25.


DEPARTMENT OF EDUCATION
UNITED STATES OF AMERICA