NCEE 2011–4005

U.S. DEPARTMENT OF EDUCATION

Classroom Assessment for Student Learning: Impact on Elementary School Mathematics in the Central Region

Final Report

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

REL CENTRAL
Regional Educational Laboratory

Classroom Assessment for Student Learning: Impact on Elementary School Mathematics in the Central Region

April 2011

Authors

Bruce Randel, Ph.D., Mid-continent Research for Education and Learning

Andrea D. Beesley, Ph.D., Mid-continent Research for Education and Learning

Helen Apthorp, Ph.D., Mid-continent Research for Education and Learning

Tedra F. Clark, Ph.D., Mid-continent Research for Education and Learning

Xin Wang, Ph.D., Mid-continent Research for Education and Learning

Louis F. Cicchinelli, Ph.D., Mid-continent Research for Education and Learning

Jean M. Williams, Ph.D., Mid-continent Research for Education and Learning

NCEE 2011-4005
U.S. Department of Education

Institute of Education Sciences

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

April 2011

This report was prepared for the National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, 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.

Disclosure of potential conflict of interest

None of the authors or other staff from REL Central involved in the study has financial interests that could be affected by the content of this report.

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

Acknowledgments

This study and its report were made possible by a collaborative effort of school districts, schools, teachers, students, researchers, reviewers, technical advisors, and editors. The authors thank the following individuals for their participation and assistance.

First, they thank all the school districts, schools, principals, teachers, and students who participated in the study for their commitment to the project, their time and effort to provide data, and their willingness to support the efforts to provide rigorous evidence of effectiveness to the field of education. This study would not have been possible without their involvement.

At REL Central, the authors thank Andrew Newman for site recruiting, site relations, instrument development, and data collection; Trudy Cherasaro for database management and site relations; Kirsten Miller for editorial review; Jessica Rainey and Carol Loredo for administrative support; Jessica Allen for data analysis; and Dawn Fries, Michael Phillips, and Mya Martin-Glenn for site recruiting, site relations, and instrument development.

The authors also thank those individuals who provided technical consultation and advice: the members of our Technical Working Group—Geoffrey Borman, Sue Brookhart, Barbara Plake, and Bob St. Pierre—and the reviewers at the Analytical and Technical Support contract administered by Mathematica Policy Research, Inc.

In addition, the authors thank Jacque Williams for site relations, and Kellie Randall and Tanya Batzel for scoring teacher work samples.

Table of contents

List of boxes

List of figures

List of tables

Executive summary

Under the No Child Left Behind Act of 2001 (NCLB), states, districts, and schools are required to ensure that all students meet the same high standards in mathematics and reading by the end of the 2013/14 school year (No Child Left Behind Act 2002). In working toward this goal, states, districts, and schools are increasingly in need of rigorous, high-quality research on efficient and effective interventions to improve student achievement.

This study was conducted by the Central Region Educational Laboratory (REL Central) administered by Mid-continent Research for Education and Learning to provide educators and policymakers with rigorous evidence about the potential of Classroom Assessment for Student Learning (CASL) to improve student achievement. CASL is a widely used professional development program in classroom and formative assessment published by the Assessment Training Institute of Pearson Education. CASL consists of the primary text of the same name (Stiggins, Arter, Chappuis, & Chappuis 2004), DVDs, ancillary books, and an implementation handbook (Chappuis 2007). Approximately 123,000 copies of the current edition of the CASL text have been sold (Stephen Chappuis, personal communication, September 10, 2009). CASL is typically implemented via teacher learning teams, in which teachers meet regularly to discuss and reflect on the content of the textbooks and DVDs and share their experiences applying the program in their classrooms. The terms "formative assessment" and "classroom assessment" have been defined in a variety of ways in the literature and in practice. For this study, formative assessment broadly includes assessment that happens during the learning process for the purpose of improving teaching and learning. Formative assessment, therefore, includes much of the assessment that happens in the classroom, but not assessment used specifically to document learning that has already happened, such as end-of-semester grades. The term classroom assessment includes all assessment that happens within the classroom regardless of its purpose. CASL predominantly emphasizes formative assessment but also addresses other types of classroom assessment, such as helping teachers understand standardized tests and how to use them productively in the classroom.

REL Central identified priority needs in the region for education research and technical assistance through the following process and with the following participants:

- Solicitation of specific regional issues and concerns from chief state school officers from the Central Region states (Colorado, Kansas, Missouri, Nebraska, North Dakota, South Dakota, and Wyoming) at their semi-annual meetings.

- Identification of needs expressed by the Regional Advisory Committee for the Central Region in their report to the U.S. Department of Education (Mid-Continent Regional Advisory Committee 2005).

- A comprehensive review of state demographic and education system data.

- Interviews with chief state school officers and key state agency staff.

- Ongoing contact with a variety of constituent groups, including policymakers, principals, and superintendents.

- In-depth interviews of state education agency staff in each of the seven Central Region states.

- Telephone interviews of a random sample of regional educators conducted by the Gallup Organization (Gallup Organization 2005).

Through the integration of these efforts, REL Central identified several priority needs to guide its work in the Central Region, including a need for guidance on research-based classroom practices and a need for improved teacher quality, particularly in light of the highly qualified teacher requirement of the NCLB Act. This study addressed the regional needs of poor performance in mathematics, the lack of the use of formative assessment, and a need for quality professional development for educators. The ultimate goal of providing scientifically based guidance on formative assessment was to help schools meet NCLB adequate yearly progress requirements.

Despite CASL's wide use, there is no direct causal evidence supporting its effectiveness in raising student achievement or improving other student and teacher outcomes. This study was designed to provide an unbiased estimate of the effectiveness of CASL to improve student achievement and other student and teacher outcomes. This study estimates the impact of CASL under conditions that would typically occur had the schools purchased CASL and implemented it without monitoring or involvement of research staff.

Research questions

This study examines the impact of CASL on the primary outcome of student achievement in mathematics. Although the intervention is applicable to all content areas, mathematics was chosen in view of the regional needs identified through REL Central's need-sensing process. In addition, the focus on this single content area was intended to reduce the data collection burden on participants and to prevent the intervention's impact from being dispersed across multiple content areas or focused on different content areas in different schools. Student mathematics achievement was measured by the statewide test administered under the NCLB Act; as a result, the impact estimate provides information about whether or not the intervention is effective in helping schools meet the goals of the NCLB Act.

The research team developed a theory of action to guide the design of the study and the development of research questions. This theory of action hypothesizes that teacher participation in CASL leads to increases in teacher knowledge of classroom assessment practices and principles, improvements in the quality of classroom assessment practice, and increases in the extent to which students are involved in classroom assessment. According to the theory of action, improvements in these three proximal outcomes lead to improved student motivation to learn and, in turn, to improved student achievement. The primary research question for this study was:

- Does CASL have a significant impact on student mathematics achievement?

According to the hypothesized theory of action that guided the design of this study, additional research questions were posed to address the impact of CASL on several intermediate outcomes, to provide additional information on the impact of the intervention, and to provide contextual information to aid in the interpretation of the impact on the primary outcome. The first intermediate outcome and its respective research question relate to student motivation:

- Does CASL have a significant impact on the extent to which students are motivated to learn mathematics?

CASL is unlikely to impact student achievement or motivation without first having an impact on teacher understanding and practice of formative assessment. The study, therefore, also addressed the following research questions about intermediate teacher outcomes:

- Does CASL have a significant impact on teacher knowledge of classroom assessment practices?
- Does CASL have a significant impact on the quality of classroom assessment practices?
- Does CASL have a significant impact on the extent to which teachers involve their students in formative assessment?

Study sample and design

Schools were recruited from across Colorado to participate in the study. Colorado was chosen as the target state primarily because it has one of the largest populations in the Central Region from which to recruit schools and because its statewide achievement test is vertically scaled. The target population for this study was public schools in Colorado large enough to have at least one grade 4 teacher and one grade 5 teacher. The study focused on grade 4 and 5 classrooms to allow for availability of baseline student achievement data from the grade 3 administration of the statewide NCLB achievement test. The focus on grade 4 and grade 5 was not based on any extant research.

From among Colorado's 178 districts, the study team identified 55 school districts that had six or more total schools (elementary, middle, and high schools). REL Central contacted these 55 school districts to request their schools' participation in the study. Separately, 332 of Colorado's elementary school principals who had signed up to be on Mid-continent Research for Education and Learning's (McREL) organizational mailing list were contacted to request their schools' participation in the study. The 55 school districts and the 332 principals were not cross-referenced; it is likely that at least some of the 332 principals were from at least some of the 55 districts.

Sixty-seven schools from 32 districts in Colorado participated in the study. These schools were not specifically selected using any sampling methodology; rather, they volunteered to participate in the study. This voluntary sample differed from all eligible Colorado elementary schools on a variety of demographic characteristics, such as percentage of students eligible for free or reduced price lunch and ethnic makeup of the student body. As such, the sample for this study does not

represent the population of Colorado elementary schools or any other broader population of schools.

Thirty-three schools were randomly assigned to the intervention group (CASL), and thirty-four schools were randomly assigned to the control group (where teachers conducted their regular professional development activities). The final student impact analysis sample included 9,596 students from the study schools. Sample size was sufficient to provide statistical power (greater than .80) to detect an impact on student mathematics achievement of 0.25 standard deviation, which would represent approximately 5 and ½ months of instruction in Colorado. Random assignment was blocked by district and resulted in two groups of schools that were found to have no statistically significant differences on a number of characteristics, for example, mathematics achievement, teacher–student ratio, and percentages of students in racial/ethnic groups.

The study included the students and teachers from all the grade 4 and 5 classrooms in the participating schools. Four-hundred-nine grade 4 and grade 5 teachers (178 intervention teachers and 231 control teachers) from a variety of large and small, urban, suburban, and rural schools participated in the study and were included in the impact analysis sample. The teacher sample included only teachers who provided direct mathematics instruction in grade 4 or grade 5. The intervention and control teachers did not differ by a statistically significant margin in their education or scores on the teacher measure of assessment knowledge measured at baseline. Intervention teachers, however, had more years of experience teaching (average =13.01, standard deviation = 9.16) than did control teachers (mean = 10.54, standard deviation = 7.95). Intervention teachers also had more years of experience in teaching mathematics (mean = 11.06, standard deviation = 8.55) than did control group teachers (mean = 8.90, standard deviation = 7.02). These differences in years of experience teaching and years of experience teaching mathematics were statistically significant and controlled in the impact analysis by including these two teacher experience variables as covariates.

Data collection procedures and instruments

Over the course of the study, data were collected to describe the fidelity of CASL implementation and the larger professional development context. Data were also collected to estimate the impacts of CASL on the student and teacher outcomes. Student achievement data were obtained directly from the Colorado Department of Education for the Colorado Student Assessment Program (CSAP) administered in April of each year; survey data were collected from teachers and students in the participating schools.

Teacher background information, such as years experience and highest degree, was collected with a short survey administered to teachers at the beginning of the study. CASL implementation fidelity data were collected with logs administered to intervention group teachers at the beginning of the study and after they completed each chapter of the CASL textbook. The logs addressed the amount of the chapter read, the activities completed for each chapter, the learning team meetings and attendance at those meetings, and the total number of hours spent training with CASL for each chapter. Data on teachers' non-CASL professional development were collected from all participating teachers with a survey administered at the end of each semester during the 2007/08 school year and the 2008/09 school year. Teachers reported the types of

professional development activities that they participated in, their frequency, duration, subject area, emphasis, perceived quality, and perceived impact on classroom practice.

Students' scale scores on the 2009 administration of the mathematics portion of the Colorado Student Assessment Program were used to estimate the impact of CASL on student achievement. The outcome of student motivation to learn was measured using the Ongoing Engagement and Perceived Autonomy (Self-Regulation) subscales of the elementary student Research Assessment Package for Schools (IRRE 1998) and the Academic Efficacy subscale of the Patterns of Adapted Learning Scales (Midgely, Maehr, Hruda, Anderman, Anderman, Freeman, et al. 2000). Student motivation was measured in May of the 2008 school year and May 2009 (posttest). Teacher knowledge of classroom assessment was measured using a 60-item test of multiple-choice, true/false, and matching items that covered teacher knowledge of, and reasoning skills regarding, generally accepted principles and practices of classroom assessment. The quality of classroom assessment practices were measured with a work sample instrument used to collect and score teacher classroom assessment artifacts. The artifacts were scored by two raters blind to the teachers' experimental group membership using a 6-dimension rubric. Finally, teachers' involvement of their students in assessment was measured with a 14-item self-report survey where teachers reported the number of days during a two-week instructional period that they involved all or most of their students in activities related to assessment, such as discussing the learning objectives, evaluating their own work using scoring guides or rubrics, and revising work to correct errors.

Intervention training and implementation

REL Central provided schools in the intervention group with the complete set of CASL professional development materials at the beginning of the study (November 2007), including a facilitation handbook (Chappuis 2007), CASL textbooks (Stiggins et al. 2004) for every participating teacher, DVD sets, and ancillary books. Teachers in the intervention group also participated in an introductory videoconference with CASL author Richard Stiggins and had access to a facilitator who had attended a training workshop conducted by the CASL developers.

The CASL program is designed to be self-executing, without a coach or external facilitator. The handbook provides guidance and developer recommendations for implementing the program (Chappuis 2007). The developers recommended implementing CASL using teacher learning teams, in which teachers meet regularly to discuss and reflect on the content of the program provided in the textbook and DVDs and share their experiences applying the program practices and principles in their classrooms. Teachers in the intervention group implemented CASL naturally, without any involvement of, or requirements from, the research team. Intervention teachers studied the CASL materials and applied the CASL principles, practices, and tools in their classrooms during the 2007/08 school year (the CASL training year). During the 2008/09 school year (the CASL implementation year), intervention teachers implemented the CASL program in their classrooms for one full school year after completing the training year.

Implementation fidelity of CASL in the intervention schools was assessed by the research team using teacher self-report participant logs. One-hundred-fifty-eight teachers out of the 175 randomly assigned to the intervention group (90.29%) provided at least some implementation

fidelity data. Although implementation fidelity varied within the intervention group, all teachers and learning teams for whom data were available were included in the teacher impact analysis regardless of their level of implementation fidelity. Implementation fidelity was assessed at both the school and individual teacher level. At the school level, 68 percent of the learning teams in schools in the intervention group met the overall quality criterion for learning teams, defined as meeting at least four of five criteria established by the CASL developers: establishing group operating principles, setting a meeting schedule, choosing a regular place to meet, establishing agreements regarding activities between meetings, and sharing a common purpose. Learning teams ranged from three to eight members, with 90 percent of learning teams made up of the recommended size of three to six members. Sixty-three percent of the 142 teachers who responded to relevant log items attended at least the recommended nine learning team meetings, with an average of 9.78 meetings during the CASL training year (SD = 8.04). Seventy-eight percent of the 121 teachers who responded to at least one relevant log item reported that their learning team meetings involved discussions on what they were learning about classroom assessment. At the individual teacher level, 42 percent of the 142 teachers who provided at least some data regarding reading the CASL textbook reported at least partially reading each CASL textbook chapter, and 40 percent reported reading each chapter fully. Based on data provided by the 142 teachers who responded to at least one relevant log item, the average amount of total time that teachers reported spending on CASL training was 31 hours (SD=19.89, minimum = 2.00, maximum = 115.00), as compared with the 60 hours recommended by the developer.

Analysis and results

Confirmatory impact analyses were conducted for the primary outcome, student mathematics achievement. Exploratory impact analyses were conducted for all of the intermediate outcomes. The student achievement impact analysis sample included all schools that were randomly assigned at the beginning of the study to either the intervention or control group.

The impact analysis samples for the intermediate outcomes were reduced by school and teacher nonresponse and attrition. Nonresponse occurred when teachers failed to complete data collection. Attrition occurred when teachers or schools withdrew from the study and were excluded from the analysis because no data were available. The Wave 3 (May 2008) student motivation sample excluded 12 schools: five intervention schools and seven control schools. The Wave 5 (posttest) student motivation sample excluded 11 schools: nine intervention schools and two control schools. All teachers from three schools were excluded from the teacher outcomes impact analysis sample because they withdrew from the study prior to baseline data collection. Forty-one teachers (9.11 percent of the total sample) were excluded from the teacher outcomes impact analysis sample (26 intervention group teachers and 15 control group teachers) because they failed to provide any data. Comparisons of the baseline characteristics of the samples of schools used in the intermediate outcome impact analyses did not reveal any statistically significant differences between the intervention and control groups.

Schools and teachers were excluded from the impact analysis samples when no data were available. For students and teachers with only partial missing data, the expectation maximization algorithm with multiple imputation method was used to impute the missing data. The expectation maximization algorithm is an iterative statistical method that replaces missing data with imputed

data (Puma, Olsen, Bell, & Price 2009). Missing data are replaced with values predicted from the relationships between all of the variables using all available observations including those with missing data. Randomly selected residual error values are then added to the imputed values to ensure that the replacement process does not incorrectly reduce the natural variation in the data.

Average student mathematics achievement as measured by the mathematics portion of the CSAP did not differ by a statistically significant margin between the intervention group (adjusted mean = 502.49, standard error = 2.53) and control group schools (adjusted mean = 501.91, standard error = 2.44) on the CSAP scale score metric. Follow-up exploratory analyses found that CASL did not have a statistically significant impact on either grade 4 or grade 5 adjusted mean mathematics achievement. At grade 4, the intervention group students' average score was 494.19 (standard error = 3.06) and the control group students' average score was 488.60 (standard error = 2.95), and at grade 5, the intervention group students' average score was 510.13 (standard error = 2.88) and the control group students' average score was 514.68 (standard error = 2.75).

Intervention and control schools did not differ at a statistically significant level on the extent to which students were motivated to learn mathematics. For Wave 3 (May 2008), the intervention group students' average rating was 3.29 (standard error = 0.02) and the control group students' average rating was 3.28 (standard error = 0.02) on the survey's 4-point scale where 1 = "not at all true" and 4 = "very true." For the Wave 5 (posttest) motivation outcome, intervention students' average rating was 3.33 (standard error = 0.02) and control students' average rating was 3.32 (standard error = 0.02).

In terms of teacher outcomes, CASL had a statistically significant impact ($p = .01$) on intervention teachers' knowledge of classroom assessment: intervention teachers answered an average of 41.36 items correctly (standard error = 0.76) on the 60-item test as compared to an average of 38.58 items (standard error = 0.60), a difference of 2.78 items or 0.42 standard deviation. Teachers from the intervention and control schools were similar in the quality of their classroom assessment practices and the extent to which they involved their students in formative assessment; no statistically significant impacts were found on these two intermediate outcomes. For classroom assessment practice, intervention teachers were given an average rating of 1.61 (standard error = 0.05) on the rubric scale where 1 represented low quality and 4 represented high quality, as compared to the control group teachers' average rating of 1.60 (standard error 0.04). For student involvement in formative assessment, intervention teachers' average response on the survey was .39 (standard error = 0.02) whereas the control teachers' average response was .34 (standard error = 0.02) where 1.00 represents students involved in formative assessment activities for 100 percent of the instructional days during the identified two-week instructional period and 0.00 represents no student involvement.

Conclusions

This cluster randomized trial of the CASL professional development program had sufficient statistical power to detect an impact of at least 0.25 standard deviation on student achievement. An intent-to-treat analysis was conducted to estimate the impact of CASL on student achievement; all schools were included in the analysis and were analyzed as randomized regardless of the level of implementation fidelity. Analysis did not reveal a statistically

significant impact of CASL on the school-level average mathematics achievement of grade 4 and grade 5 students. Results from sensitivity analyses revealed that the impact estimates on student achievement were robust to decisions regarding the inclusion of covariates, estimation method, and the treatment of missing data. In other words, design and analysis decisions made by the research team did not change whether the impact results would have been statistically significant.

Interpretation of this study is subject to several limitations. The first set of limitations concerns generalizability of results. Results from this study generalize only to implementation of CASL with similar degrees of intensity at both the school level (e.g., number of learning team meetings) and at the teacher level (number of chapters read and average number of total hours spent on CASL program activities). Results from this study do not generalize to formative assessment practices in general. Study results also only generalize to the voluntary sample included in the study and to student grade 4 and 5 mathematics achievement as measured by the Colorado statewide achievement test. Second, although attrition was not an issue for the student achievement outcome, the attrition and nonresponse for the student motivation outcome and the teacher outcomes, however, exceeded levels considered acceptable (What Works Clearinghouse 2008). According to the What Works Clearinghouse, unacceptable attrition is attrition that results in the estimated impact of the intervention deviating from the true impact (What Works Clearinghouse 2008). The multiple imputation method used to impute missing data for the teacher outcomes resulted in a teacher impact analysis with levels of attrition expected to result in an acceptable level of bias (What Works Clearinghouse 2008). Finally, it should be pointed out that it is not correct to interpret impact estimates that were not statistically significant as evidence of no impact. Rather, these estimates failed to provide any evidence of an impact.

Chapter 1. Introduction and study overview

Under the No Child Left Behind Act of 2001 (NCLB), all students are expected to attain proficiency in state content standards by the end of the 2013/14 academic year. To help students reach this goal, educators need information on effective and efficient research-based interventions.

The terms "formative assessment" and "classroom assessment" have been defined in a variety of ways in the literature and in practice. For this study, formative assessment is assessment that happens during the learning process with the purpose of assisting teaching and learning (Cizek 2010). Formative assessment includes much of the assessment that happens in the classroom; not all assessment in the classroom, however, is used to improve learning. Rather, some assessment in the classroom is used exclusively for documenting learning that has already occurred, such as end-of-semester grades. The term classroom assessment, then, includes all assessment that happens within the classroom, whether its purpose is to improve learning or document learning.

Prior research suggests that quality formative assessment in the classroom can improve student achievement. Black & Wiliam's (1998a) oft-cited review of studies of the impact of classroom assessment interventions on student learning stated that "innovations that include strengthening the practice of formative assessment produce significant and often substantial learning gains" (p. 140). In a review of Black & Wiliam's (1998a) study, Bennett (2009) noted the value of the qualitative synthesis of a broad array of research on practices related to formative assessment while recognizing that Black & Wiliam's study was not a meta-analysis of a single, well-defined set of interventions. Bennett (2009) concluded that school effectiveness research suggests that the practices associated with formative assessment can, under the right conditions, facilitate learning.

Other research has revealed differential effects of formative assessment on low-achieving students as compared with high-achieving students, with low-achieving students realizing larger gains, which suggests that effective classroom assessment may help reduce achievement gaps (White & Frederiksen 1998). More recently, a non-experimental comparison study on the effects of professional development in classroom assessment found an average effect size of 0.32 on student achievement across teachers after six months of teacher training (Wiliam, Lee, Harrison, & Black 2004).

Although research provides some suggestions that the use of formative assessment holds promise for raising student achievement, prior research also suggests that teachers often receive little training in classroom assessment or related topics as part of their teacher preparation experience (O'Sullivan & Chalnik 1991; Schaffer 1993). Many in-service teachers lack training, knowledge, and skills in classroom assessment (Marso & Pigge 1993; Plake, Impara, & Fager 1993; Plake & Impara 1997). For example, only one-quarter to one-third of middle school teachers were found to administer coherent assessments, defined as assessments aligned with learning goals and evaluation criteria (Aschbacher 1999).

Cizek (2010) recently presented a constellation of key characteristics of formative assessment that includes practices such as requiring students to take responsibility for their own learning, encouraging students to self-monitor, and providing feedback that is nonevaluative. If teachers typically receive little preservice training in formative assessment and if in-service teacher formative assessment literacy is low, realizing the promise of formative assessment for raising student achievement will likely require effective professional development programs to improve teacher practice of formative assessment.

According to the NCLB Act, high-quality professional development is sustained, intensive, content-focused, aligned with standards, and increases teacher understanding of the subjects they teach and research-based teaching strategies for those subjects. Recent research supports this definition of high-quality professional development. Yoon, Duncan, Lee, Scarloss, & Shapley (2007) conducted a quantitative meta-analysis of studies addressing the effects of in-service professional development of any type on student achievement in mathematics, science, and language arts and found nine studies that met What Works Clearinghouse standards. The analysis found that teachers who received substantial professional development (an average of 49 hours total over the course of the entire training program) raised their students' achievement by an average of 21 percentile points.

Although research suggests that both formative assessment and professional development are promising approaches to improving teacher practice and student achievement, little rigorous research has examined the impact of professional development in formative assessment on teacher knowledge and classroom assessment practices (Schneider & Randel 2010). More rigorous effectiveness research on professional development in formative assessment is needed to determine whether or not formative assessment can be used to increase student achievement or other student and teacher outcomes.

Study purpose

This study was conducted at Mid-continent Research for Education and Learning under its Central Region Educational Laboratory (REL Central) contract with the Institute of Education Sciences in the U.S. Department of Education. The study was designed to respond to specific regional issues and concerns expressed by chief state school officers from the Central Region states (Colorado, Kansas, Missouri, Nebraska, North Dakota, South Dakota, and Wyoming). This study addressed the regional needs of poor performance in mathematics, the lack of the use of formative assessment, and a need for quality professional development for educators. The ultimate goal of providing scientifically based guidance on formative assessment was to help schools meet NCLB adequate yearly progress requirements.

The concerns of the chief state school officers were echoed in in-depth interviews of state education agency staff in each of the seven states, telephone interviews of a random sample of regional educators by the Gallup Organization (Gallup Organization 2005), and the Regional Advisory Committee for the Central Region. Each of these sources of guidance indicated needs in broad policy and practice areas rather than requests for specific studies or information on specific interventions. Poor performance in mathematics, the lack of the use of formative assessment, and the need to improve classroom practice were all mentioned as priorities. The

need for quality professional development and the need to help schools make adequate yearly progress were identified as high-priority issues in a Gallup survey of regional educators (Gallup Organization 2005) commissioned by REL Central.

Despite its widespread use and anecdotal evidence of effectiveness, no direct causal evidence was available regarding the effectiveness of CASL on student achievement or on the other student and teacher outcomes identified above. The aim of the study was to provide educators and policymakers with rigorous evidence regarding the impact of a specific widely used, professional development program in classroom and formative assessment on student mathematics achievement, student motivation, and teacher knowledge and practice of formative assessment. The purpose of this study was not to evaluate the practices of formative assessment in general.

Overview of the intervention

The intervention for this study was a program of professional development in classroom assessment for classroom teachers: Classroom Assessment for Student Learning (CASL). CASL is a widely used professional development program in classroom and formative assessment published by the Assessment Training Institute of Pearson Education. CASL consists of the primary text of the same name (Stiggins et al. 2004), DVDs, ancillary books, and an implementation handbook (Chappuis 2007). CASL was chosen as the intervention for this study because it is a program for improving classroom assessment that has been widely used for more than 10 years, with anecdotal evidence of its effectiveness. The first edition of the CASL textbook was published in 1994; the current, fourth edition was published in 2004 and since then approximately 123,000 copies have been sold (Stephen Chappuis, personal communication, September 10, 2009). According to the definitions of formative assessment and classroom assessment at the beginning of this chapter, CASL predominantly emphasizes formative assessment but also addresses other types of assessment that occur in the classroom, such as helping teachers understand standardized tests and how to use them productively in the classroom.

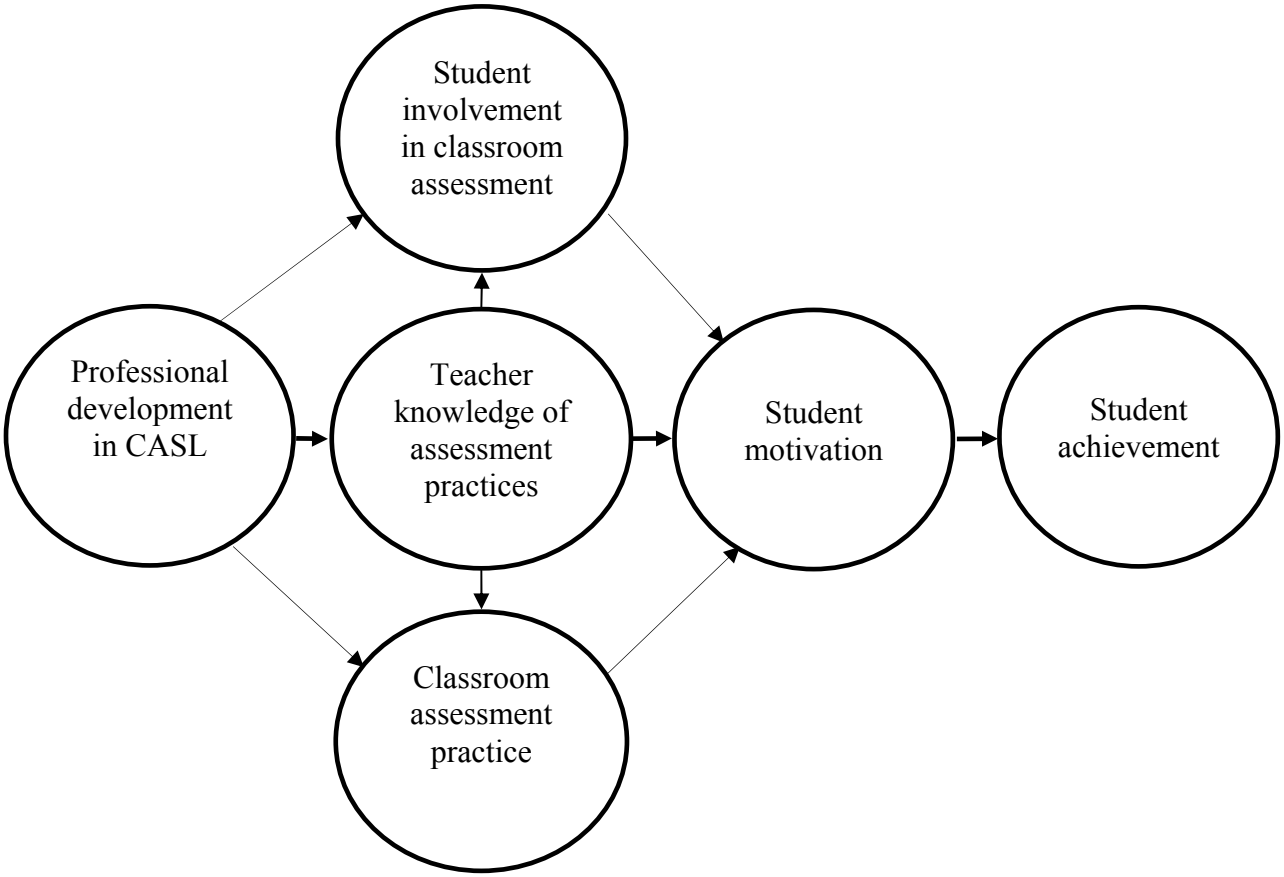
The CASL intervention includes studying the CASL textbook (Stiggins et al. 2004) and other materials; applying sound classroom assessment principles, practices, and tools in the classrooms; and receiving support and problem-solving guidance from learning teams. Learning teams are made up of groups of three to six teachers "who have committed to meet regularly to for an agreed amount of time guided by a common purpose: to help all members increase classroom assessment competence through collaboration during team meetings, and individual study and practice between meetings" (Chappuis 2009, p. 19). Learning teams discuss and reflect on the content of the program provided in the textbook and DVDs and share their experiences in applying the program practices and principles in their classrooms. The CASL program is designed to be self-executing so that an external agent, such as a training coach, is not part of the program. In other words, schools typically buy the program materials and use the handbook to "successfully conduct learning teams around the study of the text" (Chappuis 2009, p. 3). A detailed description of the CASL intervention can be found in Chapter 3.

REL Central's research team developed a theory of action to guide the design of the study and the development of research questions and to serve as a conceptual representation (figure 1.1); the theory of action is not meant as a literal depiction of the causal pathways through which the intervention achieves its hypothesized impacts. This theory of action purports that teacher participation in textbook and materials study, classroom application, and learning teams leads to increases in teacher assessment knowledge and practice, quality of classroom assessment practice, and the extent of student involvement in classroom assessment. Improvements in these three proximal outcomes were hypothesized to result in improved student motivation to learn and, in turn, to improved student achievement.

Teacher knowledge of classroom assessment practices

The CASL approach to improving student achievement begins by improving teacher knowledge of sound classroom assessment practices and principles. Teachers study and discuss the material in the CASL textbook to increase their assessment literacy. The CASL textbook content is organized around five key components of classroom assessment: assessment purposes, clear learning targets, sound classroom assessment practices, communication and management of results, and student involvement in classroom assessment. As teachers progress through the CASL program, they acquire the knowledge and reasoning skills expected to help them improve their classroom assessment practices.

Figure 1.1. Theory of action



Classroom assessment practices

In addition to helping teachers increase their knowledge of the practices and principles of classroom assessment, CASL endeavors to improve teacher assessment practices. CASL's emphasis on improving assessment practice is intended to help teachers obtain accurate information regarding student performance. Accurately assessing student achievement includes matching the assessment method to the learning target, appropriately sampling the domain based on the breadth and depth of learning targets, creating well-written assessment items, developing quality scoring rubrics, and reducing measurement bias.

CASL is intended to help teachers learn how to align assessment methods with learning targets by improving teacher understanding of different types of learning targets and the assessment methods that are most appropriate for each type of learning target. CASL is intended to help teachers learn how to develop and use multiple-choice and short-answer tests, extended response assessments, performance assessments, and personal communication as assessment. According to the developers, CASL also helps teachers develop assessments that reflect the relative importance of the learning targets in terms of the number of items on the assessment and the number of score points on the assessment assigned to each learning target. CASL further provides teachers with guidance and practice activities intended to improve the quality of the assessments they develop so assessments are well written and have minimal potential for bias.

Accurate information regarding student performance obtained from classroom assessments is thought to influence student learning. Airasian & Jones (1993) contend in their description of classroom assessment that assessment results can help teachers make sound instructional decisions. Developing assessments that measure clear learning targets linked to content standards was found in an experimental study to provide teachers with information regarding student progress toward the content standards (Bergan, Sladeczek, Schwarz, & Smith 1991). As a result, teachers are more readily able to adjust their instruction to help students meet those goals, as reported by Black, Harrison, Lee, Marshall, & Wiliam's (2003) descriptive research into teacher practices. Bergan et al. (1991) also found that accurate information regarding student progress toward standards enhance teacher confidence and ability to make instructional decisions.

One of the primary purposes of obtaining accurate information from classroom assessments is to communicate that information to the student. The CASL program emphasizes the importance of feedback. Feedback is defined as information about a student's current understanding or performance as compared with the desired level of understanding or performance. Such information is purported to be useful to help the student move closer to the desired performance level (Ramaprasad 1983). Stiggins et al. (2004) state that feedback should reflect "student strengths and weaknesses with respect to the specific learning targets they are trying to hit in a given assessment" and should convey "information that provides insight so one can continue to improve one's work" (pp. 43, 363).

CASL focuses on helping teachers provide relevant feedback. Indicators of relevance include timeliness and accuracy. Relevant feedback is a function of how well the knowledge and skills being assessed are understood and articulated as developmental trajectories, including prerequisite knowledge, advanced proficiencies, and common errors along the way. Feedback

that is inaccurate or that comes too late to be used to improve work is not relevant, and inaccurate feedback may even be harmful to student learning. CASL also helps teachers provide feedback that focuses on the task, not on the individual or any reference group of individuals. This type of feedback is thought to be effective in changing student behavior (Black & Wiliam 1998a). In other words, CASL helps teachers provide students with descriptive feedback.

Kluger & DeNisi (1996) found in a meta-analysis of results from 131 empirical articles that descriptive feedback is more effective in raising student achievement and student motivation than is evaluative feedback. Evaluative feedback (such as grades) is often norm- or peer-referenced and used for administrative purposes—for example, to determine a student's rank in class, as defined by Marzano, Gaddy, & Dean (2000) in their instructional guide. Evaluative feedback rarely provides a description of the quality of the work being graded. Descriptive feedback, by contrast, gives students information about their achievement relative to specific learning targets and emphasizes improvement, as found in an empirical study by Tunstall and Gipps (1996). Bangert-Drowns, Kulik, Kulik, & Mogran's (1991) meta-analysis of 40 studies of feedback found that descriptive feedback provides students with information that helps them correct errors. Descriptive feedback can confirm students' self-assessments or help students become aware of errors, as found by Taras (2003) in an empirical study of British university students. Contemporary understanding of feedback in classroom learning emphasizes the active role of the student: effective feedback encourages "mindfulness" in the student's response to the feedback (Black & Wiliam, 1998a, p. 28).

Student involvement in classroom assessment

CASL is hypothesized to improve student achievement by increasing students' active involvement in the learning process. According to the CASL approach, teachers are taught to involve students in all aspects of classroom assessment, from test development, writing practice items, and reviewing test plans to self-assessment, revisions, and grading. To help students understand the learning targets, teachers can share test plans with students, identify weak and strong examples of student work, ask students to match learning propositions with test plans, ask students to develop learning propositions or test items, have students develop rubrics, or have students score their own or one another's work according to rubrics.

A well-written scoring rubric can be used as part of self-assessment where students identify areas of strengths and weakness that may require more work. CASL encourages students to record what they have learned in a journal or log to keep track of their progress and to identify examples of their own work that represent important steps toward the learning target. Self-assessment helps students understand not only the learning target but also their performance in relation to the learning targets. In a recent experimental study, the test scores of students who graded themselves showed large, statistically significant increases, whereas the scores from students who graded their peers did not increase by a statistically significant margin (Sadler & Good 2006).

CASL also encourages focused revision, in which students revise only a single aspect of the product at a time, as a method for increasing student involvement. Students can also be asked to create a revision plan or explain how they would improve their own or others' work. Revising

and refining their work under their teachers' guidance and using their understanding of the learning targets is recommended as a strategy in Charter's (1984) teachers' guide to help students acquire strategies for closing the gap between the target and their current levels of performance. Both Boud (1986) and Lindeman (1982) contend in their separate teachers' guides that students can learn new ways to solve problems by explaining to others their reasoning and the processes they used to produce their work. Because CASL requires students to be active participants in formative assessments, they are more aware of their learning targets, their strengths and weaknesses relative to the learning targets, and how they can close the gap between what they know and are able to do and what they are expected to know and be able to do.

Student motivation

Stiggins et al. (2004) asserted that classroom assessment can help increase student achievement by increasing student motivation. According to the CASL authors, formative assessment increases motivation by giving students the confidence that they can learn. Sound formative assessment describes what students know, what the learning targets look like, and how to close the gap between the current quality of their work and where it needs to be. With this knowledge, students feel more self-efficacious and in control of their learning and engage more actively in the learning process. Self-efficacy, as defined by Bandura (1994, 1997) is an individual's perceived competence for a particular task. Students with high self-efficacy have confidence that they know what it takes to be competent in a particular activity or subject area and that they can learn what they need to learn even if it is difficult (see Institute for Research and Reform in Education [IRRE] 1998 for a literature review; Skinner, Wellborn, & Connell 1990 for correlational research). In short, self-efficacy gives students the confidence to engage in learning, and self-efficacy's positive relationship with achievement has been well documented by correlational research (see, for example, Eccles, Wigfield, Flanagan, Miller, Reuman, & Yee 1989; Schunk 1991; Wigfield 1994).

Student engagement can include both behavioral and emotional components. Engaged students are behaviorally involved in learning activities and exhibit positive emotions while so involved. They initiate action when possible, choose challenging activities, work with effort and concentration, and stay on task. Their positive emotions include interest, enthusiasm, optimism, and curiosity (see Skinner & Belmont 1993 for correlational research and path analysis). Student engagement is hypothesized to be essential to achievement because engaged students are doing the work necessary for learning. As demonstrated in empirical research, high levels of student engagement can even explain why some at-risk students achieve at high levels (Connell, Spencer, & Aber 1994; Finn 1993).

Self-regulation deals with how students take in extrinsic contingencies (such as assignments) and values (such as being hardworking) and transform them into personal values and motivations. Students who are self-regulated will engage in learning tasks because they think the material is important and want to understand it or because doing well in school is important to them, not because they are being made to do it by others. They may report that learning tasks are fun or enjoyable. According to theory, students with high self-regulation tend to exhibit high levels of engagement as well and are more likely to persist in the face of difficulty (IRRE 1998; Ryan & Deci 2000).

Study design overview

This study was designed to estimate the impact of CASL on the primary outcome of student achievement and to estimate CASL's impact on intermediate student and teacher outcomes. Sixty-seven schools from 32 districts in Colorado were randomly assigned to either the intervention or control group. In the intervention group schools received the CASL program materials and implemented the program as they would have if they had not been participants in a research study. In the control group, teachers continued with their regular professional development activities. To aid in recruiting of schools and to balance resources provided to the intervention schools, the research team provided each control group school with $1,000 at the beginning of the study. During recruiting, all schools were informed that they would receive either the intervention materials or the approximate financial equivalent of the intervention materials at the beginning of the study. This strategy was used to encourage the participation of schools wary of being assigned to the control group and not receiving any benefit of participation until after the study was completed. Because the CASL intervention essentially consists of materials with no direct coaching or facilitation, any impact could be attributed to the receipt of these additional resources in the intervention group. To eliminate the alternative hypothesis that any potential intervention impact was simply the result of the availability of more resources to the intervention group, the control group was provided with the approximate equivalent in financial resources of the intervention materials to use as they saw fit.

The study focused on grade 4 and 5 classrooms to allow for availability of baseline student achievement data from the grade 3 administration of the statewide NCLB achievement test. The focus on grade 4 and grade 5 was not based on any extant research. All grade 4 and 5 classrooms, teachers, and students in each school were required to participate in the study. The study included 409 teachers from a variety of large and small, urban, suburban, and rural schools. Approximately 9,600 students were included in the impact analyses.

The study was conducted in the participating schools during the 2007/08 and 2008/09 school years. During the 2007/08 school year, intervention teachers studied the CASL materials and applied the CASL principles, practices, and tools in their classrooms (the CASL training year). The 2008/09 school year was included in the study to allow intervention teachers one full school year to implement the CASL program in their classrooms after completing the training.

Research questions

The primary focus of this study was to estimate the impact of CASL on student achievement. Although the intervention is applicable to all content areas, mathematics was chosen as the focus of this study to address regional concerns regarding mathematics achievement. In addition, collecting data for one subject area reduced the burden on participants and helped prevent the intervention's impact from being dispersed across multiple content areas or being focused on different content areas in different schools. Student mathematics achievement was measured by the mathematics portion of the statewide test administered as part of the NCLB Act. As a result, the impact estimate provides information about whether the intervention is effective in helping schools meet the goals of the NCLB Act. The primary research question was:

- Does CASL have a significant impact on student mathematics achievement?

Additional research questions were posed to address the impact of CASL on several intermediate outcomes included in the theory of action. The examination of the intermediate outcomes was intended to provide additional information about the impact of the intervention and aid in the interpretation of the results of the primary research question. The first intermediate outcome, student motivation, was included in the study per the theory of action and addressed with the following research question:

- Does CASL have a significant impact on the extent to which students are motivated to learn mathematics?

CASL is unlikely to impact student achievement or motivation without first affecting teacher understanding and practice of formative assessment. CASL's impact on teacher knowledge and practice as intermediate outcomes was thus of interest because it could inform interpretations of the estimated impact of CASL on student achievement. The following research questions regarding intermediate teacher outcomes were also addressed in this study:

- Does CASL have a significant impact on teacher knowledge of classroom assessment practices?
- Does CASL have a significant impact on the quality of classroom assessment practices?
- Does CASL have a significant impact on the extent to which teachers involve their students in formative assessment?

Content and organization of this report

This study estimates the impact of CASL on student achievement in mathematics. Impact estimates for the intermediate outcomes were also addressed. Chapter 2 describes the study sample, the study design, data collection, and impact estimation approach. Chapter 3 describes the implementation of the CASL intervention and participating teachers' non-CASL professional development experiences. Chapter 4 presents impact estimates of CASL on primary outcomes. Chapter 5 presents exploratory analyses estimating the impact of CASL on the intermediate outcomes. Chapter 6 summarizes the study's key findings and discusses its limitations.

Chapter 2. Method and study design

The goal of this study was to estimate the impact of Classroom Assessment for Student Learning (CASL) when implemented in real-world conditions. Although teachers learn and apply CASL individually, the majority of sites that use CASL (75%) purchase the school pack, which is structured for implementation through a learning team. This study used a cluster randomized design in which schools were randomly assigned to either the intervention or control group to allow teacher teaming and collaboration to continue across and within grades through learning teams. Principals and other instructional leaders in schools in both the intervention and control groups were able to continue their usual practices of encouraging input and collaboration. Random assignment of whole schools also reduced the risk of crossovers between classrooms in the same building (for example, control group members acquiring intervention materials and using them in their classroom). In the intervention group teachers formed learning teams, received the professional development materials, and implemented the CASL program (see chapter 3 for a complete description of the implementation). In the control group, teachers participated in their regular professional development activities.

This chapter describes the research methods and study design. The chapter is organized around five major sections: study timeline, study sample, attrition and nonresponse, data collection, and data analysis methods.

Study timeline

Over the course of the study, research staff visited each participating site to conduct study orientations, deliver intervention materials to intervention schools, and provide instructions for each wave of data collection (table 2.1). The CASL program is designed to be completed in approximately one school year. Teachers in the intervention schools studied the CASL materials and applied the CASL practices and principles in their classrooms from November 2007 to May 2008 (the CASL training year). The study also included the entire 2008/09 school year (the CASL implementation year) to allow intervention teachers one full school year to implement the CASL program in their classrooms after completing the training. (See table 2.16 for a detailed description of the data collection schedule and respondents.)

Study sample

This section covers sample recruiting, random assignment of schools, and the student achievement, student motivation, and teacher samples.

Sample recruiting

The study team focused recruiting efforts in Colorado for several reasons. Colorado is one of the most populous states in the Central Region and therefore provided a large pool from which to recruit schools. The statewide assessment, the Colorado Student Assessment Program (CSAP), provides reliable scale scores on a vertical scale, allowing test data from grades 4 and 5 to be combined in a single impact analysis. In addition, conducting the study in one state eliminated

any issues that might result from attempting to aggregate achievement data across states using different achievement tests.

Table 2.1. Timeline of key activities

Date	Activity
April 2007	2007 Colorado Student Assessment Program (CSAP) administration
October 2007	Participation agreements (district and school memoranda of understanding and teacher consent forms)
November 2007	Random assignment of schools
	Study participant orientation
	Wave 1 (baseline) data collection
	Delivery of Classroom Assessment for Student Learning (CASL) materials to intervention schools
	Beginning of CASL training year
December 2007	Wave 2 site visits and data collection
April 2008	2008 CSAP administration
May 2008	Wave 3 site visits and data collection
August 2008	Beginning of CASL implementation year
December 2008	Wave 4 site visits and data collection
April 2009	2009 CSAP administration
May 2009	Final site visits and teacher Wave 5 (posttest) data collection

Note: Baseline refers to the time point at the beginning of the study, at the first wave of data collection and prior to the delivery of CASL materials to intervention schools. Only teachers provided data at baseline.

The target population for this study consisted of public schools in Colorado of sufficient size to have at least one teacher in grade 4 and one teacher in grade 5. Fifty-five of Colorado's 178 school districts had at least six total schools (elementary, middle, and high); these districts were contacted directly by the study team by telephone and email to provide them with information about the study and request their participation. A flyer containing information about the study was mailed to the 332 Colorado elementary school principals who had signed up to be on Mid-continent Research for Education and Learning's (McREL) organizational mailing list, and follow-up telephone calls were made to these principals to solicit their participation. The 55 school districts and the 332 principals were not cross-referenced; it is likely that at least some of the 332 principals were from at least some of the 55 districts. Additional recruiting strategies included posting information about the study on the McREL website, providing study information in the Colorado Association of School Executives monthly bulletin, and posting information about the study in the Colorado League of Charter Schools newsletter. Recruiting at the district level helped ensure district administration support for the study and facilitated access to student-level achievement data. Recruiting at the school level helped ensure that schools from small districts had the opportunity to participate.

Districts and schools that wished to participate in the study were required to meet several criteria. First, districts and schools needed to provide permission for the study team to use student-level achievement data from the statewide student achievement test. Districts and schools also had to understand, agree to, and abide by the random assignment of schools to either the intervention or control group. Schools assigned to the control group had to agree not to purchase or implement the CASL program during the duration of the study, the 2007/08 and 2008/09 school years. Participating schools also were required to form a learning team that included all grade 4 and grade 5 teachers and included at least three individuals (principals and other teachers could also

be members of the learning team, but did not participate in data collection). Teachers had to be willing to complete all of the study requirements. As noted in chapter 1, the study design required all teachers in grades 4 and 5 providing direct instruction in mathematics to be included in the study to prevent crossover or selection bias within intervention schools.

During the recruiting process, districts, schools, and teachers were offered the following benefits for participating in the study:

- All CASL materials for intervention schools at the beginning of the study.
- All CASL materials for control schools at the end of the study.
- $1,000 to control schools at the beginning of the study to counterbalance the resources available to intervention schools.
- The opportunity to send one district-level staff member to the intervention developers' training seminar for every four schools participating, regardless of assignment to either the intervention or control group.
- The opportunity for intervention group teachers to earn up to four hours of graduate credit.

Sixty-seven schools from 32 districts agreed to participate in the study. The study sample of elementary schools represented 6.83 percent of all eligible elementary schools in Colorado. Results of statistical power analyses estimated that 67 schools would provide sufficient power (greater than .80) to detect impacts of 0.25 standard deviation on student mathematics achievement. (Appendix A presents detailed information on the power analyses and assumptions of cluster size, effect sizes, and variation.) The districts and schools were located across Colorado and provided a diverse sample in terms of district and school size, levels of urbanicity (urban, suburban, town, and rural), and student achievement.

Table 2.2 shows the results of a comparison of the study sample of schools with the eligible elementary schools in Colorado that did not participate in the study. The average mathematics achievement among study schools was higher for grade 5 in 2007 than the average mathematics achievement in eligible Colorado schools that did not participate in the study. There were proportionally fewer suburban schools and more rural schools in the study sample than in the group of nonparticipating eligible schools. The percentage of students eligible for free or reduced-price lunch was higher in study sample schools than in the nonparticipating eligible Colorado schools, and the proportion of schools in the study sample that were Title I eligible was higher than the proportion of schools that were eligible but that did not participate in the study. The study sample schools also had a higher percentage of Hispanic students and lower percentages of Asian American and American Indian students than did the eligible schools Colorado schools that elected not to participate in the study. These results suggest that the sample of schools that participated in the study was different from the group of remaining Colorado eligible schools that did not participate in the study. Results from this study, therefore, may not generalize to the broader population of Colorado elementary schools.

Table 2.2. Comparison of study schools with eligible Colorado nonstudy schools, by preintervention characteristic

Characteristic	Schools not in study[a]	Schools in study	*p*-value
	Mean	**Mean**	
Average 2007 CSAP mathematics achievement			
Grade 4	491	493	.18
	(78)	(75)	
Grade 5	519	525	<.01
	(78)	(76)	
Number of students per school (average)	392.94	387.94	.83
	(190.67)	(187.92)	
Number of students per teacher (average)	16.03	15.18	.17
	(10.22)	(4.28)	
Students eligible for free or reduced-price lunch (percent)	39.75	47.13	.03
	(27.59)	(26.13)	
Student population (percent)			
White, non-Hispanic	60.63	55.96	.18
	(27.70)	(27.29)	
Black, non-Hispanic	5.35	3.87	.11
	(9.29)	(7.25)	
Hispanic	29.61	37.23	.02
	(25.71)	(26.76)	
Asian/Pacific Islander	3.17	2.06	<.01
	(3.08)	(3.06)	
American Indian/Alaska Native	1.25	0.88	<.05
	(2.76)	(1.34)	
	Percent	**Percent**	
School urbanicity (percent of schools)			<.0001
City	30.99	31.43	
Suburb	32.88	7.14	
Town	10.26	17.14	
Rural	25.86	44.29	
Schools receiving Title I funding (percent of schools)			
Title I–eligible school	47.02	76.81	<.0001
Schoolwide Title I	62.81	47.17	.03

a. Includes regular schools (public schools that do not focus primarily on vocational, special, or alternative education) classified as primary in the Common Core of Data.

Note: Sample sizes were suppressed in this table because data were not available for all schools for some of the variables. Numbers in parentheses are standard deviations. Significance tests for 'math achievement' were one-sample *t*-tests using grade-level mean scores for Colorado as the comparison values. Significance tests for 'students per teacher', 'students per school', 'free-reduced lunch', and 'student population' were *t*-tests between group means. Significance tests for 'school urbanicity' and 'Title I funding' were *chi-square* tests of group frequencies. The schools that participated in the study included three sets of two schools each that were configured as grades K–4 and grades 5–8. In these three sets of schools all students in grade 4 transitioned to the grades 5–8 school. To reduce crossover, each set of schools was randomly assigned as a single school unit and was included as a single school unit in the impact analyses.

Source: The student achievement statistics were obtained by aggregating grade 4 and grade 5 achievement scores from the 2007 Colorado Student Assessment Program data; school demographic statistics were obtained from school-level data for entire schools from the 2007/08 Common Core of Data (http://nces.ed.gov/ccd/pubschuniv.asp).

Random assignment of schools

Recruited schools were randomly assigned to either the intervention group or control group within district or pseudo-district. The schools in six districts were randomly assigned within their respective districts: blocks 1 through block 6. When only one school from a district volunteered to participate in the study or the school was the only elementary school in the district, schools were blocked by pseudo-district (a group of schools similar in terms of locale, location in Colorado, and the date they volunteered to participate): blocks 7, 8, and 9. Block 9 included four schools that volunteered in late fall 2007, after the schools in blocks 7 and 8 had been randomly assigned. The largest blocks had 14 schools, and between one and thirteen districts were represented in each block. In each block, schools were first assigned a random number from a random number generator (RANUNI in SAS). Schools were then ordered by the random number. The first half of the schools were assigned to the intervention group; the second half were assigned to the control group. When the block contained an odd number of schools, the additional school was assigned to the control group.

Although schools were randomly assigned to the intervention or control group, the random assignment could yield two groups that differed by chance at baseline on some key characteristics. The intervention and control groups were compared to determine whether the groups differed on the following school-level characteristics: student mathematics achievement, urbanicity, percentage of students eligible for free or reduced-price lunch, Title I status, and percentage of students in racial/ethnic groups (table 2.3). No statistically significant differences were found between the two groups of schools on any of these characteristics.

Table 2.3. Comparison of intervention and control schools, by preintervention characteristic

Characteristic	Intervention Mean	Control Mean	Difference	Test statistic	*p*-value
Mathematics achievement[a]	508.28	505.26			
	(76.04)	(78.43)	3.02	0.41	.69
Number of students per school (average)	377.54 (190.72)	398.34 (187.26)	–20.80	0.46	.65
Number of students per teacher (average)	15.03 (4.55)	15.33 (4.05)	–.30	0.29	.78
Students eligible for free or reduced-price lunch (percent)	46.59	47.67			
	(26.86)	(25.76)	–1.08	0.17	.86
Student population (percent)					
White, non-Hispanic	56.35 (28.18)	55.56 (26.76)	0.79	0.12	.91
Black, non-Hispanic	3.67 (7.67)	4.08 (6.89)	–0.41	–0.24	.81
Hispanic	36.73	37.74			
	(27.24)	(26.65)	–1.01	–0.16	.88
Asian/ Pacific Islander	2.23 (3.72)	1.88 (2.23)	0.35	0.47	.64
American Indian/ Alaska Native	1.02 (1.73)	0.73 (0.75)	0.29	0.90	.37
	Percent	**Percent**			
School urbanicity (percent of schools)				0.27	.87
City	28.57	34.29	–5.72		
Suburb/Town[b]	25.71	22.86	2.85		
Rural	45.71	42.86	2.85		
Schools receiving Title I funding (percent of schools)					
Title I–eligible school	80.00	73.53	6.47	0.41	.52
Schoolwide Title I	42.86	52.00	–9.14	0.44	.51

a. Test statistics and *p*-values for the mathematics achievement comparisons were adjusted for clustering of students within schools.
b. Data for suburb and town locales were combined to prevent disclosure.
Note: Sample sizes were suppressed in this table because data were not available for all schools for some of the variables. Numbers in parentheses are standard deviations. Significance tests for 'mathematics achievement', 'students per teacher', 'students per school', 'free-reduced lunch', and 'student population' were *t*-tests between group means; significance tests for 'school urbanicity' and 'Title I funding' were *chi-square* tests of group frequencies.
Source: The student achievement statistics were obtained by aggregating grade 4 and grade 5 achievement scores from the 2007 Colorado Student Assessment Program data; school demographic statistics were obtained from school-level data for entire schools from the 2007/08 Common Core of Data (http://nces.ed.gov/ccd/pubschuniv.asp).

Student achievement sample

CASL was implemented in grade 4 and grade 5 classrooms during both the CASL training year and the CASL implementation year to help ensure the availability of pretest achievement data from the grade 3 CSAP administration the prior year. Pretest data are data that were used in the impact analysis to control for achievement prior to exposure to the intervention. Because the CSAP is administered in the spring of every school year, no student achievement data were

collected at study baseline, November 2007. All student achievement data were obtained directly from the state in September 2009, as described in the data collection section of this report. Grade 3 classrooms were not included in the study because achievement test data from the end of grade 3 would already reflect the potential impact of the intervention for students in grade 3 had they been included in the study.

Over the course of the study, three different cohorts of students in the study schools were administered the CSAP (table 2.4). The first cohort of students (Cohort 1) was enrolled in grade 4 during the CASL training year and in grade 5 during the CASL implementation year. Cohort 1 students in the intervention schools were exposed to CASL for two years. The second cohort of students (Cohort 2) was in grade 5 during the CASL training year. Cohort 2 students in the intervention schools were exposed to CASL for only one year and left the study schools after the training year; they were not present in the schools during the CASL implementation year. The third cohort of students (Cohort 3) was enrolled in grade 4 during the implementation year. Cohort 3 intervention students experienced CASL in the classroom only during the implementation year.

Table 2.4. Student cohorts

Grade level	CASL training year (2007/08)	CASL implementation year (2008/09)
Grade 4	Cohort 1	Cohort 3
Grade 5	Cohort 2	Cohort 1

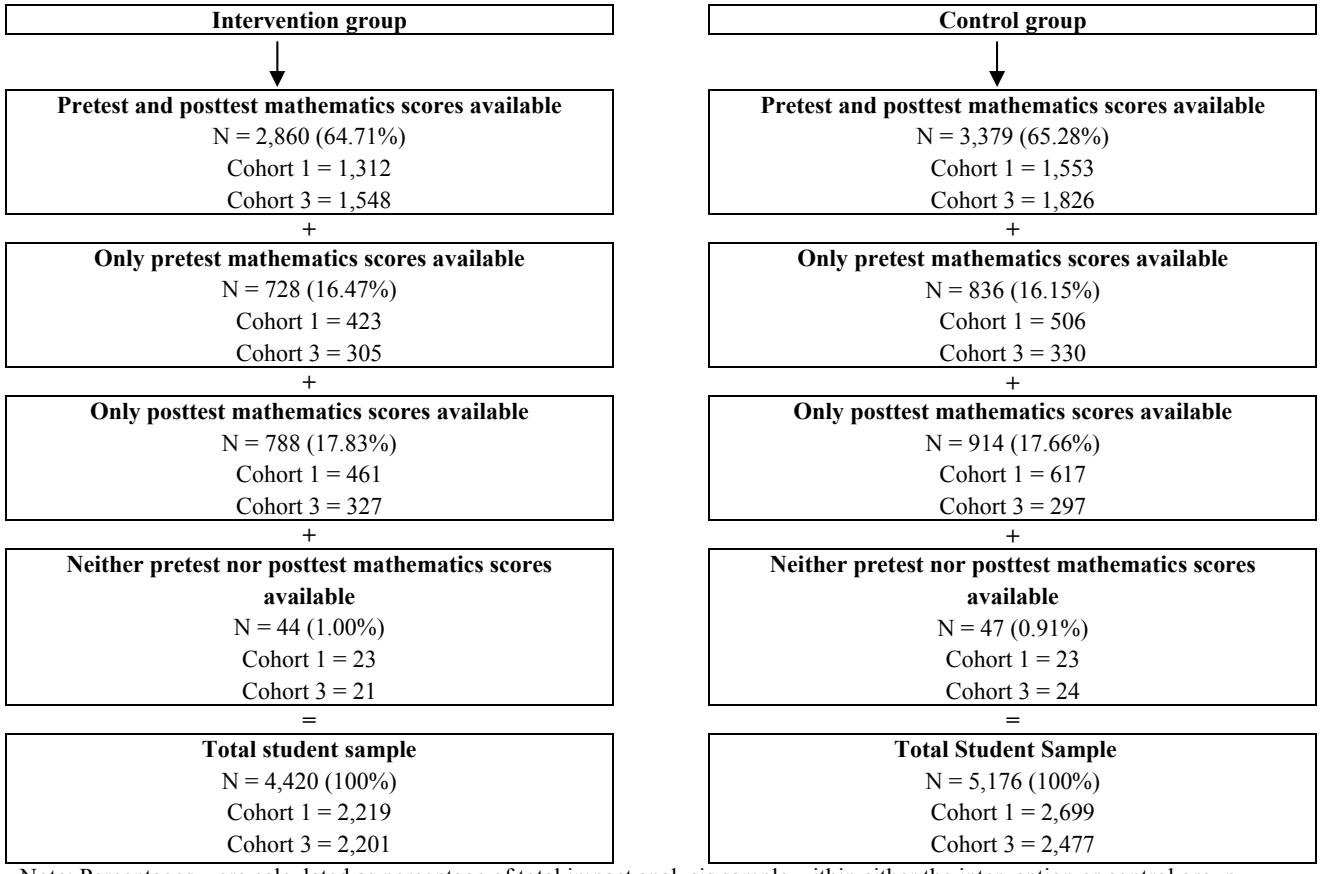
The student achievement sample was defined as students for whom CSAP data were available from the study schools for any content area (mathematics, reading, or writing) from either the pre- or posttest administration. For Cohort 1, the pretest was the grade 3 CSAP administered in the spring of 2007, and the posttest was the grade 5 CSAP administered in the spring of 2009. For Cohort 3, the pretest was the grade 3 CSAP administered in the spring of 2008, and the posttest was the grade 4 CSAP administered in the spring of 2009. Students in Cohort 2 were not included in the student achievement sample because the Cohort 2 students in the intervention schools were exposed to the intervention during the CASL training year only, had moved on to middle school before the CASL implementation year began, and were not enrolled in the study schools at the time of the April 2009 (posttest) administration of the CSAP. Less than 1 percent of students had neither pre- nor posttest mathematics scores but were included in the impact analysis sample with mathematics scores imputed from the other test score data. Students who were in the study schools at the time of random assignment but left the study schools during the course of the study were included in the study sample to help ensure that the impact analysis sample included all students as randomly assigned. Students who entered the study schools during either the CASL training year or the CASL implementation year were included as part of the student sample so that the impact analysis would test whether overall student achievement in grade 4 and grade 5 in the intervention schools improved as a result of the assignment to the intervention.

To test the impact of CASL on student achievement after a full year of implementation following the training year, Cohort 1 students (in grade 4 during the CASL training year and grade 5 during the CASL implementation year) were combined with the Cohort 3 students (in grade 3 during the

CASL training year and grade 4 during the CASL implementation year) to form the student achievement impact analysis sample. These two samples of students were analyzed together to estimate the impact of CASL at the school level even though Cohort 1 students and Cohort 3 students in the intervention group experienced different levels of exposure to CASL. Cohort 2 students were included in one sample of student motivation, described in the next section.

The universe of eligible students included those students who had CSAP scores from any content area (mathematics, reading, or writing) from either the pre- or posttest administration dates. Figure 2.1 illustrates how the student achievement impact analysis sample was constructed using available CSAP data. The pretest was the grade 3 CSAP administered in the spring of 2007 for Cohort 1 and the grade 3 CSAP administered in the spring of 2008 for Cohort 3. The posttest was the grade 5 CSAP administered in the spring of 2009 for Cohort 1 and the grade 4 CSAP administered in the spring of 2009 for Cohort 3.

Figure 2.1. Construction of the student achievement impact analysis sample: number of students in the impact sample as a function of the availability of pretest and posttest CSAP mathematics achievement scores.



Note: Percentages were calculated as percentage of total impact analysis sample within either the intervention or control group. Pretest scores for Cohort 1 were from the grade 3 2007 CSAP; pretest scores for Cohort 3 were from the grade 3 2008 CSAP. Posttest scores were from spring 2009 for both cohorts: grade 5 for Cohort 1 and grade 4 for Cohort 3. Where neither pretest nor posttest mathematics scores were available, data were imputed from available reading and/or writing scale scores.
Source: 2007, 2008, and 2009 Colorado Student Assessment Program data.

Student mobility into or out of study schools over the course of the study resulted in either missing pretest data or missing posttest data. Student achievement data also may have been missing if a student enrolled in a study school was not administered the pretest or posttest for other reasons, such as illness on the day of testing. For students included in the impact analysis sample, both pre- and posttest CSAP mathematics scale scores were available for 64.71 percent of students in the intervention group and 65.28 percent of students in the control group.

Of the intervention students tested in the study schools at baseline, 16.47 percent were not tested in the study schools at posttest, and 16.15 percent of control group students tested at baseline were not tested at posttest. In the intervention group, 19.06 percent of Cohort 1 students tested at baseline did not have posttest scores, as compared with 13.86 percent of students in Cohort 3. Conversely, posttest data were not available for 18.75 percent of Cohort 1 students in the control group and 13.32 percent of Cohort 3 students in the control group. Missing mathematics scale scores, for either pretest or posttest, were imputed (see the section below on the treatment of missing data).

Over the course of the study, student mobility also occurred between study schools. There were four possible patterns of student mobility within the study schools: moving from an intervention school to an intervention school, moving from a control school to a control school, moving from an intervention school to a control school, and moving from a control school to an intervention school (table 2.5). In cases where students moved between study schools, students were analyzed as a member of the school they attended at the time of random assignment.

Table 2.5. Within-study student mobility

Mobility patterns	Number of students	Percent of impact sample
Intervention to intervention	137	1.43
Control to control	103	1.07
Intervention to control	57	0.59
Control to intervention	78	0.81
Total	375	2.90

Source: 2007, 2008, and 2009 Colorado Student Assessment Program data.

The sample used to estimate the impact of CASL on student achievement included all 67 schools randomly assigned at the beginning of the study. To provide an unbiased estimate to CASL's impact on student achievement, all schools were included as randomly assigned, regardless of the level of CASL implementation for intervention schools, and regardless of their participation in teacher data collection. The impact analysis sample included 9,596 students, of which 51.25 percent were in Cohort 1 and 48.75 percent were in Cohort 3 (table 2.6). The 67 schools included in the impact analysis sample varied in terms of the number of students in grades 4 and 5 (table 2.7).

Table 2.6. Impact analysis sample size by grade and experimental group

	Intervention		Control		Total	
Grade level	Number	Percent	Number	Percent	Number	Percent
Grade 4	2,201	49.80	2,477	47.86	4,678	48.75
Grade 5	2,219	50.20	2,699	52.14	4,918	51.25
Total	4,420	100.00	5,176	100.00	9,596	100.00

Source: 2009 Colorado Student Assessment Program data.

Table 2.7. Number of students per school in impact analysis sample, by intervention and control group

	Intervention (Schools = 33)		Control (Schools = 34)	
Number of students per school	Grade 4	Grade 5	Grade 4	Grade 5
Mean	67	67	73	79
Standard deviation	38	40	39	40
Lowest quartile	43	39	48	52
Median	59	58	67	77
Highest quartile	90	88	93	109

Note: Numbers in this table were generated by first creating a frequency distribution of the number of students per study school by the treatment condition and grade level. Then, a descriptive statistics analysis, including means, standard deviations, first quartile, median, and third quartile, was conducted on the resulting frequency distribution by treatment condition and grade level.
Source: 2009 Colorado Student Assessment Program data.

Student motivation sample

In addition to the student achievement sample, two samples were used to test the impacts of CASL on student motivation. Student motivation data came from a survey administered by participating teachers to the students in their classrooms (see the data collection section below for a description of the survey). To reduce the data collection burden on students, the student motivation survey was administered in a posttest-only design (no pretest measures were administered). Motivation data were collected at Wave 3 (May 2008) at the end of the CASL training year and Wave 5 (May 2009, posttest) at the end of the CASL implementation year. Student responses to the motivation survey were anonymous, and student data from the two administrations could not be matched. As such, two separate samples were available to estimate the impact of the intervention on student motivation: one sample that included all students in grades 4 and 5 (Cohort 1 and Cohort 2) who were administered the motivation survey in Wave 3 (May 2008) and one sample that included all students in grades 4 and 5 (Cohort 1 and Cohort 3) who were administered the motivation survey at Wave 5 (posttest). (See table 2.1 for timeline of key activities and table 2.16 for a description of the data collection schedule.) The Wave 3 data were used to estimate the impact of CASL after the training year, and the Wave 5 (posttest) data were used to estimate the impact after the CASL implementation year.

Although all schools randomly assigned were included in the impact analysis sample for student achievement, attrition occurred at the school level for the student motivation sample. Three schools withdrew from the study at the time of study orientation, prior to baseline data collection. No student motivation surveys were administered in these three schools. Thus, the

student motivation sample was reduced to 64 schools. In addition, student motivation surveys were not obtained for all classrooms in the remaining schools. In fact, no motivation data were obtained from some schools, further reducing the number of schools in the student motivation sample. There were 57 more classrooms in the total sample at Wave 5 (posttest) than at Wave 3 (table 2.8) because a number of motivation survey packets were returned at Wave 3 by teachers who failed to identify themselves or their school.

Table 2.8. Number of schools and classrooms participating in student motivation survey

	Intervention		Control		Total	
Time point	**Schools**	**Classrooms**	**Schools**	**Classrooms**	**Schools**	**Classrooms**
Random assignment	33	150	34	175	67	325
Baseline	32	144	32	164	64	308
Wave 3 (May 2008)	28	85	27	130	55	215
Posttest (May 2009)	24	107	32	165	56	272

Source: Survey of Student Motivation.

Although there was attrition in the school samples for both the Wave 3 impact analysis and the Wave 5 (posttest) impact analysis, the impact analysis samples did not differ by statistically significant margins in their characteristics at baseline for either the Wave 3 sample of schools (table 2.9) or the Wave 5 (posttest) sample of schools (table 2.10).

Table 2.9. Baseline characteristics of intervention and control schools in Wave 3 student motivation impact analysis sample

Characteristic	Intervention Mean	Control Mean	Difference	Test statistic	*p*-value
Mathematics achievement[a]	508.27	506.70			
	(74.05)	(79.23)	1.57	0.21	.84
Number of students per school	352.43	422.93			
	(186.46)	(184.05)	–70.5	–1.44	.16
Number of students per teacher	14.99	15.61			
	(4.91)	(4.04)	–0.62	–0.53	.60
Students eligible for free or reduced-price lunch (percent)	46.91 (24.39)	45.64 (25.22)	1.27	-0.19	.85
Student population (percent)					
White, non-Hispanic	58.33	57.95			
	(27.24)	(24.27)	0.38	0.05	.96
Black, non-Hispanic	2.93	4.71			
	(6.55)	(7.42)	–1.78	–0.95	.35
Hispanic	35.40	34.56			
	(26.55)	(23.50)	0.84	0.12	.90
Asian/Pacific Islander	2.11	2.11			
	(3.66)	(2.38)	0.00	0.00	.99
American Indian/Alaska Native	1.23	0.66			
	(1.88)	(0.59)	0.57	1.52	.14
	Percent	**Percent**			
School locale (percent of schools)				1.63	.44
City	21.43	33.48	–12.05		
Suburb /Town[b]	21.43	24.14	–2.71		
Rural	57.14	41.38	15.76		
Schools receiving Title I funding (percent of schools)					
Title I–eligible school	82.14	71.43	10.71	1.08	.30
Schoolwide Title I	39.13	55.00	–15.87	0.90	.34

a. Test statistics and *p*-values were adjusted for clustering of students within schools.
b. Data for suburb and town locales were combined to prevent disclosure.
Note: Sample sizes were suppressed in this table because data were not available for all schools for some of the variables. Numbers in parentheses are standard deviations. Significance tests for 'mathematics achievement', 'students per teacher,' 'students per school,' 'free-reduced lunch,' and 'student population' were *t*-tests between group means; significance tests for 'school urbanicity' and 'Title I funding' were *chi-square* tests of group frequencies.
Source: The student achievement statistics were obtained by aggregating grade 4 and grade 5 achievement scores from the 2007 Colorado Student Assessment Program data; school demographic statistics were obtained from school-level data for entire schools from the 2007/08 Common Core of Data (http://nces.ed.gov/ccd/pubschuniv.asp).

Table 2.10. Baseline characteristics of intervention and control schools in posttest student motivation impact analysis sample

Characteristic	Intervention Mean	Control Mean	Difference	Test statistic	*p*-value
Mathematics achievement[a]	508.49	504.90	3.59	.46	.65
	(73.85)	(79.21)			
Number of students per school	372.68	396.15	–23.47	–0.47	.64
	(183.15)	(192.56)			
Number of students per teacher	15.66	15.21	0.45	0.39	.70
	(4.79)	(4.10)			
Students eligible for free or reduced-price lunch (percent)	47.44	47.41	0.03	0.00	.99
	(25.59)	(24.90)			
Student population (percent)					
White, non-Hispanic	57.64	56.10	1.54	0.21	.83
	(28.20)	(26.12)			
Black, non-Hispanic	3.02	4.16	–1.14	–0.61	.54
	(6.89)	(7.09)			
Hispanic	35.86	37.18	–1.32	–0.18	.85
	(27.41)	(26.07)			
Asian/Pacific Islander	2.31	1.94	0.37	0.42	.68
	(3.83)	(2.28)			
American Indian/Alaska Native	1.17	0.61	0.56	1.59	.12
	(1.87)	(0.59)			
	Percent	**Percent**			
School locale (percent of schools)				0.23	.89
City	28.00	33.33	–5.33		
Suburb/Town[b]	24.00	24.24	–0.24		
Rural	48.00	42.42	5.58		
Schools receiving Title I funding (percent of schools)					
Title I–eligible school	80.00	75.00	5.00	0.20	.66
Schoolwide Title I	40.00	50.00	–10.00	0.44	.51

a. Test statistics and *p*-values were adjusted for clustering of students within schools.
b. Data for suburb and town locales were combined to prevent disclosure.
Note: Sample sizes were suppressed in this table because data were not available for all schools for some of the variables. Numbers in parentheses are standard deviations. Significance tests for 'mathematics achievement', 'students per teacher', 'students per school', 'free-reduced lunch', and 'student population' were *t*-tests between group means; significance tests for 'school urbanicity' and 'Title I funding' were *chi-square* tests of group frequencies.
Source: The student achievement statistics were obtained by aggregating grade 4 and grade 5 achievement scores from the 2007 Colorado Student Assessment Program data; school demographic statistics were obtained from school-level data for entire schools from the 2007/08 Common Core of Data (http://nces.ed.gov/ccd/pubschuniv.asp).

On the day of survey administration, participating teachers administered the motivation survey to all students in their classrooms whose parents or guardians had provided passive consent for their child to complete the survey. Given attrition at the school and classroom levels and the possibility of students not receiving passive consent, the student motivation impact analysis sample was smaller than the student achievement sample (table 2.11).

Table 2.11. Comparison of student sample size between motivation impact sample and achievement data

	Intervention		Control	
Data source	**Cohort 1**	**Cohort 2**	**Cohort 1**	**Cohort 2**
Wave 3 student achievement data (May 2008)	1,830	1,773	2,091	2,197
Wave 3 student motivation data (April 2008)	662	517	1,320	1,259
	(36.17%)	(29.16%)	(63.13%)	(57.31%)
	Cohort 1	**Cohort 3**	**Cohort 1**	**Cohort 3**
Wave 5 (posttest) student achievement data (May 2009)	1,784	1,888	2,181	2,135
Wave 5 (posttest) student motivation data (April 2009)	1,045	971	1,556	1,614
	(58.58%)	(51.43%)	(71.34%)	(75.60%)

Note: Percentages are calculated as number of students with motivation data divided by the number of students with achievement data for each experimental group by grade and by wave of data collection.
Source: 2008 and 2009 Colorado Student Assessment Program data and Survey of Student Motivation.

Teacher sample

Successfully completing this study depended on collecting quality data from participating teachers, including teacher background data, outcome data (required to answer several research questions), and contextual and implementation fidelity data to inform the interpretation of results.

The teacher sample was defined as all teachers in grades 4 and 5 in the study schools who provided direct instruction in mathematics. Teachers who entered the study schools at the beginning of the CASL implementation year (late entries) also were included as part of the teacher sample, as were teachers who were in the study schools at the time of random assignment but who left the study schools during the course of the study. This definition was used to help ensure that teachers were included in the impact analysis sample as randomly assigned and that the impact analysis would test whether overall teacher outcomes in the intervention schools improved as a result of the assignment to the intervention. Late-entry teachers were provided their own copies of the primary CASL textbook, joined the learning teams in their respective schools, and participated in whatever learning team activities occurred in their schools during the implementation year. Chapter 3 provides details on the activities of intervention schools during the implementation year.

The samples of teachers in the intervention and control groups were compared on the following baseline characteristics: years of overall teaching experience, years of experience teaching mathematics, teacher education level, and teacher knowledge of classroom assessment (table 2.12). This comparison also included late-entry teachers. A statistically significant difference was found between intervention and control teachers on years of teaching experience and years of experience in teaching mathematics. Intervention group teachers had 13.01 years (standard deviation = 9.16) of teaching experience, on average, as compared to an average of 10.54 years (standard deviation = 7.95) among the control group teachers. For years teaching math, intervention teachers averaged 11.06 years (standard deviation = 8.55), whereas control group teachers averaged 8.90 years (standard deviation = 7.02). Because the intervention and control groups were formed using random assignment, these baseline differences are random errors, not biases (Bloom 2008). The correlation between years teaching and years teaching mathematics

was .53 ($p > .01$), meaning that these two variables were related but measured different aspects of teacher experience. To control for these differences in teaching experience between the intervention and control group teachers, both years of teaching experience and years of teaching mathematics were added as covariates to the analysis models of teacher impacts.

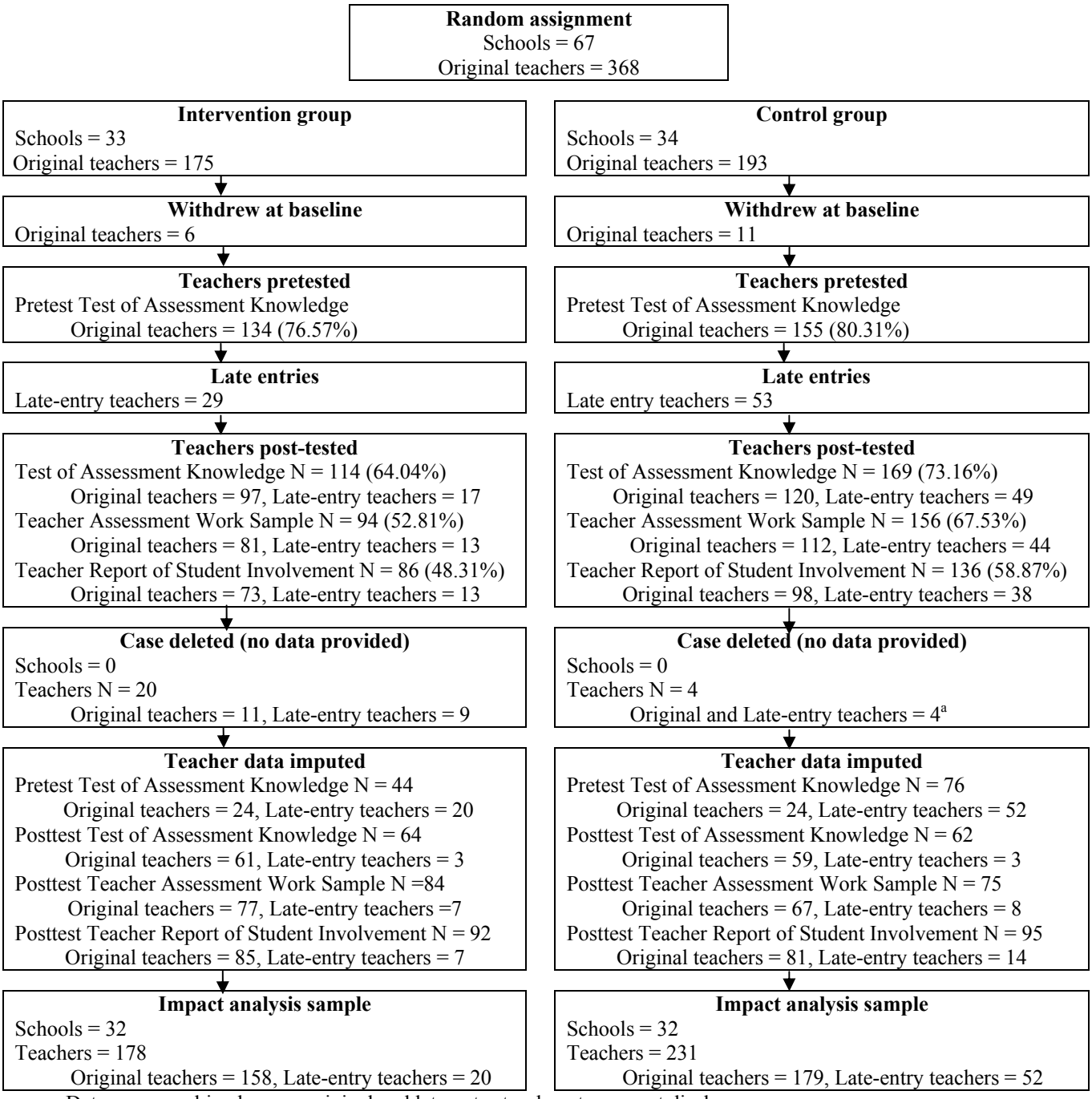
Table 2.12. Comparison of intervention and control teachers on preintervention characteristics

Characteristic	Intervention	Control	Difference	Test statistic	*p*-value
Years teaching experience					
Mean	13.01	10.54	2.47	2.12	.038
Standard deviation	9.16	7.95			
Number in sample	168	213			
Years of teaching math					
Mean	11.06	8.90	2.16	2.16	.035
Standard deviation	8.55	7.02			
Number in sample	166	212			
Percentage with master's degree or higher					
Mean	53.01	48.33	3.68	0.50	.62
Standard deviation	50.06	50.09			
Number in sample	166	209			
Score on test of assessment knowledge					
Mean	36.52	35.95	0.57	0.79	.43
Standard deviation	6.23	5.81			
Number in sample	130	154			

Note: Number in sample includes original and late-entry teachers. Some teachers did not provide all background data. Test statistics and *p*-values were adjusted for clustering of teachers within schools. Data do not include teachers from the three schools that withdrew prior to data collection.
Source: Teacher background data.

At random assignment, the teacher sample included all of the grade 4 and 5 teachers in the participating schools, for a total of 368 teachers: 175 teachers in the intervention group and 193 teachers in the control group (figure 2.2). The withdrawal of three schools from the study at the time of study orientation reduced the sample of participating teachers to 350. (Student achievement data, however, were obtained for all schools randomly assigned at the beginning of the study, including the three schools that withdrew.) At baseline, 76.57 percent of intervention teachers and 80.31 percent of control teachers completed the pretest (see appendix B for complete information on teacher response rates.) At the beginning of the CASL implementation year, 29 teachers transferred into the intervention group schools (late entries), and 53 teachers transferred into the control group schools; these late-entry teachers were invited to provide background information and complete the Wave 5 (posttest) instruments. After collecting Wave 5 (posttest) data, a single teacher impact analysis sample was created for testing the impact of the intervention on the three teacher outcomes. The impact analysis sample included all original teachers who completed the pretest and/or any of the posttests and any late-entry teachers who completed any of the posttests, for a total of 178 teachers in the intervention group and 231 teachers in the control group. Twenty teachers in the intervention group and four teachers in the control group were excluded from the impact analysis sample because they failed to provide any data. Missing pre- and posttest data were imputed (see the section on treatment of missing data below).

Figure 2.2. Flow of schools and teachers from random assignment to impact analysis



a. Data were combined across original and late-entry teachers to prevent disclosure.

Note: Percentages in parentheses are responses rates; see appendix B for complete response rate information. Does not include intermediate waves of data collection.

Source: Teacher background, pretest, and posttest data.

Three schools withdrew from the study during study orientation and prior to baseline data collection, so no teacher data were available from the teachers at these schools (figure 2.2). These three schools represent 4.5 percent of the schools randomly assigned at baseline and were excluded from the teacher impact sample through case deletion. To examine the possibility that

the loss of these three schools introduced bias into the teacher impact analysis sample, the study team compared the baseline characteristics of the schools included in the teacher impact analysis sample. No statistically significant differences of baseline characteristics were found between the intervention and control schools included in the teacher impact analysis (table 2.13).

Table 2.13. Baseline characteristics of intervention and control schools in teacher impact analysis sample

	Intervention	Control			
Characteristic	**Mean**	**Mean**	**Difference**	**Test statistic**	***p*-value**
Math achievement [a]	507.40	504.89	2.51	.33	.75
	(76.77)	(79.21)			
Number of students per school	371.48	396.15	–24.67	–0.52	.61
	(194.75)	(192.56)			
Number of students per teacher	15.24	15.21	0.03	0.03	.98
	(4.61)	(4.10)			
Students eligible for free or reduced-price lunch (percent)	48.93	47.41	1.52	0.24	.81
	(25.84)	(24.90)			
Student population (percent)					
White, non-Hispanic	54.72	56.10	–1.38	–0.20	.84
	(28.21)	(26.12)			
Black, non-Hispanic	3.85	4.16	–0.31	–0.17	.87
	(7.87)	(7.09)			
Hispanic	38.14	37.18	0.96	0.14	.89
	(27.43)	(26.07)			
Asian/Pacific Islander	2.23	1.94	0.29	0.36	.72
	(3.83)	(2.28)			
American Indian/Alaska Native	1.06	0.61	0.45	1.37	.18
	(1.77)	(0.59)			
	Percent	**Percent**			
School locale (percent of schools)				0.25	.88
City	30.30	33.33	–3.03		
Suburb/Town [b]	21.21	24.24	–3.03		
Rural	48.48	42.42	6.06		
Schools receiving Title I funding (percent of schools)					
Title I–eligible school	81.82	75.00	6.82	0.45	.50
Schoolwide Title I	44.44	50.00	–5.56	0.16	.69

a. Test statistics and *p*-values were adjusted for clustering of students within schools.
b. Data for suburb and town locales were combined to prevent disclosure.
Note: Sample sizes were suppressed in this table because data were not available for all schools for some of the variables. Numbers in parentheses are standard deviations. Significance tests for 'mathematics achievement', 'students per teacher,' 'students per school,' 'free-reduced lunch,' and 'student population' were *t*-tests between group means; significance tests for 'school urbanicity' and 'Title I funding' were *chi-square* tests of group frequencies.
Source: The student achievement statistics were obtained by aggregating grade 4 and grade 5 achievement scores from the 2007 Colorado Student Assessment Program data; school demographic statistics were obtained from school-level data for entire schools from the 2007/08 Common Core of Data (http://nces.ed.gov/ccd/pubschuniv.asp).

Attrition and nonresponse

Student attrition occurred as students transferred out of study schools and when students' CSAP scores were not available as part of the school's CSAP dataset. Student scores also may not have

been available if enrolled students were not tested as part of NCLB testing. For example, a student may have been absent on the day of testing. In the intervention group, posttest CSAP scores were not available for 17.47 percent of students who were tested in the study schools at baseline. In the control group, posttest CSAP scores were not available for 17.06 percent of students who were tested at baseline (table 2.14). Attrition rates were higher for students in Cohort 1 than for students in Cohort 3. In the intervention group, 20.10 percent of students in Cohort 1 who were tested at baseline were missing posttest scores (posttest scores for this cohort reflected two years in the study), as compared with 14.81 percent of students in Cohort 3 who were missing posttest scores that reflected one year in the study. For the control group, posttest data were not available for 19.60 percent of students in Cohort 1 and 14.29 percent of students in Cohort 3. Missing data rates between the intervention and control groups did not differ by a statistically significant margin for the total sample, Cohort 1, or Cohort 3. The levels of overall and differential attrition for the student achievement data fall within the range of attrition considered by the What Works Clearinghouse to result in acceptable levels of bias (that is, 0.05 standard deviation of the outcome measure) even under conservative assumptions (What Works Clearinghouse 2008). Using the multiple imputation procedure described in the section below on the treatment of missing data, however, resulted in a student achievement impact analysis sample that did not exclude any students because of missing pretest or missing posttest data.

Table 2.14. Available and missing posttest mathematics scores for intervention and control groups

	Intervention				Control				*p*-value
	Posttest math scores available		Posttest math scores missing		Posttest math scores available		Posttest math scores missing		
Group	Sample size	Percent	Sample size	Percent	Sample size	Percent	Sample size	Percent	
Total	3,648	82.53	772	17.47	4,293	82.94	883	17.06	.54
Cohort 1	1,773	79.90	446	20.10	2,170	80.40	529	19.60	.66
Cohort 3	1,875	85.19	326	14.81	2,123	85.71	354	14.29	.61

Note: *p*-values are from 2 x 2 chi-square tests comparing frequency counts of instruments submitted versus not submitted for treatment versus control groups.
Source: 2007, 2008, and 2009 Colorado Student Assessment Program data.

Teacher attrition occurred because teachers withdrew from the study schools or failed to provide any pretest or outcome data. This attrition response resulted in the exclusion of 41 teachers (9.11 percent of the total sample) from the impact analysis sample (table 2.15). Twenty-six teachers from the intervention group (12.75 percent) were excluded from the impact analysis sample, and fifteen teachers from the control group (6.10 percent) were excluded from the impact analysis sample.

Table 2.15. Teacher attrition by intervention and control group

Reason for attrition	Intervention	Control	Total
Dropped at baseline	6	11	15
Original and late entry teachers case deleted (no data provided)[a]	20	4	24
Total	26	15	41

a. Data for original and late-entry teachers combined to prevent disclosure.
Note: See table B1 in appendix B for more details regarding teacher response rates to all instruments at each wave.

Teacher attrition also occurred as wave nonresponse, which occurs when participants fail to complete data collection instruments at one or more waves of data collection but provide data in other waves. Wave nonresponse is different from item nonresponse, which occurs when respondents fail to respond to individual survey items (Graham, Cumsille, & Elek-Fisk 2003; Puma et al. 2009). Appendix B contains detailed information regarding the response rate for each teacher instrument and wave. Response rates varied by instrument and wave. The response rates of the intervention group teachers were lower than the response rates of the control group teachers by a statistically significant margin for 15 of 31 comparisons (see table B1 in appendix B). There were no instances where the control group response rate was lower than the intervention group response rate by a statistically significant margin.

Statistically significant differences in response rates between intervention and control group teachers ranged from 9.08 percent to 22.69 percent, with the intervention group response rate always lower than the control group response rate. The intervention teachers had the added data collection burden of responding to the CASL participant logs, an obligation the control group teachers did not have. For the teacher outcomes used in the impact analysis, the difference in response rates ranged from 9.12 percent for the Test of Assessment Knowledge to 14.72 percent for the Teacher Assessment Work Sample. The level of overall nonresponse (greater than 30 percent) and the levels of differential nonresponse are expected to yield biases greater than 0.05 standard deviation in the outcome and require the establishment of baseline equivalence of the analysis sample in order to warrant a rating of "meets evidence standards with reservations" (What Works Clearinghouse 2008).

Using the multiple imputation procedure described in the section below on treatment of missing data, however, resulted in an impact analysis sample that excluded 26 teachers from the intervention group (6 for withdrew at baseline, 20 for no data provided) and 15 from the control group (11 for withdrew at baseline, 4 for no data provided). This resulted in overall attrition of 9.11 percent and differential attrition of 6.65 percent for the impact analysis sample. These levels of attrition are expected to result in bias of less than 0.05 standard deviation in the outcome even under conservative assumptions and can warrant a rating of "meets evidence standards" (What Works Clearinghouse 2008).

The section below on data collection discusses response rates for each instrument in the study. The treatment of wave nonresponse (that is, missing pre- and posttest data) is described in the section below on treatment of missing data.

Data collection

Data were collected throughout the course of the study to describe the fidelity of CASL implementation and the larger professional development context in the study schools and to estimate the impacts of CASL on the student and teacher outcomes. Student achievement data were obtained directly from the Colorado Department of Education for the CSAP administered in April of each year; survey data were collected from teachers and students in the participating schools (table 2.16). Teacher baseline data were used as field test data for the Teacher assessment work sample (described in this section), baseline comparisons of the intervention and control groups, and as covariates in the teacher impact analyses. Response rates for each

instrument by wave are included in appendix B. Copies of the data collection instruments are included in appendix C.

Table 2.16. Data collection schedule

Instrument	Source	Baseline November 2007	Wave 2 December 2007	Wave 3 May 2008	Wave 4 December 2008	Wave 5 (posttest) May 2009
Teacher background information	CRESST and REL Central	X				
Survey of Professional Development Activities	REL Central		X	X	X	X
CASL participant log [a]	REL Central		I	I		
Test of Assessment Knowledge	REL Central	X		X		X
Teacher Assessment Work sample	CRESST	FT				X
Teacher Report of Student Involvement	REL Central		X	X	X	X
Student Survey of Motivation	RAPS-SE & PALS			X		X

X = Instrument completed by intervention and control teachers.
I = Instrument completed by intervention teachers only.
FT = Field test.
CRESST is the National Center for Research on Evaluation, Standards, and Student Testing. REL Central is the Central Regional Educational Laboratory.
a. Completed each time the intervention teachers completed a chapter of the CASL textbook.

At the study orientation, conducted in the fall of 2007, the study team provided all participating teachers with the full data collection schedule and the two-week response window for each data collection wave. This advance schedule was intended to help participants understand the data collection commitment and to plan for and incorporate the data collections into their schedules. Prior to each wave of data collection, the study team conducted site visits to each participating school to provide instructions for each wave of data collection and deliver paper-pencil instruments. Only the Teacher Assessment Work Sample and the Survey of Student Motivation were paper-pencil instruments; all other teacher instruments were administered online. At the beginning of each wave of data collection, teachers received an email message providing them with a link to the data collection instrument and a requested timeline for completion. Email reminders to complete instruments were sent to each nonrespondent until the response was received, the participant requested not to be contracted further, the data collection window closed, or the school closed for the year.

Teacher background data

Information on teacher background characteristics was collected at baseline (November 2007). Teachers were asked to provide background information (including gender, race/ethnicity, and years of teaching experience) through an online survey. Category definitions for race/ethnicity were consistent with Office of Management and Budget statistical classifications. Late entry teachers who transferred into the study schools at the beginning of the CASL implementation year also were asked to complete the survey. The total response rate for this survey was 86.14 percent for original sample teachers and 74.39 percent for late entry teachers. Original intervention group teachers' response rate was 86.00 percent, compared with 88.08 percent for

original control group teachers, and the response rate for late-entry intervention group teachers was 62.07 percent, compared with 81.13 percent for late-entry control group teachers.

Fidelity and contextual data

Teachers assigned to the intervention group were asked to document their study experiences using the CASL participant logs, and all participating teachers were asked to document their professional development activities with the Survey of Professional Development.

CASL participant logs. Recordkeeping and reflection are theoretically sound and common practices when teachers are engaged in professional development (York-Barr, Sommers, Ghere, & Montie 2001). Teachers in the intervention group completed brief logs describing their study of the CASL materials during the CASL training year. These logs were developed for this study to provide data on implementation fidelity and were not part of the CASL materials delivered to intervention group teachers. The purpose of these logs was to capture the degree to which teachers in the intervention group implemented the CASL program according to the developers' recommendations. Teachers completed an initial log that included questions regarding the establishment of their learning team. Teachers also were asked to complete a log after completing each of the 13 chapters of the CASL textbook, reporting on each segment of their work with CASL while it was still fresh in their minds and thereby improving the quality of data collected. Log items focused on the amount of the chapter read, the activities completed for each chapter, the learning team meetings and attendance at those meetings, and the total number of hours spent training with CASL for each chapter. Each chapter log was administered online, contained approximately nine items, and took less than 10 minutes to complete. Data from the intervention schools' teacher log entries describing variations in the implementation of the CASL program are reported in chapter 3. Response rates across the 14 logs ranged from 69 percent to 79 percent. In addition, during the final site visits conducted in preparation for Wave 5 (posttest) data collection, teachers described how they had implemented CASL in the classroom during the CASL implementation year.

Survey of professional development. The purpose of the Survey of Professional Development was to gather contextual information regarding all teachers' professional development activities—outside of the CASL program—over the course of the study. Professional development was defined broadly to include any workshop, institute, conference, college course, teacher network, internship or immersion activity, teacher committee or task force, teacher study group, or work with a mentor or coach outside of the CASL program activities. All teachers, both intervention and control, were asked to complete the survey at the end of each semester (Wave 2, Wave 3, Wave 4, and Wave 5). The survey collected information on non-CASL professional development activities that occurred during that semester; the Wave 2 and Wave 4 administrations also included summer professional development activities. The survey directions instructed teachers not to include CASL professional development activities in their responses to this survey. Teachers recorded the types of professional development activities that they participated in, their frequency, duration, subject area, emphasis, perceived quality, and perceived impact on classroom practice. Any control group teacher reports of participation in professional development activities originating from the CASL developer or including CASL materials could represent crossover. The survey was administered online, included

approximately 17 items, and took less than 10 minutes to complete. Data from the Survey of Professional Development are presented in chapter 3. Total response rates ranged from 69.44 percent to 79.62 percent across the four administrations; intervention teachers' response rates ranged from 64.61 percent to 74.86 percent; control teachers' response rates ranged from 83.94 percent to 73.59 percent.

Outcome data

To the extent possible, study outcomes were measured using existing instruments with documented reliability and validity. Instruments were selected with construct validity in mind and were adapted as necessary to help ensure sensitivity to the intervention. No existing instruments thought to be appropriate to the teacher outcomes existed at the beginning of the study. Instruments to assess teacher outcomes, therefore, were adapted or developed. (See Appendix D for details on the development, reliability, and validity of the three teacher outcomes.)

Student achievement data. The study was conducted in Colorado, where the CSAP serves as the statewide NCLB assessment. Students' scale scores on the 2009 administration of the mathematics portion of the CSAP were used to estimate the impact of CASL on student achievement. Using the state assessment as the measure of the primary outcome—student achievement—allowed for estimation of the extent to which implementation of CASL impacted student achievement in relation to the goals of NCLB. In addition, students in participating schools were thought to be more likely to be motivated to perform well on the statewide assessment than on a separate achievement test administered solely for the purposes of this study.

The items on the CSAP are developed to cover the Colorado Model Content Standards and Assessment Framework and are reviewed by content review committees comprised of Colorado teachers, community members, and Colorado Department of Education department staff (CTB/McGraw-Hill 2009). As a result, the study team decided that the CSAP was the achievement test most likely to be aligned with the instructional content taught across all the study schools.

The 2009 CSAP mathematics test provides scale scores on a vertical scale from grades 3–8 through an anchor item design that horizontally equates each grade level to the vertical scale (CTB/McGraw-Hill 2009). The internal consistency of both the grade 4 mathematics test and the grade 5 mathematics test was .94 (CTB/McGraw-Hill 2009). The 2009 CSAP mathematics tests also showed similar internal consistency across NCLB subgroups; alpha coefficients ranged from .84 to .95 across the various subgroups in both grades 4 and 5 (CTB/McGraw-Hill 2009).

Although mathematics was the focus of the analysis, scores from all content areas were requested and obtained to aid in imputing missing data. Other demographic data on students—such as eligibility for free or reduced-price lunch and disability status—were requested and obtained in the achievement data files.

The impact of CASL on the primary outcome of student achievement was estimated using the student scale scores on the mathematics portion of the spring 2009 CSAP. CSAP scale scores were also used as pretest covariates in the analysis. For Cohort 1 students, who were in grade 3 in spring 2007, mathematics scale scores from the spring 2007 CSAP administration were used to control for their prior achievement. For Cohort 3, student scores from the CSAP mathematics test administered in spring 2008, when these students were in grade 3, served as the control for their prior achievement. Cohort 2 students, who were in grade 5 in spring 2008, were not included in the impact analyses because they were not enrolled in the study schools in spring 2009.

Student motivation. The developers of CASL claim that the intervention impacts student achievement through changes in student motivation, in that sound classroom assessment practices increase both student motivation to learn and their engagement in learning activities. The Survey of Student Motivation was used to collect data to estimate the impact of CASL on student motivation. The Survey of Student Motivation was comprised of the Ongoing Engagement and Perceived Autonomy (Self-Regulation) subscales of the elementary student Research Assessment Package for Schools (RAPS-SE; IRRE 1998) along with the Academic Efficacy subscale of the Patterns of Adapted Learning Scales (PALS; Midgely, Maehr, Hruda, Anderman, Anderman, Freeman, et al. 2000). These student motivation instruments have been used extensively in education research, with results published in peer-reviewed journals. The RAPS-SE was tested for reliability and validity with students in grades 4 and 5, the same age targeted in the present study. The reported alpha for engagement in the RAPS-SE was .66, and the reliabilities for perceived autonomy were .78 and .80 (IRRE 1998). The Academic Efficacy subscale of the PALS was validated with students in grade 5; the reported reliability coefficient alpha for this subscale was .78 (Midgely et al. 2000).

Items from both motivation instruments were adapted by the study team to refer specifically to mathematics to be consistent with the focus on mathematics and with the theory that self-efficacy (Bandura 1997; Pajares 1996), engagement, and perceived autonomy (Ryan & Deci 2000) are domain-specific. The words "in math" or "math class" were substituted for words such as "in school" or "in class." In the PALS manual, Midgely et al. (2000) state that the middle and high school versions of the instrument were adapted to be domain-specific; they did not originally do this for elementary schools only because elementary students typically learn different subjects in the same classroom with the same teacher. Midgely et al. (2000) report that the alpha coefficients for the domain-specific scales were as high as or higher than the general scale, which is expected because of the domain-specific nature of the constructs. Making the scales domain-specific should, therefore, increase their reliability. The alpha coefficients for the survey used in this study were .90 for the Wave 3 data and the Wave 5 (posttest) data.

The Survey of Student Motivation included 20 items and was administered at the end of each year (Wave 3, Wave 5) only to reduce the data collection burden on students. The RAPS-SE and the PALS were on a Likert-scale format in which students responded to numeric score categories from 1 to 4. The mean score across all 20 items was used as the measure of student academic motivation in the impact analysis. Motivation surveys were received from 58.42 percent of classrooms at Wave 3 (48.57 percent of intervention group classrooms and 67.36 percent of control group classrooms) and 66.50 percent of classrooms at Wave 5 (posttest) (60.11 percent of

intervention group classrooms and 71.86 percent of control group classrooms). (See the Attrition and nonresponse section above for information regarding the implications of response rates and attrition for interpreting findings.)

Teacher assessment knowledge. Teacher knowledge of classroom assessment was measured using a test designed to be sensitive to the intervention but not overaligned to the content of the CASL program. Few instruments were available for measuring teacher knowledge of classroom assessment (for example, Mertler & Campbell 2005; Mertler 2009; Plake, Impara, & Fager 1993; Zhang & Burry-Stock 2003), and the available instruments tended to either cover a broader range of topics than presented in CASL or measure perceived ability in classroom assessment rather than actual knowledge of classroom assessment. An established and proven instrument that was sufficiently well aligned with the construct under investigation did not exist at the beginning of this study. The lack of alignment between existing instruments and the construct under study was determined to be a threat to construct validity in that the existing instruments were not sensitive to the CASL training program or its impact. Although one existing instrument contained some relevant subscales (Mertler & Campbell 2005), these subscales sampled only a portion of the content domain under investigation. Using only this subscale would improperly sample from the domain of CASL.

The study team developed a test to help ensure that the measure of teacher knowledge of assessment was sensitive to the intervention impact on this domain. The Test of Assessment Knowledge included multiple-choice, true/false, and matching items. The test items covered teacher knowledge of, and reasoning skills regarding, generally accepted principles and practices of classroom assessment. The test gave more weight to topics that are described in depth by the CASL program. Although the test was designed to be sensitive to the CASL program, steps were taken to ensure that the test was not overaligned; it did not include materials, text, idiosyncratic wording, or terminology from the CASL program or program materials. The test used common language and sampled from the general domain of classroom assessment. (See appendix D for details regarding the development, reliability, and validity of this instrument.)

The 60-item test was administered online and took approximately 40 minutes to complete. The test was administered three times over the course of the study. The first administration provided pretest data on teacher assessment knowledge and was used to check for group equivalence at baseline on teacher knowledge of assessment (see table 2.4). Data from the test administered at baseline also was used as a pretest covariate in the impact analysis on all three teacher outcomes. Scores from the Test of Assessment Knowledge administered at baseline could not be influenced by the content of the intervention; thus, any risk of possible overalignment of the measure to the intervention, however small, could not invalidate the use of this measure as a pretest covariate. The Wave 5 (posttest) administration was used to measure sustained gains in teacher assessment knowledge after the CASL implementation year. The score used in the impact analyses was the number of correct responses to the 60-item test from the Wave 5 (posttest) administration. Items missing a response were treated as incorrect. If a teacher omitted more than 10 percent of the items, the total score was treated as missing. Total response rates ranged from 69.19 percent to 78.53 percent across the three administrations. The Wave 5 (posttest) response rate was 64.04 percent for intervention group teachers and 73.16 percent for control group teachers. (See the

attrition and nonresponse section above for information regarding the implications of response rates and attrition for interpreting findings.)

Teacher assessment practice. An important teacher outcome for this study was classroom assessment practice. Classroom assessment practice includes clear communication of learning targets; provision of accurate assessment results for use by students, teachers, and parents; and feedback to students that describes strengths and needs in relation to learning targets. Systematically collecting samples of graded student work can provide an efficient way to understand what is happening in classrooms (Matsumura, Garnier, Slater, & Boston 2008). Samples of graded student work directly capture teacher thinking and classroom assessment practice as opposed to the snapshot provided by a classroom visit and observation, during which there is no guarantee that assessment will even occur.

The measure of teacher assessment practice for this study was adapted by the study team from a work sample instrument developed at the National Center for Research on Evaluation, Standards, and Student Testing (CRESST). The original CRESST instrument consisted of an elementary and secondary language arts assignment rating system (Matsumura, Patthey-Chavez, Valdés, & Garnier 2002). The CRESST instrument has been researched extensively and has been demonstrated to be a valid method for measuring classroom practice (see, for example, Aschbacher 1999; Clare 2000; Clare, Valdés, Pascal, & Steinberg 2001; Matsumura, Patthey-Chavez, et al. 2002; Matsumura et al. 2008).

The Assessment Work Sample consisted of written instructions and a rubric. Packets containing the instructions and envelopes for collecting and returning artifacts were delivered to participating teachers prior to data collection. Per the written instructions, teachers collected and submitted artifacts for three types of assessment (homework/seatwork, quiz, and performance assessment); artifacts included four graded student papers and a cover sheet describing the goal of the assessment and the assessment method for each of the three types of assessment. Although this measure relied on teachers self-selecting the samples of their assessment for submission to the study team, all instructions were identical between the intervention and control teachers, making it unlikely that the instrument could introduce any bias between intervention and control groups. Research on this instrument has found that the samples submitted by teachers in this way provide a valid measure of classroom practice (Aschbacher 1999; Clare 2000; Clare, Valdés, Pascal, & Steinberg 2001; Matsumura, Patthey-Chavez, et al. 2002; Matsumura et al. 2008). The Teacher Assessment Work Sample rubric was used to score the samples on six dimensions of assessment quality: focus of goals on student learning, alignment of learning goals and task, alignment of learning goals and assessment criteria, clarity of the assessment criteria for students, type of feedback, and feedback eliciting student involvement. Details on the process for collecting and scoring the Teacher Assessment Work Sample are provided in appendix E.

The work samples were scored independently by two raters blind to the teachers' experimental group membership, and the final score was the average of each rater's score. Inter-rater reliability coefficients for the Wave 5 (posttest) work samples calculated as the correlation between rater scores by dimension and assessment type ranged from .66 to .82. Approximately 64.95 percent of teachers provided a work sample for the field test collected at baseline, and 61.12 percent provided a sample for the teacher assessment practice outcome at Wave 5. The

Wave 5 (posttest) response rate was 52.81 percent for the intervention group teachers and 67.53 percent for the control group teachers. (See the attrition and nonresponse section above for information regarding the implications of response rates and attrition for interpreting findings.)

Teacher reports of student involvement in assessment. The CASL program aims to improve student achievement by increasing student involvement in classroom assessment. Including this outcome was necessary to assess all the hypothesized impacts of CASL. Teachers in both the intervention and control schools were asked to complete the survey at the end of each semester (Wave 2, Wave 3, Wave 4, and Wave 5) in order to measure the frequency with which they involved most or all of their students in a number of activities related to classroom assessment. Teachers were asked to record the number of days in the previous two-week period during which they involved their students in assessment-related activities. Teachers also were asked to record the total number of instructional days for that two-week period. The survey included 14 items addressing activities such as discussing the learning objectives, evaluating their own work using scoring guides or rubrics, and revising work to correct errors. The survey was administered online and required approximately 10 minutes to complete.

The teacher reports of student involvement in assessment are self-report data. These teacher self-reports are limited by teachers' accuracy in recalling the extent to which they involved their students in assessment during the specified time period. Teachers may over-report student involvement if they believe there is a social bias toward these activities.

The research team took steps to minimize the limitations of self-report data for this instrument. One limitation of the teacher self-reports about teaching practice is that teachers might have different interpretations of involvement, especially if the survey used a scale with responses categories of "almost never," "sometimes," "often," and "almost always." The research team minimized this limitation by asking teachers to report the number of days during a specific period. This technique reduces the risk of different interpretations. In addition, the specific period allows teachers to use lesson plans to refresh their memories regarding their instructional activities.

A frequently-used alternative to self report is observations. While observations may address the limitation of self-reports, they are more costly and require much more frequent data collection to produce reliable estimates of instructional activities (Rowan, Camburn, & Correnti, 2004). Observational data is also subject to bias, for example, when a teacher alters his or her instruction, consciously or unconsciously, when an outside observer is present.

It should also be noted that this instrument and its administration were identical between the intervention and control groups. It is unlikely that limitations of self-reports would differentially impact the two groups, thus making comparisons between the two groups—the primary purpose of this study—unlikely to be adversely impacted by the instrument.

Each item from the survey was recoded by dividing the item score (number of days for the activity) by the total number of instructional days so that the recoded score reflected the percentage of instructional days during which teachers involved their students in each respective classroom assessment activity. This was done to account for the possibility that the teachers'

two-week reporting period may have included fewer than 10 instructional days because of in-service or other noninstructional time. The total score on the survey was the mean percentage across all 14 items. Total response rates ranged from 54.28 percent to 72.01 percent across the four administrations, although intervention group teachers' response rates were between approximately 10–17 percent lower than control group teachers' response rates. (See the Attrition and nonresponse section above for information regarding the implications of response rates and attrition for interpreting findings.)

Data analysis methods

This section discusses the impact estimation method, sensitivity analysis, treatment of missing data, and multiple hypothesis testing.

Impact estimation method

The primary purpose of the analysis was to provide an unbiased estimate of the impact of CASL on student achievement in mathematics. Consistent with the random assignment of schools to either the intervention or control group, the impact was estimated at the school level, using a mixed-model approach to account for the sources of variability in the data that resulted from the nested structure of the school environment. Variance components at the student and school levels were estimated to confirm the assumption of the nested structure of the data.

A two-level model in which students were nested within schools was used to estimate the impact of CASL on student mathematics achievement. Classroom-level teacher effects were not included in the analysis of student achievement for several reasons. First, the intervention was implemented over the course of two years, so students were exposed to two teachers over the course of the study. Students, therefore, were not nested in the same classroom during the CASL training year and the CASL implementation year. Second, the student achievement data did not include information linking students to teachers or any information identifying students (other than an encrypted identification number), so grouping students by classroom was not possible.

Although the data for this study could be conceived as containing three levels—students nested in classrooms and classrooms nested in schools—not including the intermediate level likely has little consequence for estimation of the impact of CASL. First, the variance components at the intermediate level (in this case, classrooms) is not lost but is distributed over the lower and upper levels (Moerbeek 2004). In addition, not including the intermediate level does not have an effect on statistical power if the variable of interest is at the top level (Moerbeek 2004), as in the case of the analytic models used in this study. Finally, simulation studies have found that clustering within intermediate units has little effect on Type I error (Murray, Hannan, and Baker 1996).

Level 1 of the model was specified as follows:

$$Y_{ij} = \beta_{0j} + \beta_{1j}(\mathrm{GRADE}_{ij} - \overline{\mathrm{GRADE}}_{..})_{ij} + \beta_{2j}(\mathrm{CSAP}_{ij} - \overline{\mathrm{CSAP}}_{.j})_{ij} + e_{ij}$$

where Y_{ij} is achievement outcome (mathematics scale score on the spring 2009 CSAP for student *i* in school *j*, β_{0j} is the adjusted mean achievement outcome for students in school *j*, and β_{1j} is the adjusted difference in achievement outcome due to student's grade, where GRADE was coded as 1 for students in grade 4 and 0 for students in grade 5 and centered on the grand mean so that the intercept for school *j* is adjusted for the proportion of students in grades 4 and 5 in the entire study sample (Enders & Tofighi 2007; Raudenbush & Bryk 2002). This method was chosen so that the model would provide the estimated impact and a test of statistical significance of the impact at the overall school level to address the primary research question on the impact of CASL on student mathematics achievement. It should be noted that with GRADE grand-mean-centered, β_{1j} does not provide an interpretable parameter estimate because it blends within and between cluster associations between grade-level and student achievement (Enders & Tofighi 2007; Raudenbush & Bryk 2002). β_{2j} is the within-school association between pretest and posttest, controlling for student grade level, and e_{ij} is the random error in the achievement outcome associated with student *i* in school *j*.

Level 2 of the model was specified as follows:

$$\beta_{0j} = \gamma_{01}(\text{CASL})_j + \gamma_{02}(\overline{\text{CSAP}}_{.j} - \overline{\text{CSAP}}_{..})_j + \gamma_{03}(\text{BLOCK 1}) + \gamma_{04}(\text{BLOCK 2}) + \gamma_{05}(\text{BLOCK 3}) + \gamma_{06}(\text{BLOCK 4}) + \gamma_{07}(\text{BLOCK 5}) + \gamma_{08}(\text{BLOCK 6}) + \gamma_{09}(\text{BLOCK 7}) + \gamma_{010}(\text{BLOCK 8}) + \gamma_{011}(\text{BLOCK 9}) + u_{0j}$$

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

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

where γ_{01} is the adjusted mean difference in the achievement outcome between schools assigned to the intervention group and schools assigned to the control group, CASL is an indicator variable for the intervention coded as 1 for schools randomly assigned to the intervention and 0 for schools randomly assigned to the control group, γ_{02} is the regression slope of the school level pretest (grand-mean-centered) to explain additional between-school variance not explained in level 1 of the model, γ_{03} through γ_{011} are the additive effects of each block used in the random assignment of schools, u_{0j} is the random error associated with school *j*'s adjusted average on the achievement outcome, γ_{10} is the average regression slope for student grade, and γ_{20} is the average regression slope of the student pretest.

As previously noted, schools were blocked by district and then randomly assigned within each block. An indicator variable was included for every block, requiring the suppression of the intercept in level 2 of the model. This is a no-intercept model for computing impact estimates when units were blocked and randomized within block (Bloom 2008).

As modeled above, only the school-adjusted average on the achievement outcome (that is, school intercept) was allowed to vary randomly across schools. This is the random intercept model (Raudenbush & Bryk 2002) and was used to estimate the variance in the intercept across schools.

The regression slope for student grade level and the regression slope for student pretest were fixed across schools.

The full mixed model used to estimate the impact of CASL on student achievement was specified as follows:

$$Y_{ij} = \gamma_{01}(\text{CASL})_j + \gamma_{02}(\overline{\text{CSAP}}_{.j} - \overline{\text{CSAP}}_{..})_j + \gamma_{03}(\text{Block 1}) + \gamma_{04}(\text{Block 2}) + \gamma_{05}(\text{Block 3}) + \gamma_{06}(\text{Block 4}) + \gamma_{07}(\text{Block 5}) + \gamma_{08}(\text{Block 6}) + \gamma_{09}(\text{Block 7}) + \gamma_{010}(\text{Block 8}) + \gamma_{011}(\text{Block 9}) + \gamma_{10}(\text{GRADE}_{ij} - \overline{\text{GRADE}}_{..})_{ij} + \gamma_{20}(\text{CSAP}_{ij} - \overline{\text{CSAP}}_{.j})_{ij} + e_{ij} + u_{0j}$$

Given this specification, the primary parameter of interest was γ_{01}, which can be interpreted as the adjusted school mean difference between intervention and control groups—that is, the estimated impact of CASL on school average mathematics achievement.

As described in the section above on the student achievement sample, the students in Cohorts 1 and the students in Cohort 3 had different levels of exposure to the intervention and different intervals between the pretest measure of student achievement and the end of the CASL implementation year. Thus, it was possible that the intervention had a differential impact on student achievement for these two cohorts of students. To test for an intervention impact that may have occurred in only one of the cohorts, follow-up exploratory analyses were conducted separately for grade 4 and grade 5 students where the benchmark impact model was fit separately to the data for the Cohort 1 students and to the data for the Cohort 3 students.

Models for estimating the impact of CASL on student motivation and teacher outcomes were similar to the model used to estimate the impact on student achievement: they were developed to provide an average impact estimate at the school level. (Appendix F provides details regarding the models used to estimate CASL impacts on intermediate outcomes.)

Two-tailed tests ($p < .05$) were conducted to assess the statistical significance of the impact estimates for the following reasons. First, although research suggests that the use of sound classroom assessment practices can have large positive impacts on student achievement (for example, Black & Wiliam 1998a), rigorous evidence regarding the impact of CASL was not available. In addition, activities within the control group may have been as effective as, or more effective than, CASL in raising student achievement.

In addition, the study team calculated effect sizes to describe the impact of CASL. First, 95 percent confidence intervals were calculated and presented to show the range within which the estimated impact is likely to lie. Second, effect sizes were calculated to place the intervention impacts on a scale that can be compared across outcome measures and other studies. (See appendix G for details).

Sensitivity analyses

Sensitivity analyses were conducted to test the robustness of the findings for student achievement to methodological decisions regarding the use of achievement as a pretest covariate, the estimation methods, and the method for dealing with missing data. In other words, the sensitivity analyses test whether the impact findings were affected by analytic decisions made by the researcher or by the analytic methods used.

First, sensitivity analyses were conducted to test the robustness of the benchmark impact estimates to the inclusion of covariates. Results of this noncovariate analysis are presented in conjunction with the benchmark analyses described above.

Second, sensitivity analyses were conducted to test the robustness of the benchmark impact estimates to the estimation method. The benchmark impact estimates for the primary outcome were estimated using SAS PROC MIXED and its default estimate method of residual (restricted) maximum likelihood. Sensitivity analyses were conducted using the maximum likelihood estimation method and the minimum variance quadratic unbiased estimation of the covariance parameters estimation method. Both these sensitivity analyses used the benchmark impact analysis model described in the section above on the impact estimation method and in appendix F.

Third, sensitivity analyses were conducted to test the robustness of benchmark impact estimates to the methods used to handle missing data. A sensitivity analysis was conducted in which the benchmark impact model described above was fit to data where students with missing pre- or posttest data were deleted. This sensitivity analysis used the benchmark impact analysis model described in the section above on the impact estimation method and in appendix F.

Treatment of missing data

As mentioned, three schools that withdrew from the study prior to baseline data collection were excluded from the analyses of impacts on teacher outcomes and student motivation (see figure 2.1). The case deletion method has been recommended when data are missing for entire schools (Puma et al. 2009). In addition, of the entire sample of teachers, 41 teachers (26 intervention and 15 control) were excluded from the impact analyses because they either withdrew from the study at baseline or failed to provide any data.

Missing data also occurred as either item nonresponse, when participants failed to respond to individual items, or as attrition (wave nonresponse), when participants failed to respond at one or more waves and may or may not have reappeared in a later wave (Graham, Cumsille, & Elek-Fisk 2003; Puma et al. 2009). (See appendix B for information regarding instrument response rates.)

To determine whether attrition resulted in an impact analysis sample that differed on preintervention characteristics, the sample of treatment and control schools used in the impact analysis were compared on baseline characteristics. As described in the sections above on

student motivation sample and teacher sample, no statistically significant differences were found between the schools included in the impact analysis samples.

According to Graham (2009), the inclusion of covariates in the missing data models is probably the single best strategy for reducing bias. The missing data models used to impute student achievement and teacher outcomes for this study included covariates. In addition, Puma et al. (2009) found that using pretest scores as regression covariates reduced bias in impact estimation. Pretest covariates were used in both the student achievement and teacher outcome impact analysis models.

The study team used a multiple imputation approach to impute values for missing data. Model- and data-based multiple imputation procedures have been recommended for situations in which missing data exceed 5 percent, and complete case analysis has been endorsed only when missing data do not exceed 5 percent (Graham, Cumsille, & Elek-Fisk 2003). In a recent study, multiple imputation performed well in a variety of missing data scenarios and was recommended as an approach to address missing pretest data, covariate data, and posttest data (Puma et al. 2009). Using pretest data as covariates, the multiple imputation approach produced impact estimates with biases lower than the What Works Clearinghouse standard of 0.05 standard deviation (Puma et al. 2009).

The expectation maximization algorithm with multiple imputation was used to impute missing scale score data for the student achievement impact analysis, item data for the student motivation impact analysis, and missing summary score data for the teacher outcomes impact analysis. The expectation maximization algorithm is an iterative statistical method that replaces missing data with imputed data (Puma et al. 2009). The imputed data are predicted from the relationships between the variables, calculated using all available observations (either students or teachers in this study), including observations with missing data. The imputed values also include some random error to ensure that the replacement process does not incorrectly reduce the natural variation in the data. Missing data imputations were conducted separately for intervention and control groups. The expectation maximization algorithm was implemented multiple times to result in 40 imputed datasets for the student achievement data, 40 imputed datasets for the student motivation data, and 40 imputed datasets for the teacher outcome data. Impact estimation was then conducted on the 40 imputed datasets produced by the expectation maximization method, and impact estimates were combined across the datasets. (See appendix G for more details regarding the treatment of missing data.)

Multiple hypothesis testing

This study was designed to estimate the impact of CASL on the primary outcome of student achievement in mathematics. Exploratory analyses were conducted to estimate the impact of CASL on the intermediate outcomes of student motivation and teacher outcomes to provide contextual information and support the interpretation of the impact on student achievement. Schochet (2008a) recommends that corrections for multiple hypothesis testing should be considered only for the primary outcomes of a study. Because this study had only one primary outcome, no corrections for multiple hypothesis testing were conducted.

Chapter 3. Implementation of the intervention

This chapter describes the Classroom Assessment for Student Learning (CASL) intervention and how it was implemented during this study. During the CASL training year, intervention teachers worked through the CASL materials in learning teams. Because it typically takes a full year for teachers to complete the training, the study included a second year, the CASL implementation year, in which the intervention teachers were asked to fully implement CASL techniques in their classrooms.

This chapter begins with a presentation of CASL as designed and implemented, which is followed by the results for and a discussion of the measures of adequate implementation fidelity in the training year, as recommended by the CASL developers. This chapter also describes the activities that intervention teachers implemented during the CASL implementation year. Finally, the chapter describes the non-CASL professional development received by teachers in both groups.

Classroom Assessment for Student Learning as designed and implemented

This section discusses the program materials, learning teams, introductory videoconference, and district facilitators.

Program materials

Central Regional Educational Laboratory (REL Central) staff provided each intervention school with all available CASL professional development materials as of fall 2007. The materials included:

- One copy of *Classroom Assessment for Student Learning* (Stiggins et al. 2004) for each member of the learning team (including all of the grade 4 and 5 teachers in the school). This book was the central text of the intervention. Teachers were asked to read the book and then discuss it during their learning team experience.
- One copy of the *Learning Team Facilitator Handbook* (Chappius 2007) per school. Teachers used this guide to plan and conduct their implementation of CASL. The handbook included instructions on readings, activities, meetings, and discussions.
- One copy per school of four supplemental books that extend the learning in selected chapters of CASL:
    - *Creating and Recognizing Quality Rubrics* (Arter & Chappius 2006).
    - *A Repair Kit for Grading: 15 Fixes for Broken Grades* (O'Connor 2007).
    - *Understanding School Assessment: A Parent and Community Guide to Helping Students Learn* (Chappius & Chappius 2002).
    - *Assessment for Learning: An Action Guide for School Leaders* (Chappius, Stiggins, Arter, & Chappius 2006).

- One set per school of seven interactive training presentations on DVD:
    - Assessment for Student Motivation.
    - Evaluating Assessment Quality: Hands-On Practice.
    - Assessing Reasoning in the Classroom.
    - Commonsense Paper and Pencil Assessments.
    - Designing Performance Assessments for Learning.
    - Grading & Reporting in Standards-Based Schools.
    - Student-Involved Conferences.
- One copy per school of the DVD, New Mission, New Beliefs: Assessment for Learning, a keynote presentation by Rick Stiggins that offers insight into the importance of sound classroom assessment.

Research staff provided the CASL materials to the intervention schools as if the schools had purchased CASL themselves and were not part of a research study. The research staff was not involved in the intervention schools' implementation of the intervention in any way, other than to collect implementation data.

Learning teams

According to the CASL developers, approximately 75 percent of CASL users participate in the program within learning teams; accordingly, the learning team model was determined to be the typical CASL implementation approach. CASL professional development time is devoted to individual reading and study between meetings, learning team meetings, and trying out assessment principles and practices in the classroom. The CASL authors recommended that teachers form learning teams of three to six members to get the most benefit from the program (Stiggins et al. 2004), and that these teams meet regularly for a period of time (such as a school year) to increase their abilities in classroom assessment (Chappuis 2007). According to the authors, CASL learning teams are intended to help all team members become assessment literate. The learning team process engages teachers intellectually in learning, planning, experimenting with strategies and techniques, and discussion of assessment (Arter 2001). The CASL learning team process includes the following cognitive and practical steps (Stiggins et al. 2004, p. 22):

- Thinking about classroom assessment.
- Reading and reflecting on new classroom assessment strategies.
- Shaping the strategies into applications.
- Trying out applications, observing, and drawing inferences about what does and does not work.
- Reflecting on and summarizing learning and conclusions from that experience.
- Sharing and problem solving with team members.

The learning team approach, as opposed to a workshop learning approach, is intended to be flexible, customizable, and low in cost. The learning team approach is also intended to be job-embedded in the classroom context, so that teachers are actively reflecting together about their use of assessment in their current teaching. Rather than a one-time event, the learning team meetings are ongoing and help develop expertise internally rather than bringing it in from the outside (Chappuis 2007).

All grade 4 and 5 mathematics teachers in each intervention school were recruited to the study and asked to form a learning team, following the recommendations in the *Learning Team Facilitator Handbook* (Chappuis 2007) about how to set up and conduct their teams. The handbook recommended that the teachers allow one to two hours for learning team meetings, schedule the meetings two to three weeks apart, and generally address one chapter from the CASL text during each meeting. Each meeting was to be led by a teacher facilitator; this role could be permanent or rotate among teachers. The recommended meeting agenda included discussion of the reading, discussion of teacher actions in the classroom since the last meeting, and preparation for the next assignment. The CASL authors recommended that between the meetings teachers spend two to four hours reading the CASL textbook or ancillary books, viewing the video, and practicing the CASL techniques in the classroom.

Introductory videoconference

When districts initially implement CASL, they typically participate in a presentation by a CASL author. For this study, Rick Stiggins conducted an orientation through a videoconference open to all intervention schools and available at eight sites around the state; these sites were chosen for their proximity to each intervention school (no more than 45 minutes by car) and for their videoconferencing capabilities and capacity to accommodate the number of teachers expected. Videoconferencing sites included education institutions, county fairgrounds, and hospitals. Stiggins began his presentation by discussing the key ideas and strategies of high-quality classroom assessment and then took questions from the participants. The entire session was recorded on DVDs, which were mailed to each intervention school so that teachers who could not attend the videoconference could view it.

District facilitators

Districts adopting CASL typically assign a district-level employee trained to provide assistance to school learning teams. For this study, an employee from each district with at least four participating schools was chosen and trained to serve as a district facilitator. District facilitators were typically instructional coaches, elementary education directors, assessment coordinators, or professional development directors. An at-large facilitator was available for all districts with fewer than four schools. The district facilitators' role was to help learning teams in intervention schools with issues related to implementing CASL (as opposed to issues related to the research, which were handled by the research team). To ensure that the facilitators were knowledgeable enough to provide this assistance, all facilitators attended the Classroom Assessment for Student Learning Workshop in November 2007. This workshop was intended to deepen participants' understanding of classroom assessment and to prepare them to lead others in classroom

assessment practice. Facilitators were to be available by telephone or e-mail to answer questions, and could also attend learning team meetings upon request.

Indicators of fidelity of implementation of Classroom Assessment for Student Learning training

Criteria and measures for defining and describing adequate CASL implementation were not based on empirical study of typical implementation of CASL but rather were defined by the developers (table 3.1). Data from the intervention logs described in chapter 2 were used to evaluate the extent to which the program was implemented with fidelity by intervention group teachers relative to these criteria.

Table 3.1. Measures of adequate fidelity of CASL implementation during the CASL training year

Indicator of adequate fidelity	Data source	Adequacy criteria
1. CASL learning teams are formed to support collaborative learning	CASL participant log learning team startup: items 5, 6, 7, 8, and 9	At least four of items 5, 6, 7, 8, and 9 answered affirmatively
2. Read the 13 chapters of *Classroom Assessment for Student Learning: Doing It Right—Using It Well*	CASL participant log chapters 1–13: item 2	All 13 chapters of the text were read
3. Nine or more CASL learning team meetings are attended	CASL participant log chapters 1–13: items 4 and 5	At least nine CASL learning team meetings were attended
4. Individual members of a CASL learning team devote 60 hours to CASL activities, including attending learning team meetings, reading the text, and trying out new ideas in the classroom	CASL participant log chapters 1–13 logs: item 9	60 hours were completed
5. Reflection on learning about classroom assessment with colleagues occurs in small groups	CASL participant log learning team startup: item 2; CASL participant log chapters 1, 8, and 13 Logs: items 8a and 8b	CASL learning teams included groups of three to six members; items 8a and 8b answered either "to some extent" or "to a great extent"

Source: Indicators of adequate fidelity are from Stiggins et al. (2004); personal communication, Judy Arter, (September 27, 2006); Chappuis (2007, p. 18); personal communication, Rick Stiggins (July 20, 2006); Arter & Busick (2001, p. 413–15).

Fidelity of implementation data

The sample of teachers and schools used for the analyses of implementation fidelity was made up of teachers who were in schools assigned to the intervention condition, were not in schools that were case-deleted because they declined to participate after random assignment, were not teachers who joined the study too late to complete the participant logs, and were eligible for participation in the study. People who may have attended learning team meetings but who were not eligible for the study included district staff, principals, teachers who did not teach grades 4 or 5, and resource teachers who had no direct responsibility for student instruction. Some of these individuals completed chapter logs to obtain the optional university course credit that was

available to all learning team participants, regardless of study eligibility, but their data were not included in the analysis described below.

Implementation data were available from 31 intervention schools. One-hundred-fifty-eight teachers out of the 175 randomly assigned to the intervention group provided at least some implementation fidelity data (6 implementation teachers withdrew at random assignment, and 11 failed to provide any implementation data). No late-entry teachers joined the study during the CASL training year; therefore, late-entry teachers did not complete the CASL participant logs.

Classroom Assessment for Student Learning training year fidelity results

This section discusses the results for the five indicators used to measure CASL training year fidelity.

Indicator 1: Forming CASL learning teams

To determine how many teachers were assigned to CASL learning teams, affirmative answers to items 5, 6, 7, 8, and 9 on the startup log were counted. These items, each of which was answered yes/no, were as follows:

- Item 5. Did your learning team establish group operating principles?
- Item 6. Did your team set a meeting schedule?
- Item 7. Has your team chosen a regular place to meet?
- Item 8. Did your team establish a reading and assignment schedule?
- Item 9. Do you and your learning team members have a common purpose or goal for your learning team?

Eighty-six of the 128 teachers (67.19 percent) responding to these items on the startup log met the implementation fidelity criterion of responding "yes" to at least four items.

All participating teachers in each intervention school were to be assigned to the same learning team. Thus, this criterion could also be applied at the school level. To determine whether the criterion was met at the school level, the responses from teachers (total counts of affirmative answers) in each school were averaged. Schools with a teacher average of at least four (of five possible) were considered to have met the criterion. Twenty-one of the 30 schools (67.74 percent) that provided startup log data met this criterion.

Indicator 2: Reading the CASL textbook

Item 2 on CASL participant logs for chapters 1–13 was used to determine to what extent the teachers read all chapters of the textbook. The item read as follows.

- Item 2. Did you read the CASL text Chapter X?

  - ☐ Yes, completely
  - ☐ Yes, partially
  - ☐ No, not at all

For each teacher the number of chapters completely read and the number partially read were counted, and the counts were summed to arrive at a number of chapters at least partially read. One hundred-forty-two of 158 teachers completed at least one of the relevant chapter log items. Seventy-one teachers completed every chapter log. The study team also analyzed data received from those teachers with incomplete data who responded to at least one relevant log item, counting missing chapter log data as chapters not read. Of the 142 teachers who provided at least some data on their reading, 59 teachers (41.55 percent) reported at least partially reading each chapter, while 23 of those 59 (39.98 percent) reported fully reading each chapter. On average, teachers who provided reading data reported that they fully or partially read 8.23 chapters (standard deviation = 5.24).

Indicator 3: Attending nine learning team meetings

According to the developers, CASL participants should attend at least nine learning team meetings. The responses to Item 5 of the CASL Participant Logs were totaled to determine how many meetings study participants attended. This item read as follows:

- Item 5. Of these meetings [the number of meetings held on each chapter], how many did you attend?

Of the 142 teachers responding to this item on at least one log, 89 teachers (62.68 percent) attended at least nine CASL learning team meetings; 16 teachers did not respond to any of the relevant items on the chapter logs. On average, teachers who responded to the item attended 9.78 meetings during the CASL training year (standard deviation = 8.04).

Indicator 4: Completing 60 hours of training

Item 9 of the CASL Participant Logs was used to determine whether teachers spent the developer-recommended 60 hours on all aspects of CASL, including reading the textbook and ancillary books, viewing the videos, attending learning team meetings, and trying out assessment techniques in their classrooms. This item read as follows:

- Item 9. How many total hours did you spend on chapter X, including reading, completing activities, trying out applications in the classroom, reflecting, and participating in learning team meetings? Round your answer to the nearest hour.

These answers were summed for each person across the 13 CASL participant logs. Of the 142 teachers responding to this item on at least one log, 12 teachers (8.45 percent) spent at least 60 hours on CASL; 16 did not respond to any of the relevant items on the chapter logs. On average, teachers reported spending 31.21 hours (standard deviation = 19.89) on CASL.

Indicator 5: Small group discussions

The developers of CASL recommended that learning teams include three to six members and that these teams reflect together on their learning about classroom assessment. To determine the number of participants in each learning team, responses to item 2 on the startup log were examined. This open-ended item read as follows:

- Item 2. How many people are on your learning team?

Because not all learning team members within a school gave the same answer, the mean of the answers within each school was used and rounded to the nearest whole number. On average, the learning teams were made up of five members (standard deviation = 1.25); the smallest had three members and the largest had eight members. Of the 31 schools providing implementation fidelity data, all but three (90.03 percent) had learning team sizes that were in the developers' acceptable range.

The assessment content of learning team meetings was evaluated using responses to two items on the CASL participant logs for chapters 1, 8, and 13:

- Item 8a. During the learning team meeting(s) for this chapter, to what extent did you share with the team ways that you have implemented the CASL techniques in your classroom?
- Item 8b. During the learning team meeting(s) for this chapter, to what extent did you share with the team results you have seen from using the CASL techniques?

Teachers responded to each item using the following scale:

☐ I did not attend any meetings

☐ Hardly at all

☐ A little

☐ To some extent

☐ To a great extent

Adequate implementation was defined as responding "to some extent" or "to a great extent" on both items. Of the 121 teachers responding to at least one relevant item on the participant logs (38 did not respond to any of the relevant questions), 94 (77.69 percent) responded "to some extent" or "to a great extent" on at least one log. On average, teachers responded "to some extent" or "to a great extent" twice out of six possible opportunities (two opportunities in each of three participant logs).

Summary of CASL training year implementation fidelity indicators

There were five indicators of implementation quality that pertained specifically to teachers participating in CASL:

1. Whether their learning team met quality standards as established by the authors
2. Whether they read the 13 chapters of the textbook
3. Whether they attended nine or more CASL meetings
4. Whether they spent 60 hours on CASL
5. Whether they reported spending learning team time discussing assessment

Therefore, the 158 participating teachers who provided some training year implementation fidelity data could have met anywhere from zero to five of these indicators of implementation quality. The number and percentage of the 158 teachers who reported each possible number of quality indicators appears in table 3.2.

Table 3.2. Number and percentage of intervention teachers achieving indicators of intervention quality during the CASL training year

Number of indicators of quality	Number of intervention teachers	Percentage of intervention teachers
0	38	24.05
1	20	12.66
2	28	17.72
3	28	17.72
4	40	25.32
5	4	2.53

There were two indicators of implementation quality that pertained specifically to schools participating in CASL: whether their learning team met quality standards, and whether the learning team had the appropriate number of members. Of the 32 intervention schools, 18 (56.25%) achieved both indicators, and 13 (40.63%) achieved one indicator.

Classroom Assessment for Student Learning implementation year fidelity results

During the CASL implementation year (the 2008/09 school year), intervention group teachers were asked to fully implement the CASL techniques in their classrooms. The CASL developers

also recommended optional learning team activities to be conducted during the implementation year (J. Chappuis, personal communication, May 28, 2008), which were provided to each implementation school. These options were as follows:

- The team focuses on assessment quality. Team members throughout the year audit each assessment they give for standards of quality.
- The team focuses on assessment for learning, planning how they will use the seven strategies of assessment for learning in each unit or segment of instruction and then meeting together to discuss what they did.
- The team works through the video series, selecting the videos that most apply to their needs, or rereads chapters of the CASL book or the ancillary books.
- Each team member selects one assessment method and revises all relevant assessments, or the team as a whole selects an assessment method.

Participant logs were not administered in the study during the CASL implementation year, as the logs had focused on the CASL training. Instead, teachers in the intervention schools who were present at the Wave 5 (posttest) preparation site visits described to the researchers how they had implemented CASL during the CASL implementation year. Not all intervention teachers were present at these visits, but all intervention schools were visited. The CASL implementation during the implementation year, therefore, is characterized only by the descriptions of the teachers present at the visits. The teachers present at 24 of the intervention schools reported using CASL techniques in their classrooms during the CASL implementation year. In eight intervention schools, all teachers present at the site visits reported no implementation of CASL during the implementation year.

During the site visits teachers were also asked what practices they had implemented in their classrooms as a result of their training with CASL. When analyzed, the answers were grouped into five general categories:[1]

- *Developing assessments differently (15 schools).* Examples included creating reasoning questions, developing assessments before instruction, using multiple forms of assessment, making assessments more authentic, altering premade assessments to align with instruction, allowing students to take assessments as many times as needed, giving immediate feedback from homework, and developing rubrics.
- *Changing instructional practices (14 schools).* Examples included providing more visual help, reteaching in groups, changing grouping practices, understanding student learning curves, adjusting lessons based on analysis of assessment results, focusing on objectives and communicating them, using CASL in other content areas such as writing and social studies, aligning learning targets and textbook, using a curriculum map to develop goals, and showing students anonymous examples.

[1] The types of practices intervention schools engaged in during the implementation year were not summarized because these were categorical variables and did not represent intensity of implementation.

- *Eliciting student involvement in assessment (13 schools).* Examples included teachers having their students set their own goals according to assessment and survey results and learning targets, be more accountable for their own learning, help make rubrics, participate in student-involved or -led conferences, do self-assessment at the beginning and end of each unit, describe what they are learning, show their learning in different ways, and provide feedback on what they are doing.
- *Changing grading practices (9 schools).* Examples included monitoring student progress every two weeks, ending the practice of taking off points for late work, allowing students to redo assignments for better grades, allowing students to grade tests as a class, not recording grades for what has not yet been mastered, giving students descriptive feedback with grades rather than just a number, providing feedback without a grade, describing grades according to objectives and skills, organizing report cards according to assessment purpose, and changing the weighting of grades.
- *Teacher collaboration (3 schools).* Examples included collaborating together about assessments, striving for consistency in assessments across teachers, and discussing examples of how they used the ideas in the CASL book.

Study participants at 16 of the 27 responding schools did not meet as a CASL learning team during the CASL implementation year, although participants at 18 schools did meet at least occasionally about improving assessment for learning. For those who did meet as a learning team during the CASL implementation year, activities included reviewing the textbook or specific chapters in the textbook again (6 schools), discussing how the techniques worked in the classroom (3 schools), and disseminating the information through blogs or discussions (3 schools).

When asked about barriers to implementing CASL, teachers at 26 of the schools said that they needed more time for the program, and teachers at 24 schools said that competing district initiatives took time away from the program. At 15 of the schools, teachers said that they would have liked more help from an external facilitator or coach. Ten schools had undergone teacher or principal personnel changes during the study period that made implementation challenging. In reacting to the CASL content, teachers at nine schools said that mathematics was not best suited to CASL because there were not enough mathematics examples in the text and because they perceived mathematics assessment to be less flexible and less amenable to the CASL strategies.

Professional development surveys

To understand the context of non-CASL teacher professional development, all intervention and control teachers were administered the Survey of Professional Development four times during the study. The survey asked participants to describe the type of professional development in which they participated (such as workshop), the subjects covered by the professional development (such as writing), the emphasis the professional development had on various aspects of instruction and assessment, and their perception of the usefulness of the professional development, for up to two professional development activities within the past semester.

Participation in professional development

The majority of teachers in both the intervention and control groups reported participation in non-CASL professional development activities during the course of the study (table 3.3). In Wave 3, the participation rate among the total sample of the intervention group teachers (80.15 percent) was 8.74 percent lower than the participation rate among the total control group sample (88.89 percent), a statistically significant difference. Otherwise, participation rates in non-CASL professional development between intervention and control teachers did not differ by statistically significant margins.

Table 3.3. Intervention and control teacher participation in non–Classroom Assessment for Student Learning professional development

	Intervention			Control				
Wave	Responded to survey	Participated in professional development	Participation rate	Responded to survey	Participated in professional development	Participation rate	Difference	*p-value*
2	117	102	87.18	154	132	85.71	1.47	.73
3	131	105	80.15	162	144	88.89	–8.74	.04
4	115	105	91.30	170	153	90.00	1.30	.72
5	115	92	80.00	171	140	81.87	–1.87	.69

Source: Survey of Professional Development.

Professional development formats

Teachers were asked to identify the formats of all non-CASL professional development activities in which they participated during the specified data collection period (table 3.4).

Table 3.4. Teachers reporting participation in specific professional development formats by data collection wave (percent)

Activity	Wave 2 December 2007			Wave 3 May 2008			Wave 4 December 2008			Wave 5 (Posttest) May 2009		
	Intervention group (118)	Control group (160)	*p*-value	Intervention group (133)	Control group (166)	*p*-value	Intervention group (116)	Control group (170)	*p*-value	Intervention group (116)	Control group (171)	*p*-value
Workshop or institute	58.13	68.64	.07	46.99	48.12	.85	66.47	68.10	.77	44.44	38.79	.34
Conference	22.88	18.13	.33	15.79	18.67	.51	27.59	28.24	.90	15.52	19.88	.35
College course	18.64	21.25	.59	19.55	19.88	.94	18.10	24.71	.19	14.66	14.62	.99
Teacher network	22.88	8.13	<.01	17.29	12.65	.26	12.93	9.41	.35	18.97	15.20	.40
Internship, immersion activity, or work with mentor	25.00	24.58	.94	22.89	19.55	.48	23.53	21.55	.66	18.71	17.24	.75
Teacher committee or task force	36.44	30.00	.26	36.09	28.92	.19	35.34	38.82	.55	37.93	28.65	.10
Teacher study group	34.75	17.50	<.01	35.34	25.30	.06	37.07	32.35	.41	34.48	38.60	.48
Other professional development activity	22.03	21.25	.88	19.55	21.69	.65	17.24	9.41	.05	10.34	9.94	.91

Note: Numbers in parentheses are number of teachers. Responses for workshop and institute and responses for internship, immersion activity, and work with mentor were collapsed to prevent disclosure due to small cell sizes.
Source: Survey of Professional Development.

There were two significant differences in professional development formats experienced by the participating teachers. In Wave 2, just over 8 percent of control teachers but 22.88 percent of intervention teachers noted that they participated in a "teacher network" activity, whereas 17.50 percent of control teachers and 34.75 percent of intervention teachers participated in a "teacher study group" activity (see table 3.4). Forty tests of statistical significance were conducted; two yielded statistically significant differences at the .05 level. This is exactly the number of tests that would be statistically significant by chance alone given the threshold for statistical significance and the number of tests. In other words, it is unlikely that there were any real differences between the percentages of teachers in the intervention and control groups who participated in the various non-CASL professional development activities.

Teachers identified each of their non-CASL professional development activities by content area (table 3.5). For teachers in both the intervention and control groups and across all waves, the smallest percentage of teachers indicated participation in health-related professional development. For teachers in both the intervention and control groups in data collection Waves 4 and 5, the highest percentage of teachers indicated participation in reading-related professional development. The only significant group difference was in Wave 2, where more teachers in the intervention group than the control group reported participating in mathematics-related professional development. Given the threshold for statistical significance and the number of tests conducted, it is possible that the difference between intervention and control teachers on Wave 2

mathematics-related professional development occurred by chance. Thus, the non-CASL professional development experiences of the teachers in the intervention and control groups likely did not differ by statistically significant margins in terms of subject area.

Table 3.5. Teachers reporting professional development activities in subject areas (percent)

	Wave 2			Wave 3			Wave 4			Wave 5 (Posttest)		
Activity	Intervention group (118)	Control group (160)	*p*-value	Intervention group (133)	Control group (166)	*p*-value	Intervention group (116)	Control group (170)	*p*-value	Intervention group (116)	Control group (171)	*p*-value
Math	54.24	41.88	.04	34.59	31.33	.55	36.52	40.00	.52	40.87	32.75	.14
Science	17.80	15.63	.63	21.80	16.87	.28	28.70	25.29	.55	22.61	23.98	.89
Reading	39.83	38.13	.77	39.10	44.58	.34	50.43	46.47	.56	41.74	46.20	.51
Writing	43.22	42.50	.90	36.09	42.77	.24	46.96	42.35	.48	34.78	37.43	.72
Social studies	14.41	11.88	.53	18.05	13.86	.32	20.00	18.82	.83	16.52	21.05	.42
Health-related topics	4.24	2.50	.42	3.76	2.41	.50	3.48	1.76	.37	3.48	2.34	.58
Activity was not specific to one subject area	26.25	30.51	.44	27.82	29.52	.75	28.70	34.12	.31	25.22	30.41	.32
Activity covered another subject	93.13	88.98	.22	100.00	98.19	.12	13.91	15.29	.88	11.30	14.04	.48

Note: Numbers in parentheses are number of teachers.
Source: Survey of Professional Development.

Professional development emphasis

On the Survey of Professional Development, teachers were asked to indicate the degree of emphasis the non-CASL professional development activities placed on aspects of teaching and learning associated with classroom assessment. Several items in this section of the survey addressed topics related to classroom assessment.

Table 3.6 shows the percentage of teachers in the intervention and control groups who reported no emphasis, minor emphasis, or major emphasis across a number of topics. The intervention and control group teachers differed by a statistically significant margin for 5 of the 44 comparisons. In all but one of these four cases, control group teachers reported greater emphasis than did intervention group teachers: on increasing student involvement in both Wave 2 and Wave 5 (posttest), on use of technology in Wave 3, and on strategies for teaching diverse student populations in Wave 3. Intervention teachers reported greater emphasis on curriculum than did control teachers in Wave 5 (posttest).

The results in tables 3.4 and 3.5 suggest that intervention teachers may have reported CASL activities in the Survey of Professional Development. The results in table 3.6, however, clearly show that intervention teachers did not report CASL activities in the Survey of Professional Development.

Table 3.6. Teacher reports of emphasis of professional development activities

Item/ wave	Intervention				Control				*p*-value[a]
	Sample size	No emphasis (percent)	Minor emphasis (percent)	Major emphasis (percent)	Sample size	No emphasis (percent)	Minor emphasis (percent)	Major emphasis (percent)	
Curriculum (such as units, textbooks, standards)									
Wave 2	102	6.86	23.53	69.61	132	5.30	30.30	64.39	.49
Wave 3	106	8.49	24.53	66.98	145	7.59	33.79	58.62	.29
Wave 4	104	6.73	21.15	72.12	152	6.58	28.29	65.13	.43
Wave 5	92	6.52	20.65	72.83	140	8.57	35.00	56.43	.04
Working with content standards (such as understanding, unpacking, simplifying, aligning instruction to standards)									
Wave 2	102	6.86	21.57	71.57	131	6.11	27.48	66.41	.58
Wave 3	106	6.60	30.19	63.21	144	9.03	28.47	62.50	.77
Wave 4	104	8.65	31.73	59.62	151	7.28	31.13	61.59	.91
Wave 5	92	10.87	26.09	63.04	140	7.86	38.57	53.57	.14
Instructional methods									
Wave 2[a]	102	5.88	13.73	80.39	131	13.74		86.26	.17
Wave 3	106	3.77	17.92	78.30	145	2.07	13.10	84.83	.39
Wave 4[a]	104	17.31		82.69	152	2.63	17.76	79.61	.60
Wave 5[a]	92	21.74		78.26	139	3.60	10.07	86.33	.11
Increasing student involvement in learning									
Wave 2[a]	102	4.90	18.63	76.47	130	7.69		92.31	<.01
Wave 3[a]	105	6.67	10.48	82.86	144	11.81		88.19	.09
Wave 4	104	0.00	20.19	79.81	151	1.99	10.60	87.42	.04
Wave 5[a]	92	20.65		79.35	139	10.07		89.93	.08
Formative assessments (e.g. developing, selecting, and using assessment in the classroom)									
Wave 2	100	16.00	40.00	44.00	131	13.74	39.69	46.56	.87
Wave 3	106	17.92	37.74	44.34	145	11.72	44.83	43.43	.30
Wave 4	104	12.50	37.50	50.00	150	9.33	39.33	51.33	.72
Wave 5	92	13.04	35.87	51.09	139	17.99	43.17	38.85	.18
Communicating assessment results to students									
Wave 2	101	21.78	46.53	31.68	131	14.50	47.33	38.17	.30
Wave 3	106	23.58	34.91	41.51	145	17.24	44.14	38.62	.26
Wave 4	103	14.56	46.60	38.83	151	9.27	49.67	41.06	.43
Wave 5	91	18.68	43.96	37.36	140	20.71	45.00	34.29	.87
Using assessment results to guide instruction (such as making adjustments in instructional strategies or lesson plans)									
Wave 2	102	17.65	34.31	48.04	132	11.36	34.09	54.55	.35
Wave 3	105	17.14	29.52	53.33	145	14.48	32.41	53.10	.80
Wave 4	103	8.74	24.27	66.99	151	7.28	29.14	63.58	.67
Wave 5	92	17.39	25.00	57.61	139	14.39	33.09	52.52	.41
Use of technology in instruction									
Wave 2	102	33.33	48.04	18.63	132	30.30	47.73	21.97	.79
Wave 3	105	35.24	31.43	33.53	145	28.28	50.34	21.38	<.01
Wave 4	103	24.27	48.54	27.18	151	19.87	54.30	25.83	.61
Wave 5	89	25.84	43.82	30.34	139	29.50	48.92	21.58	.33
Strategies for teaching diverse student populations									
Wave 2	102	17.65	27.45	54.90	131	13.74	35.11	51.15	.41
Wave 3	106	16.04	27.36	56.60	146	5.48	33.56	60.96	.02
Wave 4	104	7.69	35.58	56.73	152	7.24	32.24	60.53	.83
Wave 5	91	7.69	30.77	61.54	139	5.76	30.22	64.03	.83
Leadership development									
Wave 2	102	28.43	47.06	24.51	132	34.09	39.39	26.52	.48
Wave 3	105	41.90	31.43	26.67	146	27.40	41.78	30.82	.05
Wave 4	104	32.69	43.27	24.04	150	30.00	42.00	28.00	.77

Wave 5	89	28.09	42.70	29.21	140	32.14	41.43	26.43	.79
Statewide assessment or standardized testing									
Wave 2	102	24.51	42.16	33.33	131	17.56	41.98	40.46	.35
Wave 3	106	29.25	40.57	30.19	144	25.00	44.44	30.56	.73
Wave 4	104	19.23	37.50	43.27	152	19.74	50.00	30.26	.08
Wave 5	91	23.08	40.66	36.26	140	27.14	43.57	29.29	.52

a. Data for no emphasis and minor emphasis collapsed to prevent disclosure due to small cell size; p-values based on un-collapsed data.
Note: Cells may not sum to 100 percent due to rounding. *p*-values based on chi-square tests comparing the frequencies among intervention and control groups. Cells were merged, where necessary, to prevent disclosure.
Source: Survey of Professional Development.

Table 3.7 shows the percentage of teachers reporting that at least one professional development activity was aligned with state content standards. For all waves, there were no statistically significant differences between groups in the percentage of teachers reporting the professional development aligned with state content standards.

Table 3.7. Teachers reporting professional development aligned with state content standards (percent)

Wave	Intervention	Control	*p*-value
2	98.04	95.45	.45
	(51)	(66)	
3	100.00	92.86	.06
	(46)	(56)	
4	78.43	83.54	.46
	(51)	(79)	
5	78.05	77.78	.92
	(42)	(63)	

Note: Numbers in parentheses are number of teachers.
Source: Survey of Professional Development.

Teachers were asked to report on the quality of the non-CASL professional development by rating each activity as "poor" (coded as 1), "fair" (coded as 2), "good" (coded as 3), or "excellent" (coded as 4) (table 3.8). There were no statistically significant differences between the intervention and control groups in their ratings of professional development quality.

Table 3.8. Teachers reporting professional development quality (percent)

	Intervention			Control			
Wave	Mean	Standard deviation	Sample size	Mean	Standard deviation	Sample size	*p*-value
2	3.74	0.44	72	3.70	0.46	93	.60
3	3.76	0.46	71	3.83	0.45	103	.34
4	3.51	0.67	72	3.46	0.76	112	.59
5	3.41	0.79	64	3.47	0.73	93	.59

Note: Responses were "poor" (coded as 1), "fair" (coded as 2), "good" (coded as 3), "excellent" (coded as 4).
Source: Survey of Professional Development.

Teachers were asked to report on the perceived impact the non-CASL professional development would have on their classroom instruction by rating the anticipated impact as "none" (coded as

1), “low” (coded as 2), “medium” (coded as 3), or “high” (coded as 4). Table 3.9 displays the mean teacher response by group and by wave. There were no significant differences.

Table 3.9. Teachers reporting anticipated professional development impact (percent)

	Intervention			Control			
Wave	Mean	Standard deviation	Sample size	Mean	Standard deviation	Sample size	*p*-value
2	3.80	0.41	73	3.74	0.47	95	.40
3	3.86	0.38	81	3.87	0.39	105	.97
4	3.58	0.71	86	3.55	0.71	120	.75
5	3.62	0.69	76	3.58	0.63	97	.68

Note: Responses were “poor” (coded as 1), “fair” (coded as 2), “good” (coded as 3), “excellent” (coded as 4).
Source: Survey of Professional Development.

Summary

Logs and surveys were administered to participating teachers over the course of the study to collect information on CASL implementation fidelity and the greater professional development context. Not all participating teachers responded to all logs and surveys. Response rates for the CASL participant logs ranged from 69 percent to 79 percent. Response rates to the Survey of Professional Development ranged from 69 percent to 80 percent.

In this study, the majority of responding treatment teachers reported learning team experiences that met the CASL developers’ criteria. More than half the responding teachers (67.18 percent) were in learning teams that met the developers’ criteria, such as developing operating principles and a common goal. All but three schools had learning teams of the appropriate size. More than half the teachers (62.68 percent) attended the minimum number of nine learning team meetings, and more than 75 percent of responding teachers reported that they discussed relevant assessment content in learning team meetings.

Many teachers did not report meeting criteria related to reading the textbook and spending time on the program. Less than half (41.55 percent) of teachers reported at least partially reading each textbook chapter. To report reading the entire textbook, however, teachers had to complete all 13 chapter logs, which only 71 teachers did. The extent to which teachers read chapters for which they did not complete chapter logs cannot be known. Only 12 teachers reported spending the recommended amount of time (60 hours) on CASL; the average time spent was just over half that, at 31.21 hours. However, because the 60-hour criterion was based on developer recommendations rather than empirical studies of program implementation, it is not possible to conclude that the teachers in this study spent less (or more) time on CASL than is customary.

In terms of non-CASL professional development, there were nine statistically significant differences between the intervention group and control group. In Wave 3, more control teachers participated in professional development than intervention teachers did, although control teachers reported less involvement in teacher network and less involvement in teacher study group activities than did intervention teachers in Wave 2. Also, in Wave 2 more teachers in the intervention group than the control group participated in mathematics-related professional

development. Control teachers reported more emphasis of professional development related to student involvement than did intervention teachers in both Wave 2 and Wave 4. In Wave 3, the control group teachers reported more emphasis of professional development on use of technology in instruction and on strategies for teaching diverse student populations than did the intervention teachers. In Wave 5 (posttest), intervention teachers reported more emphasis on curriculum than did control teachers.

It should be noted, however, that 124 tests of statistical significance were conducted and presented in this chapter comparing the professional development experiences of the teachers in the intervention and control groups. Nine tests resulted in statistically significant differences. Given that the tests for statistical significance were conducted at the .05 level, approximately six tests out of the 124 would yield a statistically significant difference by chance alone. Few (9 of 124) statistically significant differences were found between the non-CASL professional development experiences of the intervention and control teachers suggesting that for professional development, the only difference between the intervention group and the control group (the counter factual) was the CASL intervention.

Chapter 4. Impacts of the intervention

This chapter presents the results of the analyses used to estimate the impact of CASL on the primary outcome of student achievement. This chapter includes a discussion of the impact analyses, a presentation of impact estimates generated by the multilevel models, and the results of the sensitivity analyses.

Impact analyses

The primary student outcome was student scale scores on the mathematics portion of the Colorado Student Assessment Program (CSAP), the statewide achievement test used to measure adequate yearly progress under the No Child Left Behind Act. Student-level CSAP data were obtained from the Colorado Department of Education. To prevent the disclosure of sensitive information, no information identifying individual students was collected, but unique identifiers were provided to allow the study team to link student test data across the different administrations of the CSAP. The student achievement data did not include any information that linked students to teachers, and student identities were masked. Individual students, therefore, could not be linked to teachers. The data did, however, include school and district identifiers. Given these properties, the student achievement impact model nested students within school and schools within districts (block). Statistical analysis confirmed that achievement scores of students within the same school were related and thus warranted the use of multilevel modeling to estimate impacts (see appendix I).

Analysis failed to yield any statistically significant impacts of CASL on student mathematics achievement. Table 4.1 shows the means for the intervention and control groups on the pretest, the posttest unadjusted for the pretest scores, and posttest adjusted for pretest scores. (Appendix J provides raw means and standard deviations for all outcomes, and appendix K provides complete results from the impact analyses models.) The comparison of posttest scores unadjusted for pretest scores was included as a sensitivity analysis, and the comparison of posttest scores adjusting for pretest scores was the benchmark model. All analyses accounted for the clustering of students within schools, and estimates were combined across the 40 imputed student achievement datasets. (See the treatment of missing data section in Chapter 2 for a description of the creation of the 40 imputed datasets.) The student achievement impact analysis sample included students in Cohorts 1 and 3. Cohort 1 students were in grade 4 during the CASL training year and grade 5 during the CASL implementation year, and their pretest scores came from the spring 2007 CSAP administration, the spring before the CASL training year. Cohort 3 students were in grade 3 during the CASL training year and grade 4 during the CASL implementation year; their pretest scores came from the spring 2008 CSAP administration, at the end of the CASL training year. All students' posttest scores came from the spring 2009 CSAP administration at the end of the CASL implementation year. As a reminder, only grade 4 and grade 5 mathematics teachers in the intervention schools studied and implemented CASL.

Table 4.1. Intervention and control group means and standard errors for student mathematics achievement scores on the Colorado Assessment of Student Progress

Measure	Intervention group mean Schools = 33 Students = 4,420	Control group mean Schools = 34 Students = 5,176	Estimated difference[a]	95 percent confidence interval	*p*-value	Effect size[b]
Pretest score	**451.79 (5.41)**	**458.95 (5.26)**	**–7.16 (7.36)**	**–21.59–7.28**	**.33**	**–0.08**
Unadjusted posttest score	499.03 (4.89)	503.96 (4.75)	-4.96 (6.66)	–17.98–8.12	.46	-0.06
Adjusted posttest score	502.49 (2.53)	501.91 (2.44)	0.58 (3.47)	–6.23–7.38	.87	0.01

a. Estimated difference may not equal difference between means because of rounding.
b. Calculated as the estimated difference divided by control group standard deviation.
Note: Numbers in parentheses are standard errors. All results are based on analysis using a mixed-model approach to account for the sources of variability in the data that resulted from the nested structure of the school environment.
Source: 2007, 2008, and 2009 Colorado Student Assessment Program data.

Intervention and control group schools did not differ by a statistically significant margin on their pretest mathematics test scores ($\gamma_{01} = -7.16$, standard error = 7.36, $p = .33$). Neither the no-covariate model nor the covariate model yielded a statistically significant impact for CASL on student achievement. The benchmark covariate model estimated the school-level intervention impact as 0.58 scale score point (standard error = 3.47, $p = .87$) on adjusted mathematics achievement test scores.

Follow-up exploratory analyses were conducted to estimate the impact of CASL separately for students in Cohort 1 and students in Cohort 3 to better understand the impact CASL had on the mathematics achievement of students who experienced the intervention for two years or one year, respectively. For these analyses, the benchmark impact model was fit twice, once to the data for the Cohort 1 students (students who were in grade 5 during the CASL implementation year and experienced two years of exposure to CASL) and once to the data for the Cohort 3 students (students who were in grade 4 during the CASL implementation year and experienced one year of exposures to CASL). These follow-up analyses also used the 40 imputed datasets described in the section on treatment of missing data in chapter 2.

Neither subgroup analysis yielded a statistically significant impact of CASL on adjusted mean mathematics test scale scores: Cohort 1, $\gamma_{01} = -4.53$ (standard error = 3.80, $p = .24$) and Cohort 3, $\gamma_{01} = 5.59$ (standard error = 4.22, $p = .19$). Although CASL did not have a statistically significant impact on achievement at either grade level, CASL affected students in Cohort 1 and Cohort 3 differently, with a greater impact in Cohort 3. A statistically significant CASL-by-grade interaction of –5.81 scale score points (standard error = 1.03, $p < .0001$) was found by estimating the benchmark model extended with an interaction of the intervention indicator with the grade indicator not grand-mean-centered and coded as –1 for students in Cohort 3 and +1 for students in Cohort 1. It should be noted, however, that this statistically significant difference between Cohort 1 and Cohort 3 does not mean that CASL had any statistically significant impacts overall or for either cohort separately.

Sensitivity analyses

Sensitivity analyses were conducted only for the primary outcome of student achievement to test the robustness of the findings to methodological decisions regarding the use of achievement as a pretest covariate in the modeling of potential intervention impacts, the estimation methods, and the methods for dealing with missing data. Unless otherwise indicated, the sensitivity analyses used the benchmark impact model presented in chapter 2 and appendix F and the multiple imputed datasets. Table 4.1 shows the results of the no-covariate model used to test the sensitivity of the impact analysis to the inclusion of achievement pretest covariates.

Estimation method

The second sensitivity analysis tested the robustness of the benchmark impact estimates to the estimation method. The benchmark impact estimates for the primary outcome used the residual (restricted) maximum likelihood method. Sensitivity analyses were conducted using the maximum likelihood estimation method and the minimum variance quadratic unbiased estimation of the covariance parameters estimation method. The maximum likelihood estimation method failed to yield a statistically significant impact estimate for CASL, $\gamma_{01} = 0.58$, standard error = 3.16, $p = .85$, confirming the benchmark analysis that also failed to find a statistically significant impact of CASL. Similarly, results from the minimum variance quadratic unbiased ($\gamma_{01} = 0.59$, standard error = 3.01, $p = .85$) also failed to yield a statistically significant impact of CASL.

Case deletion treatment of missing data

To test the robustness of the findings to the treatment of missing data, the benchmark impact model was fit to data where students with missing pre- and/or posttest data were deleted. All 67 schools in the study were included in these sensitivity analyses. Total sample sizes for these analyses were 1,548 grade 4 intervention students, 1,312 grade 5 intervention students, 1,826 grade 4 control students, and 1,553 grade 5 control students. The intervention impact was estimated as 0.51 scale score point (standard error= 4.38, $p = .91$). The impact estimate from this sensitivity analysis was similar to the impact estimate from the benchmark model of 0.58 scale score points (standard error = 3.47) shown in table 4.1. Neither the benchmark impact estimate nor the sensitivity analysis estimate was statistically significant. The similarity of these estimates suggests that the benchmark impact estimate was robust to the method of treating missing data. In other words, results regarding the impact of CASL on students' mathematics achievement were not affected by the treatment of missing data.

Summary

No statistically significant difference was found between the mathematics achievement of students in the CASL group and students in the control group. Results of the sensitivity analyses revealed that the statistical significance of the impact estimate were invariant to the use of covariates in the analytic model, the estimation method, and the method used to treat missing data.

Chapter 5. Exploratory analysis of intermediate outcomes

This chapter presents the results of the exploratory analyses used to estimate the impact of CASL on the intermediate outcomes: student motivation and teacher assessment knowledge, teacher assessment practice, and teacher involvement of students in assessment.

Student motivation impact analyses

The student motivation impact model nested students within schools and schools within districts (block). Prior to fitting the student motivation data to the model, an unconditioned model with no covariates was estimated to provide estimates of variance components (see appendix I). These estimates confirmed that motivation survey scores of students within the same school were related to each other and thus warranted the use of multilevel modeling to estimate impacts.

Table 5.1 compares schools in the intervention and control groups on their means from the Survey of Student Motivation, unadjusted for pretest data. (Appendix J provides raw means, and appendix K provides complete results from the impact analyses.) Because student motivation was examined without the use of a pretest covariate, no comparisons were conducted between pretest means or posttest means adjusted for pretest data. No statistically significant parameter estimates were found for either the intervention impact for Wave 3 (at the end of the CASL training year) or Wave 5 (at the end of the CASL implementation year). The Wave 3 student motivation impact analysis sample included students in Cohort 1 (grade 4 during the CASL training year) and Cohort 2 (grade 5 during the CASL training year). The Wave 5 (posttest) student motivation impact analysis sample included students in Cohort 1 (grade 5 during the CASL implementation year) and Cohort 3 (grade 4 during the CASL implementation year). All analyses accounted for the clustering of students within schools and estimates were combined across the 40 imputed student motivation datasets.

Table 5.1. Intervention and control group means and estimated differences on student motivation at posttest and estimated impact of Classroom Assessment for Student Learning on student motivation

Measure	Intervention group mean	Control group mean	Estimated difference[a]	95 percent confidence interval	*p*-value	Effect size[b]
Wave 3 unadjusted score	3.29 (0.02) Schools = 28 Students = 1,179	3.28 (0.02) Schools = 27 Students = 2,579	0.01 (0.03)	–0.05–0.06	.79	0.02
Wave 5 (posttest) unadjusted score	3.33 (0.02) Schools = 24 Students = 2,016	3.32 (0.02) Schools = 32 Students = 3,170	0.01 (0.03)	–0.04–0.06	.63	0.03

a. Estimated difference may not equal difference between means because of rounding.
b. Calculated as the estimated difference divided by control group standard deviation.
Note: Numbers in parentheses are standard errors. All results are based on analysis using a mixed-model approach to account for the sources of variability in the data that resulted from the nested structure of the school environment.
Source: Survey of Student Motivation.

Teacher outcome impact analysis

The teacher outcome impact model nested teachers within school and schools within districts (block). Prior to fitting the teacher outcome data to the model, an unconditioned model with no covariates was estimated to provide estimates of variance components (see appendix I). These estimates confirmed that scores on the outcomes of teachers within the same school were related to each other and thus warranted the use of multilevel modeling to estimate impacts.

Table 5.2 shows the comparison of schools in the intervention and control groups on the teacher pretest measure and the three teacher outcome measures adjusted for the pretest data. The only statistically significant impact of CASL was on the Test of Assessment Knowledge; intervention teachers scored 2.78 points higher than did control teachers (standard error = 0.99, $p = .01$). All analyses accounted for the clustering of teachers within schools, and estimates were combined across the 40 imputed teacher outcome datasets.

Table 5.2. Intervention and control group means and estimated differences on teacher outcomes at pre- and posttest and estimated impact of Classroom Assessment for Student Learning on teacher outcomes

Measure	Intervention group mean Schools = 33 Teachers = 178	Control group mean Schools = 34 Teachers = 231	Estimated difference[a]	95 percent confidence interval	*p*-value	Effect size[b]
Pretest score (Baseline: November, 2007)	36.49 (0.54)	35.91 (0.47)	0.58 (0.69)	–0.78 to 1.94	.40	0.09
Test of Assessment Knowledge (Wave 5 posttest: May 2009)	41.36 (0.76)	38.58 (0.60)	2.78 (0.99)	0.84 to 4.72	.01	0.42
Teacher Assessment Work Sample (Wave 5 posttest: May 2009)	1.61 (0.05)	1.60 (0.04)	0.01 (0.06)	–0.10 to 0.13	.85	0.03
Teacher Report of Student Involvement (Wave 5 posttest: May 2009)	0.39 (0.02)	0.34 (0.02)	0.05 (0.03)	–0.01 to 0.11	.10	0.26

a. Estimated difference may not equal difference between means because of rounding.
b. Calculated as the estimated difference divided by control group standard deviation.
Note: Numbers in parentheses are standard errors. Estimated differences may not equal difference between means due to rounding.
Source: Teacher outcome data.

Chapter 6. Summary of findings and study limitations

This chapter summarizes the findings from the Classroom Assessment of Student Learning (CASL) implementation analyses and the analyses of the professional development surveys. Also summarized are the impact of CASL on student achievement, the sensitivity analyses, and the impact of CASL on intermediate student and teacher outcomes. Finally, this chapter describes the context and the limitations of the study.

Findings of Classroom Assessment for Student Learning implementation

Most intervention teachers (67.18 percent) in this study reported learning team experiences that met the CASL developers' criteria, but less than half (41.55 percent) reported at least partially reading each textbook chapter. Teachers reported spending, on average, 31.21 hours on CASL activities, as compared with the 60 hours recommended by the developer. Therefore, many participating teachers did not complete the program as the developers had recommended. Because the criteria were based on developer recommendations rather than empirical studies of program implementation, however, it is not possible to conclude that the teachers in this study spent less (or more) time on CASL than is customary in program implementation.

In terms of non-CASL professional development experiences, more control teachers participated in professional development in Wave 3 than intervention teachers did. Control teachers reported less involvement in teacher networks and less involvement in teacher study group activities in Wave 2 than did intervention teachers. Also in Wave 2, more teachers in the intervention group participated in mathematics-related professional development than did teachers in the control group. Control teachers reported more emphasis of professional development related to student involvement in both Wave 2 and Wave 4 than did intervention teachers. Control teachers also reported more emphasis of professional development on use of technology in instruction and on strategies for teaching diverse student populations in Wave 3 than did intervention teachers. Intervention teachers reported more emphasis in their non-CASL professional development on curriculum in Wave 5 (posttest) than did control teachers. There were nine statistically significant differences between the intervention group and control group in terms of their non-CASL professional development experiences, out of 124 comparisons. It is possible that these were chance differences because six statistically significant differences would be expected by chance given the minimum threshold for statistical significance (.05) and the number of comparisons (124). These results suggest that the only difference between the professional development experiences of the teachers in the intervention group and the teachers in the control group was the CASL intervention.

Impact of Classroom Assessment for Student Learning on student achievement

Results from this study show that CASL did not have a statistically significant impact on student mathematics achievement. The average school impact was estimated at less than one scale score point on the mathematics portion of the CSAP. In addition, separate follow-up exploratory

analyses failed to yield a statistically significant impact of CASL for either Cohort 1 or Cohort 3 on mathematics achievement.

Sensitivity analyses were conducted to test the robustness of the findings for student achievement to methodological decisions regarding the use of achievement as a pretest covariate, the estimation methods, and the method for dealing with missing data. In other words, the sensitivity analyses test whether the impact findings were affected by analytic decisions made by the researcher or by the analytic methods used. Results from the sensitivity analyses did not differ from the benchmark model in the statistical significance of the impact findings: none of the sensitivity analyses yielded a statistically significant impact of the intervention. These sensitivity analyses results suggest that the benchmark findings regarding the impact of CASL on student achievement are robust to the use of covariates, estimation methods, and treatment of missing data. That is, decisions made by the research team about the analytic model, the method for dealing with missing data, and the statistical estimation method used in the analysis did not influence the main findings of the study.

Impact of Classroom Assessment for Student Learning on intermediate outcomes

CASL was not found to have a statistically significant impact on student motivation to learn. Differences between intervention and control student scores on the Survey of Student Motivation were estimated to be less than one point on the survey scale. CASL group teachers scored higher on the Test of Assessment Knowledge by a statistically significant margin than did teachers assigned to the control group. CASL did not have a statistically significant impact at the school level on the other two teacher outcomes: assessment practice or involvement of students in classroom assessment.

Context of the study of Classroom Assessment for Student Learning

This study was designed to test the impact of CASL when implemented under real-world conditions. The purpose was to estimate the impact of CASL on student achievement and other outcomes as if schools had purchased the intervention on their own and implemented it without interference or guidance from the research team. No steps were taken by the research team to influence the level of intervention implementation undertaken by the participants in the intervention schools. Results of this study suggest that if schools purchase CASL and implement it under similar conditions as those found in this study, little impact on student performance on state mathematics tests may be realized from two years of CASL implementation.

This study had sufficient statistical power to detect an impact on student scores on the statewide achievement test of 0.25 standard deviation, which represents approximately 5 and ½ months of instruction in Colorado. The original power analysis conducted to estimate the sample size necessary to achieve the desired level of statistical power (described in Appendix A) assumed attrition rates at the school and student level that were not experienced in this study. In the case of all outcomes—both primary and intermediate—sample sizes at the school, student, and teacher levels exceeded the samples sizes estimated by the original power analyses as necessary

to achieve sufficient statistical power. Attrition and non-response, therefore, did not reduce the study's statistical power to below what was originally estimated.

Impact analyses on student achievement included all schools analyzed as originally randomized at the beginning of the study. As a result, the analyses provided unbiased estimates of the impact of CASL, implemented under real-world conditions by the study sample schools, on student mathematics achievement in grades 4 and 5. Although attrition reduced response rates on the intermediate outcomes, the analysis samples used to estimate impact on the intermediate outcomes did not differ by a statistically significant margin on any school characteristics. In addition, multiple imputation methods were used to impute missing data, so few students or teachers were excluded from the impact analysis samples.

Study limitations

Although this study was designed to provide an unbiased estimate of the impact of CASL on mathematics achievement and other outcomes, there are limitations that should be considered when interpreting the study findings. This study was designed to have the statistical power to detect an impact of 0.25 standard deviation on student achievement and was not designed to have the statistical power to detect effects smaller than .25 standard deviation. It also should be pointed out that it is not correct to interpret impact estimates that were not statistically significant as evidence of no impact. Rather, these estimates failed to provide any evidence of an impact. Additional limitations are discussed in two broad categories: external validity/generalizability and potential for bias because of attrition and nonresponse.

External validity

External validity concerns whether the findings can be generalized to variations in the implementation or CASL or to different settings, populations, or outcomes.

- Results of this study do not generalize to implementation of CASL under different conditions, including either lower levels of implementation or higher levels of implementation. Thus, results presented in this report provide evidence of the impact of CASL only under the implementation conditions observed in this study.
- Results of this study do not generalize to practices of formative assessment in general.
- Results from this study are applicable only to student mathematics achievement in grades 4 and 5.
- This study used a sample of convenience: all participating schools, teachers, and students were volunteers. This limits generalizability of the study findings to this voluntary sample. It may be that the study findings generalize to relatively disadvantaged urban and rural schools with high proportions of Hispanic students (see table 2.2).
- This study examined a single primary outcome measured by the statewide test of mathematics. Findings may not generalize to other states with different achievement tests.

Potential for bias because of nonresponse and attrition

Nonresponse and attrition can create the potential for bias when the participants (schools, students, or teachers in this study) who are included in the impact analysis for one group differ systematically from the participants included in the impact analysis sample for the other group. The scope of potential bias because of attrition was minimized by the use of multiple imputations for missing data for the student achievement outcome and for all teacher outcomes.

- *Impact analysis on student achievement.* Nonresponse and attrition were not an issue with the student achievement outcome and were at levels expected to result in acceptable levels of bias (What Works Clearinghouse 2008). See the section on attrition and nonresponse in chapter 2 for details.
- *Impact analysis on Wave 3 student motivation.* Nonresponse exceeded levels considered acceptable and required the establishment of baseline equivalence of the impact analysis sample (What Works Clearinghouse 2008). Comparisons of the baseline characteristics (grade 4 and grade 5 student mathematics achievement from the spring 2007 CSAP and school demographic characteristics from the Common Core of Data) of the sample of schools in the impact analysis failed to yield any statistically significant differences between the intervention and control groups.
- *Impact analysis on Wave 5 (posttest) student motivation.* Similar to the Wave 3 motivation data, nonresponse exceeded levels considered acceptable (What Works Clearinghouse 2008). The sample of schools included in the impact analysis did not differ by statistically significant margins on their baseline characteristics (grade 4 and grade 5 student mathematics achievement from the spring 2007 CSAP and school demographic characteristics from the Common Core of Data).
- *Impact analyses on teacher outcomes.* Nonresponse exceeded levels considered to result in acceptable levels of bias (What Works Clearinghouse 2008) and required the establishment of baseline equivalence of the analysis sample. The impact analysis included intervention and control schools that did not differ by a statistically significant margin in terms of their baseline characteristics (grade 4 and grade 5 student mathematics achievement from the spring 2007 CSAP and school demographic characteristics from the Common Core of Data). In addition, teachers in the intervention group had the added burden of completing the CASL participant logs, which control group teachers did not have. Nonresponse was higher among intervention group teachers than control group teachers for each teacher outcome. Missing teacher data were imputed using a multiple imputation approach that excluded 41 teachers and created an impact analysis sample of 409 teachers, resulting in an acceptable level of attrition (What Works Clearinghouse 2008).

Appendix A. Power analysis

To determine the sample size necessary to detect the impact of the intervention, the study team conducted two power analyses, one for student achievement and one for teacher outcomes. All power analyses were conducted using Optimal Design software (Liu, Spybrook, Congdon, Martinez, & Raudenbush 2006), made specifically for power analyses for hierarchical cluster randomized designs. Random assignment of schools to the intervention or control groups was blocked by district. Sample and cluster size were chosen to achieve a high level of power, greater than .80. The study team chose conservative parameter estimates for the analyses to avoid overestimating power. Rationales for the estimates for effect size, intraclass correlation, and the covariate are described below. Power analyses were conducted for fixed effects. Power analyses were adjusted to reflect the use of covariates to increase precision.

The assumed minimum detectable effect for this study was 0.25 standard deviation, for a number of reasons. First, given the cost of the Classroom Assessment for Student Learning (CASL) intervention—approximately $1,000 per school—an effect of this size on student achievement was considered worth detecting. An increase of one-quarter of a standard deviation in student achievement represents a practically significant effect, equivalent to an increase of 10 percentile points. An effect of .25 standard deviation also represents approximately 5 and ½ months of instruction at grades 4 and 5, based on the growth in scale scores on the CSAP between the two grades. Second, while no empirical evidence was available on the impact of CASL, an effect size of 0.25 is at the lower end of what is reported in the literature for classroom assessment practices and strategies. Effect sizes from the literature on classroom assessment vary according to the type of assessment intervention and the outcome measure. In their summary of the formative assessment literature, Black & Wiliam (1998b) report that the effects of classroom assessment on student achievement range from 0.40 to 0.70. A recent study on the effects of professional development in classroom assessment found an average effect size of 0.32 after six months of teacher training (Wiliam et al. 2004). Additionally, a study of mathematics students in grades 5 and 6 found an effect size of 0.40 for student self-evaluation (Ross, Hogaboam-Gray, & Rolheiser 2002), which is included in CASL's student involvement component. The study team chose a conservative effect size because the effect sizes found in the literature come from many different types of studies that vary both in implementation length and degree of rigor. In addition, the Wiliam et al. (2004) study and the Ross et al. (2002) study included potentially more intensive training than might be experienced with CASL and used outcome measures likely to be more sensitive than the standardized test used for this study. Third, this study was designed so that teacher training in CASL occurred during the academic year prior to fully implementing the program during the next academic year. Students were exposed to the intervention practices and principles for at least an entire year. Wiliam et al. (2004) found an average effect on student achievement of 0.32 after six months of teacher training, so anticipating an effect size of 0.25 is not unreasonable due to the level of exposure to the intervention. Fourth, classroom assessment literacy is low (Plake, Impara, & Fager 1993), which suggests that there is room for improvement in classroom assessment practices. Because the research literature provides some suggestions that sound formative assessment can impact student achievement, an effect of 0.25 standard deviation seemed reasonable but still conservative and worthwhile to detect.

A conservative value of .15 was selected for the intraclass coefficient based on a review of the following sources. Raudenbush, Spybrook, Liu, & Congdon (2006) cite typical intraclass correlation coefficients for educational achievement between .05 and .15. Schochet (2008b) states that the intraclass coefficients for standardized test scores often range between .10 and .20. Bloom et al. (2005) found intraclass coefficients in Grade 5 reading and math ranged from .12 to .29 across five different districts.

Prior achievement was selected as a cluster-level covariate. Schochet (2008b) concludes that the proportion of variance explained by pretest measures is at least .50 when student-level data are used. Bloom, Bos, & Lee (1999) found similar values. Bloom, Richburg-Hayes, & Black (2005) found values ranging from .33 to .81 across five districts for school-level pretests. The proportion of .50 of posttest variance explained by pretest scores was chosen for this power analysis to be an appropriately conservative estimate.

A power analysis for the outcome of student achievement was conducted using the above parameter values and an assumption that 60 students would be nested within each school. The assumption of 60 students per school assumed 15 students per classroom and four classrooms per school. This number accounts for student mobility and the potential attrition of teachers within schools. Optimal Design software calculated that 47 clusters (approximately 24 intervention and 24 control clusters) were necessary to achieve the desired power of greater than .80 for the student achievement outcome. To account for possible attrition at the school and teacher/classroom levels, a recruiting target of 64 schools was set.

The study team also conducted power analyses to determine the sample size necessary to detect the impact of the intervention on the intermediate teacher outcomes. Parameter estimates for effect size, intraclass correlation, and proportion of posttest variance explained by the pretest measure were chosen for the following reasons.

First, few rigorous studies were found that explicitly examined the impact of any type of professional development on teacher outcomes. A study of teacher assessment competencies using a national sample found an effect size of 0.20 on a test of knowledge favoring teachers who had taken a graduate-level measurement course as compared with teachers who had not taken a course (Plake, Impara, & Fager 1993). O'Sullivan & Johnson (1993) found that teacher assessment competencies increased an average of one standard deviation after taking a graduate course in assessment. They also found that teachers who had completed a graduate course in assessment scored two standard deviations higher on classroom assessment performance tasks than did teachers who had not completed a graduate course in assessment. Evaluations of the effects of training in standards-aligned classrooms found effects ranging from approximately 0.50 to 1.00 standard deviations for the effect on teacher familiarity and use of standards in instruction and assessment (Wolfe & Jarvinen 2002, 2003). An estimated effect size of 0.50 on teacher outcomes was assumed based on these findings.

Second, little empirical evidence could be found regarding estimates of intraclass correlations or covariation for teachers. A value of .10 was used as the estimate of the intraclass correlation based on the assumption that there would be slightly less shared variance between teachers than between students. A conservative value of .20 was assumed for the correlation between teacher

pretest scores on the test of assessment knowledge and teacher outcomes for two reasons. First, teacher scores were assumed to be relatively unstable due to variations in implementation of the training as well as variations in other professional development initiatives across schools. Second, scores on the Test of Assessment Knowledge administered at baseline were used as the pretest covariate for all teacher outcomes and the correlation between this measure and the other outcome measures was not known.

Finally, an assumption of four teachers per cluster was used to estimate final sample size for the power analysis for teacher effects. This analysis was also based on an estimated effect size of 0.50, a proportion of postintervention variance explained by preintervention test scores of .20, and an intraclass correlation of .10.

Using this set of assumptions, Optimal Design software was used to estimate that 41 clusters were necessary to achieve a power of greater than .80 for the teacher outcomes. Given that more schools were estimated to be needed for the analysis of student achievement and that student achievement was the primary outcome, the target sample size was set at 64 schools.

Appendix B. Response rates by data collection wave, instrument, and experimental group

Table B1. Response rates by data collection wave, instrument, and experimental group

Wave	Total			Intervention			Control			
	Sample size (N)	Participated (N)	Response rate (percent)	Sample size (N)	Participated (N)	Response rate (percent)	Sample size (N)	Participated (N)	Response rate (percent)	*p*-value[a]
Baseline										
Teacher Background Information[b]	368	317	86.14	175	147	84.00	193	170	88.08	.26
Test of Assessment Knowledge[b]	368	289	78.53	175	134	76.57	193	155	80.31	.38
Teacher Assessment Work Sample	368	239	64.95	175	104	59.43		135	69.95	.04
Wave 2: December 2007										
Survey of Professional Development	368	271	73.64	175	117	66.86	193	193	154	79.79	<.01
Teacher Report of Student Involvement	368	246	66.85	175	101	57.71		193	145	75.13	<.01
Wave 3: May 2008										
Survey of Professional Development	368	293	79.62	175	131	74.86		193	162	83.94	.03
Test of Assessment Knowledge	368	287	77.99	175	123	70.29		193	164	84.97	<.01
Teacher Report of Student Involvement	368	265	72.01	175	112	64.00		193	153	79.27	<.01
Student Survey of Motivation	368	215	58.42	175	85	48.57		193	130	67.36	<.01
Late-entry baseline										
Teacher background information[c]	82	61	74.39	29	18	62.07		53	43	81.13	.06
Wave 4: December 2008										
Survey of Professional Development	409	285	69.68	178	115	64.61		231	170	73.59	.05
Original teachers	337	221	65.58	158	99	62.66		179	122	68.16	.29
Late-entry teachers	72	64	88.89	20	16	80.00		52	48	92.31	.14
Teacher Report of Student Involvement	409	252	61.61	178	100	56.18		231	152	65.80	.05
Original teachers	337	197	58.46	158	88	55.70		179	109	60.90	.33
Late-entry teachers	72	55	76.39	20	12	60.00		52	43	82.69	.04
Posttest: May 2009										
Survey of Professional Development	409	286	69.93	178	115	64.61		231	171	74.03	.04
Original teachers	337	220	65.28	158	98	62.03		179	122	68.16	.24
Late-entry teachers	72	66	91.67	20	17	85.00		52	49	94.23	.20
Test of Assessment Knowledge	409	283	69.19	178	114	64.04		231	169	73.16	<.05
Original teachers	337	217	64.39	158	97	61.39		179	120	67.04	.28
Late-entry teachers	72	66	91.67	20	17	85.00		52	49	94.23	.20
Teacher Assessment Work Sample	409	250	61.12	178	94	52.81			156	67.53	<.01
Original teachers	337	193	57.27	158	81	51.27		179	112	62.57	.04
Late-entry teachers	72	57	79.17	20	13	65.00	231	52	44	84.62	.07

Teacher Report of Student Involvement	409	222	54.28	178	86	48.31	231	136	58.87	.03
Original teachers	337	171	50.74	158	73	46.20	179	98	54.75	.09
Late-entry teachers	72	51	70.83	20	13	65.00	52	38	73.08	.50
Student Survey of Motivation	409	273	66.75	178	107	60.11	231	166	71.86	.01
Original teachers	337	210	62.31	158	91	57.59	179	119	66.48	.09
Late-entry teachers	72	63	87.50	20	16	80.00	52	47	90.38	.23

a. *p*-values are from 2 x 2 chi-square tests comparing frequency counts of instruments submitted versus not submitted for treatment versus control groups.
b. Includes original teachers only.
c. Includes late-entry teachers only.
Source: Teacher surveys.

Appendix C. Data collection instruments

Survey of Teacher Background

Please answer the following questions for the 2008-2009 school year.

1. What is your primary role?

- ☐ 4th grade teacher
- ☐ 5th grade teacher
- ☐ Title I teacher
- ☐ Special education teacher
- ☐ Gifted and talented teacher
- ☐ Principal
- ☐ Assistant principal
- ☐ Other

2. Do you teach math to 4th graders?

- ☐ Yes
- ☐ No

3. Do you teach math to 5th graders?

- ☐ Yes
- ☐ No

4. Including this year, how many years have you:

- 4a. been a teacher? _____
- 4b. taught your current grade level(s)? _____
- 4c. taught math? _____
- 4d. taught math to your current grade level(s)? _____
- 4e. worked at your current school? _____

5. How many students are enrolled in your class? _____

6. Approximately what percentage of your students have been in your class since the beginning of the school year? _____

7. How similar are the math curriculum and instruction in your class to that of other teachers at your grade level in your school?

- ☐ Not at all similar
- ☐ Somewhat similar
- ☐ Very similar

8. What is your most advanced degree?

- ☐ Bachelor's

- ☐ Master's
- ☐ Ph.D. or Ed.D.
- ☐ Other

9. What is your gender?
   - ☐ Male
   - ☐ Female

11. What is your race? Select one or more races to indicate what you consider yourself to be.
   - ☐ American Indian or Alaska Native
   - ☐ Asian
   - ☐ Black or African-American
   - ☐ Hispanic (non-white)
   - ☐ Native Hawaiian or other Pacific Islander
   - ☐ White

12. Do you need to elaborate on any of these questions? If so, please indicate the question number and your explanation.

_____
_____
_____

Thank you for taking our survey. Your response is very important to us.

Survey of Professional Development

1. Did you participate in any professional development activities at any time from May 10, 2008 to December 5, 2008?

   ☐ Yes

   ☐ No

1a. If yes, please check all that apply. (Please do not include CASL if you are participating in CASL training.)

   ☐ Workshop

   ☐ Institute

   ☐ Conference

   ☐ College course

   ☐ Teacher network

   ☐ Internship or immersion activity

   ☐ Teacher committee or task force

   ☐ Teacher study group

   ☐ Work with a mentor or coach

   ☐ Other professional development activity

A1. Please give the name of the professional development activity you participated in during the time period between May 10, 2008 to December 5, 2008 that you feel impacted your practice the most. Please do not include CASL if you are participating in CASL training.

Count a program of on-going professional development that took place on different dates over several weeks or months, such as a summer institute with follow-up workshops or an on-going teacher study group, as ONE professional development activity.

_____

_____

_____

A2. Have you reported on this activity in a previous log entry?

   ☐ Yes

   ☐ No

   ☐ I do not remember

A3. Please briefly describe the topic and purpose of this activity

A4. Over what period of time was/is the activity spread, including the main activity and any formal preliminary or follow-up sessions?

   ☐ Less than one day

   ☐ One day

   ☐ Two to four days

   ☐ A week

- ☐ Two to three weeks
- ☐ A month
- ☐ Two to six months
- ☐ Seven months to a year
- ☐ More than a year

A5. As part of this activity, did you meet regularly, over the course of several weeks or months, with a group of educators to discuss and reflect on the material being learned?

- ☐ Yes
- ☐ No

A6. Which subject area(s) did the activity cover? (Check all that apply.)

- ☐ Math
- ☐ Science
- ☐ Reading
- ☐ Writing
- ☐ Social studies
- ☐ Health related
- ☐ Activity was not specific to any one subject area
- ☐ Other subjects (s) (please specify):

How much emphasis did the activity give to...

A7. Curriculum (e.g., units, textbooks, standards)?

- ☐ No emphasis
- ☐ Minor emphasis
- ☐ Major emphasis

A8. Working with content standards (e.g., understanding, unpacking , simplifying, aligning instruction to standards)?

- ☐ No emphasis
- ☐ Minor emphasis
- ☐ Major emphasis

A9. Instructional methods?

- ☐ No emphasis
- ☐ Minor emphasis
- ☐ Major emphasis

A10. Increasing student involvement in learning?

- ☐ No emphasis
- ☐ Minor emphasis
- ☐ Major emphasis

A11. Formative assessments (e.g., developing, selecting, and using assessment in the classroom)?

☐ No emphasis

☐ Minor emphasis

☐ Major emphasis

A12. Communicating assessment results to students?

☐ No emphasis

☐ Minor emphasis

☐ Major emphasis

A13. Using assessment results to guide instruction (i.e., to make adjustments in instructional strategies or lesson plans)?

☐ No emphasis

☐ Minor emphasis

☐ Major emphasis

A14. Use of technology in instruction (e.g., computers, graphing calculators)?

☐ No emphasis

☐ Minor emphasis

☐ Major emphasis

A15. Strategies for teaching diverse student populations (e.g., students with disabilities, from underrepresented populations, economically disadvantaged, range of abilities)?

☐ No emphasis

☐ Minor emphasis

☐ Major emphasis

A16. Leadership development?

☐ No emphasis

☐ Minor emphasis

☐ Major emphasis

A17. State-wide assessment or standardized testing?

☐ No emphasis

☐ Minor emphasis

☐ Major emphasis

A18. Other - Please describe:

_____

_____

_____

☐ No emphasis

☐ Minor emphasis

☐ Major emphasis

A19. Between May 10, 2008 and December 5, 2008, including the main activity and any preliminary activities or formal follow-up sessions, how many hours were you engaged in this activity overall? Round your answer to the nearest whole hour. _____

A20. Please indicate if you engaged in any of the following during this activity. (Check all that apply.)

- ☐ Had someone observe and provide feedback on your teaching
- ☐ Presented material or instructed others
- ☐ Led a discussion

A21. Have you discussed or shared what you learned with others in your school who did NOT attend the activity?

- ☐ Yes
- ☐ No

A22. Was this activity consistent with your own goals for your professional development?

- ☐ Yes
- ☐ No

A23. Was this activity aligned with state content standards?

- ☐ Yes
- ☐ No

A24. Please rate the overall quality of the activity, including the main activity and any preliminary activities or formal follow-up sessions.

- ☐ Excellent
- ☐ Good
- ☐ Fair
- ☐ Poor

A25. Please indicate the degree of impact you expect the activity to have on your classroom practices.

- ☐ High
- ☐ Medium
- ☐ Low
- ☐ None

Thank you for taking our survey. Your response is very important to us.

CASL participant logs

Learning team startup

1. Today's date (enter in this format: MM/DD/YYYY): _____

2. How many people are on your Learning Team? _____

3. Is membership on the Learning Team voluntary or mandatory?

- ☐ Voluntary
- ☐ Mandatory

4. Will any team member(s) be acting as manager or facilitator of the team?

- ☐ Yes
- ☐ No

4a. How is the manager role assigned?

- ☐ One or two person(s) will serve as manager for the entire life of the team
- ☐ The manager role rotates amongst members
- ☐ Different members fill different managerial responsibilities
- ☐ Other (describe) _____

4b. What are the team manager(s)' responsibilities? Check all that apply.

- ☐ Post the schedule of team meetings
- ☐ Complete the team meeting log
- ☐ Monitor meeting time so all members have an opportunity to share
- ☐ Bring materials needed for the meeting
- ☐ Other (describe) _____

5. Did your learning team establish group operating principles?

- ☐ Yes
- ☐ No

5a. What group operating principles did your team establish? Check all that apply.

- ☐ Commit to doing the work—the reading & activities we select
- ☐ Commit to attending all meetings
- ☐ Stick to the topic or task during the meeting
- ☐ Keep the focus on the students
- ☐ Involve everyone; make sure all voices are heard
- ☐ Be an active listener; seek to understand as well as be understood
- ☐ Other (describe)

6. Did your team set a meeting schedule?

- ☐ Yes
- ☐ No

7. Has your team chosen a regular place to meet?

☐ Yes

☐ No

8. Did your team establish a reading and assignment schedule?

☐ Yes

☐ No

9. Do you and your Learning Team members have a common purpose or goal for your Learning Team?

☐ Yes

☐ No

9a. If your team has a common purpose or goal, please state it below.

_____
_____
_____

Thank you for taking our survey. Your response is very important to us.

CASL participant log

For Chapter 1: Classroom Assessment: Every Student a Winner!

1. Today's date (enter in this format: MM/DD/YYYY): _____

2. Did you read the CASL text Chapter 1 (“Classroom Assessment: Every Student a Winner!”)?

☐ Yes, completely

☐ Yes, partially

☐ No, not at all

3. Please check which Individual Study and Reflection activities you completed.

☐ 1.1 Program Introduction

☐ 1.2 Emily's Interview

☐ 1.3 Case Comparison: Emily and Krissy

☐ 1.4 Case Comparison: Emily and Mr. Heim's Class

☐ 1.5 Evaluating Assessment Quality

☐ 1.6 Watch Video, "Assessment for Student Motivation"

☐ 1.7 Classroom Assessment Confidence Questionnaire

4. How many Learning Team meetings did your team have on this chapter? _____

5. Of these meetings, how many did you attend? _____

6. How useful was/were the Learning Team Meeting(s) on this chapter? _____

☐ I did not attend any meetings.

☐ Not at all useful

☐ A little useful

☐ Somewhat useful

☐ Very useful

7. To what extent did the Learning Team meeting(s) for this chapter focus on important content of the chapter?

- ☐ I did not attend any meetings.
- ☐ Hardly at all
- ☐ A little
- ☐ To some extent
- ☐ To a great extent

8a. During the Learning Team meeting(s) for this chapter, to what extent did you share with the team ways that you have implemented the CASL techniques in your classroom?

- ☐ I did not attend any meetings.
- ☐ Hardly at all
- ☐ A little
- ☐ To some extent
- ☐ To a great extent

8b. During the Learning Team meeting(s) for this chapter, to what extent did you share with the team results you have seen from using the CASL techniques?

- ☐ I did not attend any meetings.
- ☐ Hardly at all
- ☐ A little
- ☐ To some extent
- ☐ To a great extent

9. How many total hours did you spend on Chapter 1, including reading, completing activities, trying out applications in the classroom, reflecting, & participating in Learning Team Meetings? Round your answer to the nearest hour. _____

10. Briefly list how you are applying what you're learning from CASL in your classroom practice:

_____
_____
_____

Thank you for taking our survey. Your response is very important to us.

CASL participant log

For Chapter 8: Personal Communication as Assessment

1. Today's date (enter in this format: MM/DD/YYYY): _____

2. Did you read the CASL text Chapter 8 (“Personal Communication as Assessment”)?

   ☐ Yes, completely

   ☐ Yes, partially

   ☐ No, not at all

3. Please check which Individual Study and Reflection activities you completed.

   ☐ 8.1 Learning Targets Best Assessed with Personal Communication

   ☐ 8.2 Generate Oral Questions

   ☐ 8.3 Practice Questioning Strategies

   ☐ 8.4 Scored Discussion

   ☐ 8.5 Journal Icons

4. Are you keeping a portfolio to track your learning with CASL?

   ☐ Yes

   ☐ No

5. Of the 10 "Additional Portfolio Entries to Represent Learning" from Parts 1 and 2 (pages 273-275), please indicate how many of the entries you completed. _____

6. How many Learning Team meetings did your team have on this chapter? _____

7. Of these meetings, how many did you attend? _____

8. How useful was/were the Learning Team Meeting(s) on this chapter?

   ☐ I did not attend any meetings.

   ☐ Not at all useful

   ☐ A little useful

   ☐ Somewhat useful

   ☐ Very useful

9. To what extent did the Learning Team meeting(s) for this chapter focus on important content of the chapter?

   ☐ I did not attend any meetings.

   ☐ Hardly at all

   ☐ A little

   ☐ To some extent

   ☐ To a great extent

10a. During the Learning Team meeting(s) for this chapter, to what extent did you share with the team ways that you have implemented the CASL techniques in your classroom?

   ☐ I did not attend any meetings.

   ☐ Hardly at all

   ☐ A little

☐ To some extent

☐ To a great extent

10b. During the Learning Team meeting(s) for this chapter, to what extent did you share with the team results you have seen from using the CASL techniques?

☐ I did not attend any meetings.

☐ Hardly at all

☐ A little

☐ To some extent

☐ To a great extent

11. How many total hours did you spend on Chapter 8 , including reading, completing activities, trying out applications in the classroom, reflecting, & participating in Learning Team Meetings? Round your answer to the nearest hour. _____

12. Briefly list how you are applying what you're learning from CASL in your classroom practice:

_____

_____

_____

13. Please indicate the amount of support your school administrator(s) has/have provided your Learning Team since you began CASL.

☐ Hardly any support at all

☐ A little support

☐ Some support

☐ A great deal of support

Thank you for taking our survey. Your response is very important to us.

CASL participant log

For Chapter 13: Practical Help with Standardized Tests

1. Today's date (enter in this format: MM/DD/YYYY): _____

2. Did you read the CASL text Chapter 13 ("Practical Help with Standardized Tests")?

☐ Yes, completely

☐ Yes, partially

☐ No, not at all

3. Please check which Individual Study and Reflection activities you completed.

☐ 13.1 Standardized Tests Used in Your District

☐ 13.2 A Definitions Pretest

☐ 13.3 Hills' Handy Hints

☐ 13.4 Interpret Your Own Standardized Test Report

☐ 13.5 Use Item Formulas to Help Students Learn

☐ 13.6 Translate Standardized Test Jargon into Student-Friendly Language

- ☐ 13.7 When Grades Don't Match the State Assessment Results
- ☐ 13.8 A Definitions Posttest

4. How many Learning Team meetings did your team have on this chapter? _____

5. Of these meetings, how many did you attend? _____

6. How useful was/were the Learning Team Meeting(s) on this chapter?
   - ☐ I did not attend any meetings.
   - ☐ Not at all useful
   - ☐ A little useful
   - ☐ Somewhat useful
   - ☐ Very useful

7. To what extent did the Learning Team meeting(s) for this chapter focus on important content of the chapter?
   - ☐ I did not attend any meetings.
   - ☐ Hardly at all
   - ☐ A little
   - ☐ To some extent
   - ☐ To a great extent

8a. During the Learning Team meeting(s) for this chapter, to what extent did you share with the team ways that you have implemented the CASL techniques in your classroom?
   - ☐ I did not attend any meetings.
   - ☐ Hardly at all
   - ☐ A little
   - ☐ To some extent
   - ☐ To a great extent

8b. During the Learning Team meeting(s) for this chapter, to what extent did you share with the team results you have seen from using the CASL techniques?
   - ☐ I did not attend any meetings.
   - ☐ Hardly at all
   - ☐ A little
   - ☐ To some extent
   - ☐ To a great extent

9. How many total hours did you spend on Chapter 13, including reading, completing activities, trying out applications in the classroom, reflecting, & participating in Learning Team Meetings? Round your answer to the nearest hour. _____

10. Briefly list how you are applying what you're learning from CASL in your classroom practice:

11. Please indicate the amount of support your school administrator(s) has/have provided your Learning Team since you began CASL.
   - ☐ Hardly any support at all
   - ☐ A little support

- ☐ Some support
- ☐ A great deal of support

Thank you for taking our survey. Your response is very important to us.

Test of Assessment Knowledge

1. The primary users of formative assessments are policy makers, program planners and school administrators.
   - ☐ a. True
   - ☐ b. False

For items 2-6, choose the type of learning that best represents the instructional objective.

	Knowledge	**Reasoning**	**Skill**	**Product**	**Disposition**
2. Choosing to read for enjoyment in language arts					
3. Correctly using lab equipment to gather data in science					
4. Comparing and contrasting cultural aspects of the English-speaking and Spanish-speaking worlds					
5. Defining prime numbers in math					
6. Intending to vote in elections in the future					

For items 7-14, choose the most appropriate form of assessment for each instructional goal.

7. Students will be able to correctly pronounce five Spanish verbs.
   - ☐ a. Portfolio
   - ☐ b. Extended Response
   - ☐ c. Rubric
   - ☐ d. Performance Assessment
8. Students will be able to describe the concept of supply and demand and how it affects prices.
   - ☐ a. Selected Response
   - ☐ b. Performance Assessment
   - ☐ c. Extended Response
   - ☐ d. Short Answer
9. Students will be able to supply two key facts about each character in the story.
   - ☐ a. Multiple-choice
   - ☐ b. Short Answer
   - ☐ c. Extended Response
   - ☐ d. Matching
10. Students will be able to identify the correct verb form for a sentence.
   - ☐ a. Multiple-choice
   - ☐ b. Short answer
   - ☐ c. True/False

☐ d. Oral Examination

11. Students will be able to lead a small-group discussion.
    - ☐ a. Essay
    - ☐ b. Portfolio
    - ☐ c. Performance Assessment
    - ☐ d. Selected Response
12. Students will be able to write a topic sentence for a paragraph they are given.
    - ☐ a. Extended Response
    - ☐ b. Short answer
    - ☐ c. Essay
    - ☐ d. Performance Assessment
13. Students will be able to describe how a bill becomes a law.
    - ☐ a. Multiple-choice
    - ☐ b. Oral examination
    - ☐ c. Extended Response
    - ☐ d. Fill-in the blank
14. Students will be able to correctly choose the definition of the word denominator.
    - ☐ a. Multiple-choice
    - ☐ b. Short answer
    - ☐ c. Extended response
    - ☐ d. True/False
15. Achievement, competence, and celebration are three basic purposes of which of the following?
    - ☐ a. Standardized tests
    - ☐ b. Oral examination
    - ☐ c. Portfolios
    - ☐ d. Extended response

For questions 16 and 17, choose whether a norm-referenced or criterion-referenced test is more appropriate for the situation.

16. Using a rubric to determine whether an essay written by a student deserves an A.
    - ☐ a. Norm-referenced test
    - ☐ b. Criterion-referenced test
17. Selecting the five lowest-performing math students for an afterschool tutoring program.
    - ☐ a. Norm-referenced test
    - ☐ b. Criterion-referenced test
18. You are writing a test to assess student learning of a set of standards. Which of the following should you consider in deciding how many items to write for each standard?
    - ☐ a. Student proficiency levels

☐ b. Balance between different assessment methods

☐ c. Possible sources of bias

☐ d. Breadth and depth of learning objectives

19. As a classroom teacher, you need to assess student knowledge of a large number of facts. Which would be the most effective assessment for this task?

☐ a. Performance Assessment

☐ b. Oral question and answer

☐ c. Essay

☐ d. Multiple-choice

20. Performance assessment is a good way to get students involved in assessment.

☐ a. True

☐ b. False

21. Possible limitations of True/False questions include:

☐ a. Can be hard to identify plausible distractors

☐ b. The process of elimination can skew scores

☐ c. Guessing can skew scores

☐ d. Cannot measure a variety of objectives

22. When writing fill-in the blank items, it is best to have only one blank per question for students to complete.

☐ a. True

☐ b. False

23. With "matching" questions (i.e., those that require students to match items in one column with the correct items in another), both columns must contain the same number of items.

☐ a. True

☐ b. False

24. Which of the following types of learning is NOT suitable to being assessed using performance assessment?

☐ a. Performance of a task

☐ b. Recall of facts

☐ c. Reasoning skills

☐ d. Production of a product

25. Which of the following is the most appropriate assessment method to use with very young students?

☐ a. Fill-in the blank

☐ b. Performance assessment

☐ c. True/ false test

☐ d. Portfolio assessment

26. Conferences, class discussions, journals, and logs are all varieties of the following:

☐ a. Selected response assessment

☐ b. Short answer assessment

☐ c. Extended written response assessment

☐ d. Personal communication assessment

27. Which of the following grading practices will provide an accurate reflection of academic achievement?

☐ a. Including work from the entire grading period in the final grade

☐ b. Assigning zeros for missed tests and/or assignments

☐ c. Including effort in the grading

☐ d. Assigning grades using preset standards

For items 28-31, please choose the type of score that is described by the item.

	Raw score	**Percentile**	**Stanine**	**Grade equivalent**	**Competency level**
28. Which score divides percentile ranks into 9 broad categories? The range goes from 1 to 9.					
29. Which score includes a number of questions answered correctly or total number of points earned? The range goes from zero to the total possible.					
30. Which score includes the level of mastery of content? Levels are set by panels of experts.					
31. Which score includes the percent of students in a norm group that scores below any particular raw score? The range goes from 0 to 99.					

32. The same test cannot provide both norm-referenced and criterion-referenced score interpretations.

☐ a True

☐ b. False

33. Which of the following practices is important when ensuring the quality of a multiple- choice assessment?

☐ a. Use a reading level targeted at the best readers in your class

☐ b. Provide grammatical hints within the item or material presented

☐ c. Highlight words such as Most, Least, and Except

☐ d. Vary the length of the response options

34. What is the primary purpose of asking students to write practice exercises and responses?

☐ a. To use during student goal setting conferences

☐ b. To provide teachers additional test items to use on alternate versions of a test

☐ c. To teach students how to offer descriptive feedback to peers

☐ d. To provide students information about areas they are not yet mastering

35. For what purpose is an extended written response assessment most effective?
    ☐ a. To assess a large number of students
    ☐ b. To test the quality of student reasoning skills
    ☐ c. To test knowledge-level learning targets
    ☐ d. To test English language proficiency levels
36. Which assessment practice is subject to bias in the form of stereotyping?
    ☐ a. Multiple-choice questions
    ☐ b. Portfolio presentations
    ☐ c. Extended written responses
    ☐ d. Personal communication
37. Which of the following would be considered descriptive feedback?
    ☐ a. Your letters, p and f, are messy
    ☐ b. Add a conclusion to your story
    ☐ c. Your letter y looks like the letter v
    ☐ d. Your main sentence is unclear
38. All assessments that result in a grade are formative assessments.
    ☐ a. True
    ☐ b. False
39. Which practice leads to a fair, accurate reflection of academic achievement?
    ☐ a. Assigning zeros for missed assignments or tests
    ☐ b. Making final grades norm-referenced
    ☐ c. Making final grades criterion-referenced
    ☐ d. Assigning higher or lower grades based on student behavior
40. Which purpose is a report card intended to achieve?
    ☐ a. Motivating students to improve performance
    ☐ b. Communicating about academic standards
    ☐ c. Ranking students in classes or schools
    ☐ d. Communicating about student performance
41. Which of the following functions are NOT served by portfolios of student work?
    ☐ a. To improve communication about complex student learning targets
    ☐ b. To promote student learning
    ☐ c. To help students reflect on their learning
    ☐ d. To collect all student work related to a project
42. Which assessment method helps students understand the depth of their learning?
    ☐ a. Report card
    ☐ b. Multiple-choice quiz

☐ c. Oral report

☐ d. Rubric

☐ e. Portfolio

Conferences fall into five general categories according to their purposes. Please use the drop-down menu to select the appropriate conference purpose for each conference category.

Category	**Purpose**
	Please Select – *(drop down list)*:
43. Goal Setting --	
44. Intervention --	
45. Demonstration of Growth --	
46. Achievement --	
47. Feedback --	

a. Reporting strengths and weaknesses
b. Sharing information about current status
c. Observing oral reading skills
d. Sharing evidence of improvement
e. Sharing how the work of one student compares with that of another
f. Planning for improvement relative to a problem
g. Guiding for next steps in learning

48. Conferences are an effective way for students to track their progress.

☐ a. True

☐ b. False

49. A scoring guide for a performance assessment should provide:

☐ a. A checklist of important criteria

☐ b. A clear picture of what constitutes quality

☐ c. Objective judgments of student work

☐ d. A method to eliminate extraneous factors from student scores

50. A performance assessment is an assessment:

☐ a. Based on observation and judgment

☐ b. Applicable to only formative assessment

☐ c. That requires the completion of only one task

☐ d. That typically involves a simple task

51. A performance assessment should do which of the following?

☐ a. Provide students with a choice of task

☐ b. Have only one correct response

☐ c. Elicit the correct behavior from the student

☐ d. Have written instructions for a writing task

52. Which of the following is true for multiple-choice assessment items?

☐ a. They assess the production of a response

☐ b. They reduce the possibility of getting the right answer by guessing

☐ c. They cannot provide diagnostic information

☐ d. They can measure a variety of learning objectives

53. Matching questions are well suited for which of the following?

☐ a. Reducing scoring time

☐ b. When there are several plausible alternative correct answers

☐ c. Measuring association of related thoughts or facts

☐ d. Reducing the process of elimination

54. Which of the following is a potential source of bias in a multiple-choice test?

☐ a. Improper sampling of the content domain

☐ b. Assigning different weights to items

☐ c. Requiring a high reading level

☐ d. Guessing

55. Which of the following is the best example of a summative assessment?

☐ a. Report card grade

☐ b. Student self-assessment

☐ c. Portfolio

☐ d. Parent-teacher conference

56. Which of the following is a use of a formative assessment?

☐ a. Certifying student competence

☐ b. Sorting students according to achievement

☐ c. Advising students about their progress

☐ d. Forming opinions on student proficiency

57. Which strategy helps clarify instructional objectives to students?

☐ a. Showing examples of strong and weak work

☐ b. Offering regular evaluation feedback on practice work

☐ c. Explaining to students their standardized test results

☐ d. Providing clear due dates for student projects

58. Choosing the best student out of 20 to receive a citizenship award is an example of a

☐ a. Norm-referenced test

☐ b. Criterion-referenced test

59. A well-designed compare/contrast test question does not use examples covered during instruction.

☐ a. True

☐ b. False

60. Many studies have advocated for the following in order to increase motivation and achievement among students:

☐ a. Reducing both evaluative feedback and descriptive feedback

☐ b. Reducing evaluative feedback and increasing descriptive feedback

☐ c. Increasing evaluative feedback and reducing descriptive feedback

☐ d. Increasing both evaluative feedback and descriptive feedback

Thank you for taking our survey. Your response is very important to us.

Teacher assessment work sample

Directions for Collecting Assignments and Student Work

Please collect 3 assignments, with 4 graded samples of student work for each assignment. You will be asked to fill out a cover sheet for each assignment. Detailed instructions are given below.

We want to describe the nature of the math tasks that students do, what is expected of them, what feedback they are given, and how grades are assigned. Our descriptions depend on what you tell us, so please be explicit and detailed so we can be as accurate as possible. Thank you.

Adapted from Clare, L., Valdés, R., Pascal, J., & Steinberg, J.R. (2001). Teachers assignments as indicators of instructional quality in elementary schools. Los Angeles: CRESST.

INSTRUCTIONS:

1. **COLLECT THE FOLLOWING 3 ASSIGNMENTS.**
   Between now and November 19, collect 3 assignments with selected examples of graded or marked student work. These examples of student work should be papers that are ready to be returned to the students, with your marks and feedback included. Use assignments that ask students to do some individual work, and that reflect your lesson objectives. Do not create new assignments specifically for this study. Please collect one of each of the following types of assignments:

   1. 1 example of homework **or** seatwork that asks students to show their work and explain their reasoning
   2. 1 quiz **or** end-of-week assessment
   3. 1 example of a performance assessment (such as creating a graph) **or** in-class project

2. **FOR EACH OF THE 3 ASSIGNMENTS, COPY 4 SAMPLES OF STUDENT WORK showing student response to the assignment.**

   - **For each assignment, choose four samples from the same class. Choose two samples of work from students who achieved the assignment objectives, and two from students who did not achieve the assignment objectives.**
     It is fine to choose different students' papers for different assignments. If there were no students who achieved the objectives on an assignment, attach a note explaining why you are not including any of those pieces of student work. In that case, please just give us samples of work from students who did not achieve the objectives.

   - Date *and* mark the time each piece of student work was completed (e.g., 11/10/07, 1:15 pm). Copy the four sets of student work for each assignment.

   - Please cross out or white out each student's name. (We prefer to receive students' work without their names so as to protect their privacy.) Please do not cover up any part of the student's work, your feedback, or grade. It is important for us to see the feedback comments or grades.

- Place an “Achieved Objectives” label on the papers of the students who achieved the objectives. Place a “Did Not Achieve Objectives” label on the papers of the students who did not achieve the objectives.

3. FILL OUT A COVER SHEET FOR EACH OF THE 3 ASSIGNMENTS.

Fill out the enclosed Cover Sheets for Teacher Assignments in the pockets in this folder.

- **Please attach the following to help us understand** the assignment and accompanying student work, such as the following:
    - Copy of the directions given to students **(please be as explicit as possible)**
    - Grading rubric or guidelines, and
    - Outline of the unit.
- Place the cover sheet with attached papers and the 4 pieces of student work in the appropriate pockets in this binder.

Teacher report of student involvement

We are interested in the frequency of various activities that occurred in your math classroom during the period between November 3 and November 14.

How many days of regular instruction were there in this time period (excluding any teacher in-service days, field trips, etc.)? _____

On how many different days during the time between November 3 to November 14 did you have most or all of your students do the following activities? Please enter the number of different instructional days for each of the following items. For example, if you had most or all of your students discuss learning objectives on three different days during the time between November 3 to November 14, you would enter '3' for Number 1.

1. Participate in a guided discussion of the learning objectives in math. _____
2. Explain in their own words what they are supposed to be learning in math. _____
3. Identify samples of their own high quality work in math. _____
4. Use a scoring guide or rubric to evaluate their own work in math class. _____
5. Revise their own math work to make it stronger in quality. _____
6. Keep a record of their own learning progress in math. _____
7. Explain in their own words what they know how to do well in math. _____
8. Explain in their own words what they need to do to improve their math skills. _____
9. Identify examples of strong and weak anonymous student work in math. _____
10. Comment on the quality of anonymous math work using a scoring guide or rubric. _____
11. Explain in their own words what was wrong with a math answer or piece of math work. _____
12. Explain in their own words how to correct a math answer or improve a piece of math work. _____
13. Work together to correct errors in their math assignments. _____
14. Make up practice math problems. _____

Thank you for taking our survey. Your response is very important to us.

Survey of Student Motivation

Read each of the following sentences. For each one, show us how true it is for YOU by filling in the box below one of the four answers:

Not At All True, Not Very True, Sort Of True, or Very True

There are no right or wrong answers.

	4th Grade	**5th Grade**
What grade are you in?		

	Not at all true	**Not very true**	**Sort of true**	**Very true**
I work very hard on my math work.				
I do my math homework because I like to do it.				
I work on my math classwork because it's interesting.				
I'm certain I can figure out how to do the most difficult mathwork.				
I don't try very hard in math.				
I do my math homework because I want to understand the subject.				
I can do almost all the work in math if I don't give up.				
I work on my math classwork because I think it's important.				
I'm certain I can master the skills taught in math this year.				
I pay attention in math class.				
I work on my math classwork because I want to learn new things.				
I do my math homework because it's fun.				
I can do even the hardest work in math class if I try.				
I don't very hard in math.				
I do my math homework because I want to learn new things.				
Even if the math work is hard, I can learn it.				
When I am in math class I just act as if I'm working.				
I work on my math classwork because doing well in math is important to me.				
I work on my math classwork because it's fun.				

	Not at all important	**Not very important**	**Sort of important**	**Very important**
How important is it to you to do the best you can in math?				

Appendix D. Development, reliability, and validity of teacher outcomes

This study of classroom assessment used three instruments to measure intermediate teacher outcomes. Teacher knowledge of classroom assessment was measured with the Test of Assessment Knowledge. Assessment practice was measured with the Teacher Assessment Work Sample. Teacher involvement of their students in classroom assessment activities was measured with the Survey of Student Involvement. Two of these measures, the Survey of Student Involvement and the Test of Assessment Knowledge, were developed for this study. The third measure, the Teacher Assessment Work Sample, was adapted from a teacher artifact instrument developed at the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) (Matsumura, Patthey-Chavez, et al. 2002).

The Test of Assessment Knowledge was developed to sample the knowledge and reasoning skills represented in the key components of the Classroom Assessment for Student Learning (CASL) intervention's content. Some key components, such as assessment purpose, focus more on conceptual knowledge. Other key components, such as assessments that yield accurate results, emphasize skill in developing classroom assessments, rubrics, and other products to be used in the classroom. Whereas the CASL program focuses on skills and products, the Test of Assessment Knowledge measured the knowledge and reasoning that were thought to be prerequisite to the skill or products targeted by the CASL intervention.

The test samples the content covered in the CASL program, giving more weight to topics that are described in depth and that comprise a large domain of information in the CASL program. Although the test was designed to be sensitive to the CASL program, steps were taken to ensure that the test was not overaligned, which typically occurs when instruments contain materials, text, or idiosyncratic wording from the intervention. This was not the case with the Test of Assessment Knowledge, because it used common language and sampled from the general domain of classroom assessment knowledge. The test items cover generally accepted principles and practices of classroom and formative assessment and avoid terminology specific only to the CASL program.

The Teacher Assessment Work Sample was used as a measure of teacher assessment practice in the classroom. Rather than conduct observations to try to measure teacher practice, the research team used an artifact-based instrument, adapted from an artifact-based instrument developed at CRESST, as the measure of teacher assessment practice. The original instrument, developed at CRESST, measures general classroom practice in elementary and secondary language arts classrooms using language arts assignments. For this study, the instrument was adapted to measure classroom assessment practice in mathematics using mathematics assessments. Two dimensions that addressed feedback, a critical aspect of formative assessment, were added to the four dimensions of the original CRESST rubric. (See appendix E for details regarding the scoring of the Teacher Assessment Work Sample.)

The Survey of Student Involvement was used to measure the extent to which teachers involved their students in assessment and assessment-related activities. The survey included a list of

assessment-related activities, drawn from the larger literature on student involvement (such as Sadler 1989), that could occur in any classroom where students were involved in the learning and assessment process; in developing the survey of student involvement, care was taken to not include any materials, text, or wording specific to CASL. The survey used language common to educators rather than language used by the CASL developers.

All teacher outcome instruments were reviewed and pilot-tested prior to administration to the study sample. Pilot-testing with nine respondents from the target population helped ensure that the directions were clear, requests for data were unambiguous, the process was efficient, the response time was known, and the online instruments were easy to use.

Baseline data collection (November 2007) presented the first opportunity to examine data from a sample of respondents larger than the pilot test of nine individuals. The psychometric properties of the teacher outcome instruments were examined using data collected at baseline. The analyses included factor analyses, item statistics, and composite score statistics (such as item score frequency distributions, item means and standard deviations, item-total correlations, composite score frequency distributions, composite score means and standard deviations, composite score intercorrelations, and internal consistency).

Data gathered at baseline were used to examine the psychometric properties of the items and overall test functioning of the Test of Assessment Knowledge. Of the 70 items administered at baseline, 60 were selected for the final test to provide the best overall test functioning, reduce the number of items with undesirable item characteristics (that is, *p*-values below .30 or above .90 and point biserial correlations below .20), and provide an appropriate representation of the construct and sample of the domain of teacher knowledge and reasoning regarding classroom assessment practices. Responses to each item on the test were scored as either correct or incorrect.

Data from the Teacher Assessment Work Sample administered at baseline was used to provide materials for training the raters, train the raters, and examine inter-rater reliability prior to scoring the posttest Teacher Assessment Work Samples. (Details regarding the field test of the Teacher Assessment Work Sample are provided in appendix E.) Data from the Survey of Student Involvement administered at baseline showed that the survey was functioning well, the items tapped into a single underlying construct, and the survey captured variation in teacher response.

Data were analyzed to provide evidence of the reliability and validity of the teacher instruments. Table D1 provides the descriptive statistics for each teacher instrument using the posttest data, and table D2 presents correlations between the same instrument administered at different data collection waves. Each instrument was administered in a different number of data collection waves. (See chapter 2 for a description of the data collection schedule.)

Table D1. Descriptive statistics of teacher outcomes at posttest

Instrument	Number of items[a]	Maximum score[b]	Number of teachers	Mean	Standard deviation	Minimum	Maximum	Mean item-total correlation	Internal consistency
Test of Assessment Knowledge	60[a]	60[b]	291	38.53	7.54	0	56	.27	.89
Teacher Assessment Work Sample Dimensions			237						
Focus of goals on student learning	3	4	237	2.20	0.77	1	4		
Alignment of learning goals and task	3	4	237	2.12	0.75	1	4		
Alignment of learning goals and assessment criteria	3	4	237	1.49	0.72	1	4		
Clarity of the assessment criteria for students	3	4	237	1.39	0.70	1	4		
Type of feedback	3	4	237	1.31	0.57	1	4		
Feedback eliciting student involvement	3	4	237	1.21	0.52	1	4		
Survey of Student Involvement	14	100%	266	34.39%	20.28	0	100	.56	.95

a. For the Teacher Assessment Work Sample, the number of items equals the number of assessments.
b. For the Teacher Assessment Work Sample, scores were averaged across each assessment on the 1–4 rubric.
Source: Teacher outcome instruments.

Table D2. Correlations between teacher instruments by data collection wave

Instrument	Wave 3	Wave 4	Posttest
Teacher Assessment Work Sample			
Baseline			.27
			(237)
Test of Assessment Knowledge			
Baseline	.30		.28
	(264)		(194)
Wave 3			.37
			(207)
Survey of Student Involvement			
Wave 2	.56	.15	.16
	(243)	(185)	(155)
Wave 3		.16	.19
		(200)	(168)
Wave 4			.56
			(222)

Note: Numbers in parentheses are number of teachers.
Source: Teacher outcome instruments.

Table D3 shows the correlations between posttest scores on the three instruments. Table D4 shows inter-rater reliability on the Teacher Assessment Work Sample by assessment and dimension using posttest data. Table D5 presents the intercorrelations between rubric dimensions from the Teacher Assessment Work Sample using posttest data.

Table D3. Intercorrelations of teacher outcomes at posttest

Instrument	Test of Assessment Knowledge	Survey of Student Involvement
Teacher Assessment Work Sample	.28	–.02
	(236)	(187)
Test of Assessment Knowledge		–.17
		(215)

Note: Numbers in parentheses are number of teachers.
Source: Posttest data from teacher outcome instruments.

Table D4. Inter-rater reliability of Teacher Assessment Work Sample by assessment and rubric dimension at posttest

Rubric dimension	Homework	Quiz	Performance assessment
Focus of goals on student learning	.79	.83	.75
	(244)	(241)	(235)
Alignment of learning goals and task	.78	.80	.75
	(244)	(241)	(235)
Alignment of learning goals and assessment criteria	.80	.77	.75
	(243)	(240)	(234)
Clarity of the assessment criteria for students	.73	.77	.77
	(243)	(240)	(234)
Type of feedback	.81	.66	.82
	(243)	(240)	(234)
Feedback eliciting student involvement	.77	.70	.81
	(243)	(240)	(234)

Note: Numbers in parentheses are number of teachers. Inter-rater reliability is measured as Pearson product moment correlations between the two raters' scores.
Source: Teacher Assessment Work Sample.

Table D5. Intercorrelations between rubric dimensions of Teacher Assessment Work Sample at posttest

Rubric dimension	Alignment of learning goals and task	Alignment of learning goals and assessment criteria	Clarity of the assessment criteria for students	Type of feedback	Feedback eliciting student involvement
Focus of goals on student learning	.93	.42	.31	.21	.17
Alignment of learning goals and task		.45	.34	.22	.17
Alignment of learning goals and assessment criteria			.78	.29	.22
Clarity of the assessment criteria for students				.31	.24
Type of feedback					.75

Note: Number of teachers is 237.
Source: Teacher Assessment Work Sample.

Appendix E. Teacher Assessment Work Sample

The Teacher Assessment Work Sample is intended to provide an accurate measure of teacher practice of classroom assessment in elementary mathematics. This appendix describes the use of the measure in this study, the approach to using a panel to identify anchor papers, the selection and training of the scorers, and the work sample scoring process.

Instrument

As compared with classroom observations, systematically collecting samples of classroom artifacts (in this case, student papers with teacher feedback) can provide an efficient way to capture teacher practice of classroom assessment, because such samples provide a lot of information about the assignments that teachers give and how they assess them. Procedures for collecting and analyzing classroom assessments in the study of Classroom Assessment for Student Learning (CASL) were adapted from an artifact-based instrument developed to characterize classroom practice. The instrument used in this study (the Teacher Assessment Work Sample) was adapted for mathematics from one developed by researchers at the National Center for Research on Evaluation, Standards, and Student Testing (CRESST), to measure teacher practice in elementary and secondary language arts (Matsumura, Garnier, et al. 2002). CRESST's research suggested that classroom artifacts and their respective scores provide a reliable and valid characterization of classroom practice, at least as reliable as classroom observations. When scorers are sufficiently trained and use clear scoring rubrics, inter-rater agreements on scores of the artifacts have been acceptable (Borko, Stecher, Alonzo, Mocure, & McClam 2005; Clare et al. 2001; Matsumura, Garnier, et al. 2002).

In the Teacher Assessment Work Sample for the CASL study, teachers were asked to copy and send in three different types of mathematics assignments that reflected their lesson objectives, with four examples of assessed student work (including teacher feedback) for each assignment. Although teachers self-selected the samples of their assessment for submission, all instructions were identical between the intervention and control teachers, making it unlikely that the instrument could introduce any bias between intervention and control groups. This instrument, including its use of self-selected work samples, has been shown to provide a valid measure of teachers' typical classroom practice (Aschbacher 1999; Clare 2000; Clare, Valdés, Pascal, & Steinberg 2001; Matsumura, Patthey-Chavez, et al. 2002; Matsumura et al. 2008). The assignments included a typical homework or seatwork assignment, a typical in-class project or performance task, and a typical quiz or end-of-week assessment. One of the homework or in-class assessments was required to ask students to show their work and explain their answers. Instructions for the Teacher Assessment Work Sample asked teachers to attach the activity's directions and indicate the following:

- The assessment and its learning goals.
- How the assessment fit with its unit.
- How it addressed the range of student skills with the assessment.

- How much time the students needed to do the assessment.
- The type of help the students received.
- How the assessment was assessed, including the scoring rubric.
- How the students performed on the assessment.

The study team then used a rubric based on CRESST's work with classroom artifacts (Matsumura, Garnier, et al. 2002), with two feedback dimensions added to score the work samples on classroom assessment practices related to feedback. The scoring rubric included six dimensions:

- Focus of goals on student learning.
- Alignment of learning goals and assessment.
- Alignment of learning goals and grading criteria.
- Clarity of grading criteria for students.
- Type of feedback (descriptive or evaluative).
- Extent to which feedback elicits student involvement.

For each dimension, there were four levels of quality, each with a description. Because the work sample was used as a teacher outcome measure, scores from the six areas and three assessments were combined, giving each teacher a single summary score from the work sample.

Scoring panel procedure

Teacher Assessment Work Samples were first collected from teacher participants at baseline data collection to field test the instrument and allow for the training of scorers. To identify anchor and qualifying papers for the scorers, the researchers assembled a five-person panel: two assessment experts (one professor emeritus and one professor), two district-level personnel experienced in teaching and assessment, and one mathematics specialist. The panel convened for a one-day meeting in July 2008 to review the sample scoring rubric, score representative samples together as a panel, and score additional samples to be used to train and qualify the scorers. Prior to this meeting, the CASL study team had assembled a set of anchor paper candidates thought to adequately represent all dimensions and levels of the rubric, so that it would not be necessary for the panel to review the entire sample.

When reviewing the rubric, the panel recommended slight changes to its wording in order to clarify the dimensions so that the samples could be scored reliably. For example, they added "for students" to "clarity of grading criteria" in order to clarify that the criteria were not just decision rules for teachers but must be known and understood by the students. Figure E1 shows the final version of the rubric used to score the work samples. They then scored, as a group, one sample of each type of assessment (homework/seatwork, end-of-week quiz/assessment, performance task/in-class project). Next, they scored seven papers individually, with each panelist scoring each paper. The panelists decided to immediately accept scores agreed upon by at least four of five scorers and conferred among themselves to reconcile the scores for papers with lower initial

agreement. The panelists scored 14 more work samples by giving each sample three ratings (that is, a paper was considered scored when any three panelists had scored it). They immediately accepted scores that were agreed upon by two of three panelists and negotiated the others. The scoring of each work sample was recorded by a facilitator, who transcribed each panelist's initial score and the negotiated scores for each rubric dimension, for each work sample.

Figure E1. Work sample rubric: final

Dimension	Rubric score			
	4	**3**	**2**	**1**
1. Focus of the goals on student learning.	Goals are very focused on student learning. Goals are very clear and explicit in terms of what students are to learn from the assignment. Additionally, all the goals are elaborated.	Goals are mostly focused on student learning. Goals are mostly clear and explicit in terms of what students are to learn from the assignment.	Goals are somewhat focused on student learning. Goals are somewhat clear and explicit in terms of what students are to learn from the assignment. Goals may be very broadly stated. Or there may be a combination of learning goals and activities.	Goals are not focused on student learning, and are not clear and explicit in terms of what students are to learn from the assignment. Or all goals may be stated as activities with no definable objective.
2. Alignment of learning goals and task. **Match in cognitive complexity should play out in alignment of goals and criteria. Look at cover sheet and assignment.**	There is **exact** alignment between the teacher's stated learning goals for students and what the task requires students to do. The task fully supports the instructional goals. The task and goals overlap completely—neither one calls for something not included in the other.	There is **good** alignment between the teacher's stated learning goals and what the task requires students to do. The task supports the instructional goals.	There is only **some** alignment between the teacher's stated goals and what the task requires students to do. The task only somewhat supports the instructional goals, or the goal may be so broadly stated that the task and goal are aligned only at a very general level.	There is very little or no alignment between the teacher's stated goals and what the task requires students to do. The task does not support the instructional goals.
3. Alignment of learning goals and *assessment* criteria. **Pairs with alignment of goals-task Look at cover sheet (also #6) and assignment**.	Excellent quality in terms of level of cognitive challenge, clarity, and application of learning goals and *assessment* criteria. *Excellent quality of match, supports fully the intended outcomes*	There is good alignment between the teacher's stated learning goals and the stated *assessment* criteria. *Majority, but not all.*	There is only some alignment between the teacher's stated learning goals and the stated *assessment* criteria.	There is very little or no alignment between the teacher's stated learning goals and the stated *assessment* criteria.

4. Clarity of the *assessment* criteria *for students* . **If we define students out of criteria, missing the point. Emphasis on *clarity to the student.***	Teacher's *assessment* criteria are very clear, explicit, and elaborated.	Teacher's *assessment* criteria are mostly clear and explicit.	Teacher's *assessment* criteria are somewhat clear and explicit.	Teacher does not specify *assessment* criteria, or it is not possible to determine the *assessment* criteria from the teacher's documents.
5. Feedback – *Type* Pairs with feedback—student involvement. **Look at assignment, if range of feedback quality choose average/ consensus. Focus on extent to which it is descriptive.**	Majority of feedback includes descriptive information about BOTH strengths AND areas needing improvement in relation to the learning target or criteria.	Majority of feedback includes descriptive information EITHER about strengths OR areas needing improvement in relation to the learning target or criteria.	There is only some inclusion of descriptive information about strengths *OR* areas needing improvement which may or may not be clearly related to the learning target or criteria. The feedback provides descriptive information but is written illegibly or not communicated clearly.	Majority of feedback is evaluative (judgmental) or norm- or peer-referenced statements (e.g., praise without specific descriptions of the work in relation to learning target or criteria). OR it is not possible to determine the quality of feedback from information submitted.
6. Feedback – Student Involvement **Look at assignment. Focus on extent to which it relates to student's learning plan. Is there involvement feedback to all students?**	*Majority of* feedback includes questions to students eliciting their involvement in reflection and planning where to go next.	*Some of* the feedback includes guidance to students suggesting doable plans for where to go next.	Feedback includes *some* guidance to students, but the suggestions are overwhelming or paralyzing by their sheer number, or otherwise not communicated clearly.	Feedback does not include statements encouraging student involvement in assessment, revisions, extensions, or planning where to go next OR it is not possible to determine the extent to which student involvement is elicited.

Note: Bold text indicates directions or clarifying language added by the expert panel.

Outcomes of the scoring panel

The work samples scored by all five panelists showed levels of initial agreement (which required four of five identical scores) between 14 percent and 43 percent. The work samples scored by three of five panelists, requiring two of three identical scores for initial agreement (a less stringent criterion), showed initial agreement ranging between 71 percent and 93 percent (table E1). Overall, the mean of the scores was 2.15 (standard deviation = .61), indicating a fairly low level of ratings on the four-point rubric in this baseline sample. The lowest scoring dimensions were the two feedback dimensions, with means of 1.67 and 1.62.

Table E1. Initial agreement with five and three scorers (percent)

Dimension	Five scorers, seven samples	Three scorers, fourteen samples
Focus of the goals on student learning	29	86
Alignment of learning goals and task	14	93
Alignment of learning goals and assessment criteria	29	71
Clarity of the assessment criteria for students	14	79
Feedback–type	43	93
Feedback–student involvement	43	86

Source: Work sample scoring panel data.

Three work samples were labeled by panelists as "challenge samples" because of qualities that led to low initial agreement across the dimensions. Those work samples were used in training as illustrations of samples whose relationships to the rubric dimension scoring levels were relatively ambiguous.

After the scoring panel completed its work, work samples had been identified for most dimensions and scores of the rubric. However, samples with a score of 4 were identified only for alignment of learning goals and task and alignment of learning goals and assessment criteria, and samples with a score of 1 were found for alignment of learning goals and task. The panelists had particular difficulty finding work samples that described the learning goals in language that was clear, explicit, and elaborated and finding samples that gave high-quality feedback, especially feedback that elicited student involvement.

Recruiting scorers

The CASL study team recruited scorers who had experience and backgrounds in both classroom teaching and education evaluation, according to the recommendation of the original developers of the work sample instrument (Matsumura, Garnier, et al. 2002). The study team's goal was to recruit two scorers. Four district-level instructional coaches from Denver-area school districts responded to the request for scorers.

At the first meeting with the four instructional coaches, the study team described the study and introduced the teacher work sample instrument. The nature of the scoring task (number of scorers needed, nature of the work, timeline, and logistics) was then explained. Following that discussion, the candidate scorers discussed the rubric and then reviewed some work samples that had been scored by the expert panel. Finally, candidate scorers scored other sample assessments and then compared their scores with the panel's scores.

Scorer training and qualifying

The first scorer training meeting started with an in-depth discussion of the work sample instrument and the dimensions and levels of the rubric. Scorers were asked to describe what samples of the various levels of quality would look like for each dimension. Next, the scorers reviewed and discussed more anchor work samples, comparing their qualities with those described on the scoring rubric. The scorers then scored a set of anchor work samples

independently and discussed each sample as a group, covering how their scores compared with those of the expert panel, what the expert panel focused on when giving ratings, and why any discrepancies might have occurred between their scores and the panel's scores. The second scorer training meeting focused on scoring more anchor work samples. Each scorer evaluated six of the anchor work samples and again discussed how the expert panel reached their decisions and how their own scoring compared.

Originally, two scorers were to be randomly selected after the second scoring meeting. However, two of the scorers decided to opt out so that their colleagues could participate and to avoid the appearance of competition.

To maintain reliability and avoid any potential scorer effects, raters were required to qualify prior to scoring work samples from the research study sample. The two qualifying meetings took place with the two remaining scorer candidates. Scorers were expected to achieve 80 percent exact agreement with the scores of the qualifying work samples to qualify. With the first set of samples, the scorers achieved an average of 50 percent exact agreement with the final anchor work sample scores from the expert panel (an average of 93.5 percent when including exact and adjacent agreement). In the second qualifying meeting, after additional discussion, the scorers were able to achieve 80 percent agreement with the second set of anchor work samples. Over the entire set of anchor work samples, the scorers achieved 58 percent exact agreement with each other (an average of 97 percent, when including exact and adjacent agreement).

Field test

The instrument was field tested at the first administration of the Teacher Assessment Work Sample at baseline (November 2007). After the work samples were collected and stamped on each page with the teacher identification number, they were put into folders identified by teacher number and the identifying cover sheets were removed. As the scorers evaluated each teacher's work sample, they entered the ratings into an online survey software system that created a database of the scores as they worked. Although the goal of the scoring was to assign a single score to each teacher, the two scorers scored each type of assessment (quiz, homework, performance) separately, scoring each one on the six dimensions of the rubric. Therefore, each teacher's work sample had 18 scores per scorer (36 scores in total). In the field test the scorers achieved 65.0 percent exact agreement at the level of the 18 individual scores per teacher (97.0 percent exact and adjacent agreement) on 153 complete work samples. The lowest level of exact agreement (50.3 percent) was on the alignment of learning goals and task on the performance assessment, while the highest level of agreement (85.4 percent) was on the student involvement feedback dimension on the quiz. This level of agreement was consistent with that found in past uses of the work sample and rubric with reading and writing; no precedent existed for mathematics. The original developers of the work sample (Clare et al. 2001; Matsumura, Garnier, et al. 2002) reported 62 percent exact agreement between three raters in reading (with a range of 48 percent to 69 percent) and 58 percent in writing (with a range of 41 percent to 76 percent). However, Clare et al. (2001) had reported 82–92 percent initial exact agreement with two scorers. Therefore, a retraining session was conducted before the scoring of the posttest work samples in the hope that a higher degree of initial agreement could be reached between the two scorers for the CASL study.

Retraining

The retraining session focused on how to distinguish adjacent scores within the rubric dimensions (between 2 and 3 on clarity of assessment criteria, for example) and on exploring and resolving work samples with discrepant scores (such as a 1 from one scorer and a 4 from the other). In all, 32 work samples from the field test were selected for discussion and retraining. In discussing how they had originally scored the work samples, the scorers realized that there were several common types of ambiguous work samples that seemed to be causing some of the discrepant scores on the rubric dimensions. Consequently, the scorers created some additional decision rules for each of the six rubric dimensions, to use when scoring the final spring 2009 work samples. Those decision rules appear in box E1.

Box E1. Decision rules

A completely elaborated goal (rubric dimension 1) is one that explains what students are doing and why. If the goal is just an activity or verb, such as "multiply," that would score a 1. If it says what to do, such as "multiply two-digit numbers" that should score a 2. If it mentions a method also, as in "multiply two-digit numbers using the XYZ method" that would be a 3. To score a 4, a goal would have to include why, as in "multiply two-digit numbers using the XYZ method in order to…" When teachers send in a summative test (such as a midterm) instead of a short quiz, and the goal is a list of the topics on the test, it is acceptable in that case.

Dimension 2 is dependent on dimension 1. If dimension 1 scored a 2, for example, then dimension 2 cannot score higher, but it can of course score lower.

As with dimension 2, if the assessment criteria are vague, the score for dimension 3 cannot really be higher than that for dimensions 1 and 2. A good rubric that is not shared with students can score well here, but it won't score well in dimension 4 (clarity of assessment criteria to students).

Dimension 4 contains two concepts—clarity to the teacher and clarity to students. It is important here to only score what is included in the work sample packet, instead of making inferences about what may have happened in the classroom, but on the other hand, to use all available evidence to decide on the score. When the teacher says they went over the assignment in class, the sample scores a 1 if that information is really vague. If there are any actual criteria shared with students, the score is at least a 2. If the only evidence of sharing criteria with students is the check box in the work sample instructions, that is a 1. When teachers indicate that they used correct/not correct or percentage correct, that scores a 2. If there is a nice rubric but no evidence whatsoever that it was shared with the students, that scores a 1, because not sharing a rubric with students is inconsistent with good classroom assessment practice (the rubric can lead to a higher score on Dimension 3 or the feedback dimensions, however). If the rubric was shared with students but is in "teacher language," that scores a 2. If the information on this dimension is limited to a general grading policy for the class, that scores a 1.

Dimension 5, the first feedback dimension, was pretty clear. In order to score a 4, a paper must have feedback about strengths AND areas needing improvement.

For Dimension 6, if nothing is written on the paper, but the teacher indicated that they went over the papers in class and students made corrections, the sample should score a 2.

Final work sample scoring

The scorers began scoring the posttest work samples in summer 2009, after all samples had been received. The work samples had been identified by teacher number only, as in the field test administration, so scorers had no information about teachers, the schools, or experimental group. Also as in the field test, the scorers entered rubric scores for each assessment type (homework, quiz, performance assessment) for each dimension of the rubric. Each work sample ($N = 227$) was scored by both scorers. Following the end of the scoring period, the study team met with the scorers to resolve data entry errors.

At the level of the 18 scores each teacher received, the mean level of exact agreement between scorers was 82.76 percent, which was within the range reported by the original work sample developers (Clare et al 2001). The agreement on individual scores ranged from 73.09 percent on

dimension 3 (alignment of learning goals and assessment criteria, with the performance assessment) to 90.00 percent on dimension 6 (feedback eliciting student involvement, with the quiz). Exact plus adjacent agreement averaged 98.93 percent and ranged from 96.58 percent to 100 percent.

Nevertheless, there were some discrepant scores on individual assessments. Of the entire set of individual scores (two scorers and 18 scores for 227 work samples), 43 pairs of scores were not in at least adjacent agreement. In most instances (n = 38) one rater scored the sample as a 3 while the other scored it as a 1; the rest were 4/2. There were no instances of one scorer giving a 4 while the other gave a 1. The study team met with the scorers to resolve the discrepant scores. The scorers determined that some were data entry errors, while others were due to differing initial reactions to ambiguity in some work sample materials submitted. At the meeting the scorers agreed on a single score for all of the discrepant cases, and the dataset was updated.

Teacher summary scores

The individual teacher work sample summary scores were then calculated from the individual scores. To create the summary scores, the researchers first averaged the two raters' scores across the six rubric dimensions and then averaged the dimension scores across the three assessments.

Appendix F. Impact analysis models

This appendix describes the models used to estimate Classroom Assessment for Student Learning (CASL) impacts on intermediate outcomes: the student achievement benchmark model, the student achievement no pretest covariate model, the student motivation model, and the teacher outcome model.

Student achievement benchmark model

Level 1:

$$Y_{ij} = \beta_{0j} + \beta_{1j}(\text{GRADE}_{ij} - \overline{\text{GRADE}}_{..})_{ij} + \beta_{2j}(\text{CSAP}_{ij} - \overline{\text{CSAP}}_{.j})_{ij} + e_{ij}$$

where Y_{ij} is achievement outcome (mathematics scale score on the spring 2009 Colorado Student Assessment Program for student i in school j); β_{0j} is the adjusted average achievement outcome for students in school j; β_{1j} is the adjusted difference in achievement outcome due to a student's grade, where *GRADE* was coded as 0 for students in grade 4 and +1 for students in grade 5 and was grand-mean-centered; β_{2j} is the regression slope of the cluster-mean-centered student pretest in school j for students in school j whose value on the outcome variable is the school average; and e_{ij} is the random error in the achievement outcome associated with student i in school j.

Level 2:

$$\beta_{0j} = \gamma_{01}(\text{CASL})_j + \gamma_{02}(\overline{\text{CSAP}}_{.j} - \overline{\text{CSAP}}_{..})_j + \gamma_{03}(\text{BLOCK 1}) + \gamma_{04}(\text{BLOCK 2}) + \gamma_{05}(\text{BLOCK 3}) + \gamma_{06}(\text{BLOCK 4}) + \gamma_{07}(\text{BLOCK 5}) + \gamma_{08}(\text{BLOCK 6}) + \gamma_{09}(\text{BLOCK 7}) + \gamma_{010}(\text{BLOCK 8}) + \gamma_{011}(\text{BLOCK 9}) + u_{0j}$$

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

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

where γ_{01} is the adjusted mean difference in the achievement outcome between schools assigned to the intervention group and schools assigned to the control group, *CASL* is an indicator variable for the intervention coded as 1 for schools randomly assigned to the intervention and 0 for schools randomly assigned to the control group, γ_{02} is the regression slope of the school level pretest (grand-mean-centered), γ_{03} through γ_{011} are the additive effects of each block used in the random assignment of schools, u_{0j} is the random error associated with school j's average on the achievement outcome, γ_{10} is the average regression slope for student grade and is fixed across schools, and γ_{20} is the average regression slope of the student level pretest and is fixed across schools.

Mixed model:

$$Y_{ij} = \gamma_{01}(\text{CASL})_j + \gamma_{02}(\overline{\text{CSAP}}_{.j} - \overline{\text{CSAP}}_{..})_j + \gamma_{03}(\text{Block 1}) + \gamma_{04}(\text{Block 2}) + \gamma_{05}(\text{Block 3}) + \gamma_{06}(\text{Block 4}) + \gamma_{07}(\text{Block 5}) + \gamma_{08}(\text{Block 6}) + \gamma_{09}(\text{Block 7}) + \gamma_{010}(\text{Block 8}) + \gamma_{011}(\text{Block 9}) + \gamma_{10}(\text{GRADE}_{ij} - \overline{\text{GRADE}}_{..})_{ij} + \gamma_{20}(\text{CSAP}_{ij} - \overline{\text{CSAP}}_{.j})_{ij} + e_{ij} + u_{0j}$$

where γ_{01} is the adjusted school mean difference between intervention and control group.

This student achievement benchmark model was also used in the three sensitivity analyses: maximum likelihood estimation method, minimum variance quadratic unbiased estimation method, and case deletion treatment of missing data.

Student achievement no pretest covariate model

Level 1:

$$Y_{ij} = \beta_{0j} + \beta_{1j}(\text{GRADE}_{ij} - \overline{\text{GRADE}}_{..})_{ij} + e_{ij}$$

Level 2:

$$\beta_{0j} = \gamma_{01}(\text{CASL})_j + \gamma_{02}(\text{Block 1}) + \gamma_{03}(\text{Block 2}) + \gamma_{04}(\text{Block 3}) + \gamma_{05}(\text{Block 4}) + \gamma_{06}(\text{Block 5}) + \gamma_{07}(\text{Block 6}) + \gamma_{08}(\text{Block 7}) + \gamma_{09}(\text{Block 8}) + \gamma_{010}(\text{Block 9}) + u_{0j}$$

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

Mixed model:

$$Y_{ij} = \gamma_{01}(\text{CASL})_j + \gamma_{02}(\text{Block 1}) + \gamma_{03}(\text{Block 2}) + \gamma_{04}(\text{Block 3}) + \gamma_{05}(\text{Block 4}) + \gamma_{06}(\text{Block 5}) + \gamma_{07}(\text{Block 6}) + \gamma_{08}(\text{Block 7}) + \gamma_{09}(\text{Block 8}) + \gamma_{010}(\text{Block 9}) + \gamma_{10}(\text{GRADE}_{ij} - \overline{\text{GRADE}}_{..})_{ij} + e_{ij} + u_{0j}$$

The parameters for the student achievement no pretest covariate model are the same as those in the student achievement benchmark model with the only difference being that the student (level 1) pretest achievement covariate and cluster (level 2) pretest achievement covariate are not included.

Student motivation model

Level 1:

$$Y_{ij} = \beta_{0j} + \beta_{1j}(\text{GRADE}_{ij} - \overline{\text{GRADE}}_{..})_{ij} + e_{ij}.$$

where Y_{ij} is the student motivation outcome, β_{0j} is the average motivation score for students in school j, β_{1j} is the adjusted difference in student motivation due to a student's grade where *GRADE* is grand-mean-centered, and e_{ij} is the random error in the student motivation outcome associated with student i in school j.

Level 2:

$$\beta_{0j} = \gamma_{01}(\text{CASL})_j + \gamma_{02}(\text{Block 1}) + \gamma_{03}(\text{Block 2}) + \gamma_{04}(\text{Block 3}) + \gamma_{05}(\text{Block 4}) + \gamma_{06}(\text{Block 5}) + \gamma_{07}(\text{Block 6}) + \gamma_{08}(\text{Block 7}) + \gamma_{09}(\text{Block 8}) + \gamma_{010}(\text{Block 9}) + u_{0j}$$

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

where γ_{01} is the adjusted mean difference in student motivation between schools assigned to the intervention group and schools assigned to the control group, *CASL* is an indicator variable for the intervention coded as 1 for schools randomly assigned to the intervention and 0 for schools randomly assigned to the control group, γ_{02} through γ_{010} are the additive effects of each block used in the random assignment of schools, u_{0j} is the random error associated with school j's average on the student motivation outcome, and γ_{10} is the average regression slope for student grade fixed across schools.

Mixed model:

$$Y_{ij} = \gamma_{01}(\text{CASL})_j + \gamma_{02}(\text{Block 1}) + \gamma_{03}(\text{Block 2}) + \gamma_{04}(\text{Block 3}) + \gamma_{05}(\text{Block 4}) + \gamma_{06}(\text{Block 5}) + \gamma_{07}(\text{Block 6}) + \gamma_{08}(\text{Block 7}) + \gamma_{09}(\text{Block 8}) + \gamma_{010}(\text{Block 9}) + \gamma_{10}(\text{GRADE}_{ij} - \overline{\text{GRADE}}_{..})_{ij} + u_{0j} + e_{ij}$$

where γ_{01} is the impact estimate of CASL on student motivation.

Teacher outcomes model

The teacher outcome model was used in each of the teacher outcome impact analyses. Only the outcome variable changed.

Level 1:

$$Y_{ij} = \beta_{0j} + \beta_{1j}(\text{ENTRY}_{ij} - \overline{\text{ENTRY}}_{..})_{ij} + \beta_{2j}(\text{ToAK}_i - \overline{\text{ToAK}}_{.j})_{ij} + \beta_{3j}(\text{YearsTeach}_{ij} - \overline{\text{YearsTeach}}_{..})_{ij} + \beta_{4j}(\text{YearsMath}_{ij} - \overline{\text{YearsMath}}_{..})_{ij} + e_{ij}.$$

where Y_{ij} is the teacher outcome (summary score on the test of assessment knowledge, teacher work sample, or survey of student involvement for teacher i in school j); β_{0j} is the average teacher outcome score for teachers in school j; β_{1j} is the adjusted difference in achievement outcome due to teachers' entry into the study (original or late entry) where *ENTRY* is grand-mean-centered to control for the proportion of original and late entry teachers across the schools;

β_{2j} is the regression slope of the cluster-mean-centered teacher assessment knowledge pretest score in school *j* for teachers in school *j* whose value on the outcome variable is the school average; β_{3j} is the regression slope of the covariate for years of teaching experience as indicated on the teacher background survey where *YearsTeach* is grand-mean-centered; β_{4j} is the regression slope of the covariate for the number of years teaching math as indicated on the teacher background survey where *YearsMath* is grand-mean-centered; and e_{ij} is the random error in the teacher outcome associated with teacher *i* in school *j*.

Level 2:

$$\beta_{0j} = \gamma_{01}(\text{CASL})_j + \gamma_{02}(\overline{\text{ToAK}}_{.j} - \overline{\text{ToAK}}_{..})_j + \gamma_{03}(\text{Block 1}) + \gamma_{04}(\text{Block 2}) + \gamma_{05}(\text{Block 3}) + \gamma_{06}(\text{Block 4}) + \gamma_{07}(\text{Block 5}) + \gamma_{08}(\text{Block 6}) + \gamma_{09}(\text{Block 7}) + \gamma_{010}(\text{Block 8}) + \gamma_{011}(\text{Block 9}) + u_{0j}$$

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

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

$$\beta_{3j} = \gamma_{30}$$

$$\beta_{4j} = \gamma_{40}$$

where γ_{01} is the adjusted mean difference in the teacher outcome between schools assigned to the intervention group and schools assigned to the control group, *CASL* is an indicator variable for the intervention coded as 1 for schools randomly assigned to the intervention and 0 for schools randomly assigned to the control group, γ_{02} is the regression slope of the school level teacher assessment knowledge pretest (grand-mean-centered), γ_{03} through γ_{011} are the additive effects of each block used in the random assignment of schools, γ_{10} is the average regression slope for teacher entry (original or late entry) fixed across schools, γ_{20} is the average regression slope of the teacher assessment knowledge pretest fixed across schools, γ_{30} is the average regression slope of the covariate for number of years teaching experience (*YearsTeach*) fixed across schools, γ_{40} is the average regression slope of the covariate for number of years teaching math (*YearsMath*) fixed across schools, and u_{0j} is the random error associated with school *j*'s average on the teacher outcome.

Mixed model:

$$Y_{ij} = \gamma_{01}(\text{CASL})_j + \gamma_{02}(\overline{\text{ToAK}}_{.j} - \overline{\text{Mean ToAK}}_{..})_j + \gamma_{03}(\text{Block 1}) + \gamma_{04}(\text{Block 2}) + \gamma_{05}(\text{Block 3}) + \gamma_{06}(\text{Block 4}) + \gamma_{07}(\text{Block 5}) + \gamma_{08}(\text{Block 6}) + \gamma_{09}(\text{Block 7}) + \gamma_{010}(\text{Block 8}) + \gamma_{011}(\text{Block 9}) + \gamma_{10}(\text{ENTRY}_{ij} - \overline{\text{ENTRY}}_{..})_{ij} + \gamma_{20}(\text{ToAK}_{ij} - \overline{\text{ToAK}}_{.j})_{ij} + \gamma_{30}(\text{YearsTeach}_{ij} - \overline{\text{YearsTeach}}_{..})_{ij} + \gamma_{40}(\text{YearsMath}_{ij} - \overline{\text{YearsMath}}_{..})_{ij} + e_{ij} + u_{0j}$$

where γ_{01} is the impact estimate of CASL on the whole school teacher outcome.

Appendix G. Calculation of effect sizes

Glass's *d* (Glass, McGaw, & Smith 1981) was calculated to present the treatment group–control group contrast in standard deviation units of the control group, the counterfactual for this study. Glass's *d* was chosen because the intervention may have impacted the standard deviation of the treatment group such that the pooled standard deviation (as used in Hedges's *g*; Hedges & Olkin 1985) might not represent the population to whom the results would be most interesting: schools that are considering using Classroom Assessment for Student Learning (CASL) to increase student achievement. Expressing the impact of CASL in standard deviation units of the control group also helps provide a policy-relevant answer to the question, "What impact can be anticipated if a school implements CASL?" The numerator for each effect size was the difference between the adjusted treatment group mean and the adjusted control group mean estimated in the multilevel impact analysis model. The denominator was the control group standard deviation, calculated as the square root of the sum of the control group level 2 variance component (τ_{00}) plus the control group level 1 variance component (σ^2) times the control group n–1 divided by the control group n where the variance components were estimated in a null model (Hedges 2007, 2009).

Appendix H. Treatment of missing data

This appendix describes the treatment of missing data and the creation of the impact analysis samples. Missing data were imputed with multiple imputations using the maximum likelihood expectation maximization algorithm with SAS PROC MI. The maximum likelihood expectation maximization algorithm with multiple imputations was used to impute missing scale score data for the student achievement data, missing item data for the student motivation data, and missing summary score data for teacher outcomes. Forty imputed datasets were constructed for the student achievement data, 40 imputed datasets for the student motivation data, and 40 for the teacher outcome data. Graham, Olchowski, and Gilreath (2007) found that a minimum of 40 imputed datasets were needed to prevent a falloff of statistical power. Impact analyses were conducted with SAS PROC MIXED on each of the imputed datasets, and impact estimates across the imputed datasets were combined using PROC MIANALYZE.

The maximum likelihood expectation maximization algorithm with multiple imputations was used for several reasons. First, this method has been recommended for situations in which both pretest and posttest data are missing (Puma et al. 2009), which was the case for this study. Second, the literature suggests that this method yields standard errors with little or no bias and produces unbiased impact estimates when imputing missing pretest and post test data. According to Puma et al. (2009), this method produced biased impact estimates only when 40 percent of student posttest data were missing, a situation that does not characterize the Classroom Assessment for Student Learning (CASL) study data. The expectation maximization method also was chosen because this method allows for imputing all variables in the imputer's model at the same time. Given the many variations in nonresponse patterns, the expectation maximization method was most efficient for imputing missing data across the various patterns of nonresponse. Finally, this method has been found to be robust to the varying amounts of missing data found in the CASL study (Puma et al. 2009).

Student achievement impact sample

The student achievement impact sample included data obtained from the state education department for all schools randomly assigned at the beginning of the study. Student achievement data were missing for some students. Table H1 shows the student mathematics achievement data available and imputed by experimental group and grade.

Table H1. Available and imputed student achievement data

	Intervention (Schools = 33)		Control (Schools = 34)	
Data	Grade 4	Grade 5	Grade 4	Grade 5
Pre- and posttest math score available	1,548	1,312	1,826	1,553
Pretest math score imputed	327	461	297	617
Posttest math score imputed	305	423	330	506
Pre- and posttest math scores imputed	21	23	24	23
Students in final impact sample	2201	2219	2477	2699

Source: 2007, 2008, 2009 Colorado Student Assessment Program data.

Missing mathematics scale scores were imputed using the expectation maximization algorithm separately for intervention and control groups. The imputations were also conducted separately by grade level (cohort) because grade was hypothesized as a potential moderator variable where the intervention may have had a differential impact for students in Cohort 1 as compared with students in Cohort 3 (see chapter 2). When imputing under a model that includes a possible moderator composed of a grouping variable such as grade, Graham (2009) recommends that imputation be conducted separately by the groups represented by the grouping variable. Thus, data were imputed separately for the Cohort 1 intervention students, Cohort 3 intervention students, Cohort 1 control students, and Cohort 3 control students. The imputer's model included the following variables: pretest Colorado Student Assessment Program (CSAP) scale scores on mathematics, reading, and writing; posttest CSAP scale scores on mathematics, reading, and writing; gender; free or reduced-price lunch status; and race/ethnicity.

Student motivation impact sample

The student motivation data consisted of students' responses to a 20-item survey. The survey was administered once at Wave 3 (May 2008) and once at posttest (May 2009). The survey was completely anonymous; no student identifiers or student demographic data other than grade were available. Student data from the two administrations could not be linked. Information was available on students' school and intervention condition. Student responses to the survey items were coded as 1 = not at all true, 2 = not very true, 3 = sort of true, or 4 = very true. For the impact analysis a student motivation composite score was computed as the mean of the 20 items. Approximately 2.8 percent of the total item data were missing for Wave 3 and 1.2 percent of the total for posttest.

The expectation maximization algorithm was used to impute the missing item data on the survey. The imputations were conducted separately for 2008 and 2009 data, by intervention group and grade level. All 20 items from the survey were included in the imputer's model because no demographic data are available other than student grade. Item response data were imputed with no restriction to range and no rounding (the typically recommended method). This approach resulted in convergence of the imputations. The algorithm did not converge when the range was restricted with no rounding.

Teacher outcomes impact samples

Teacher data included data from each outcome administered at multiple waves (see chapter 2 for a description of the data collection schedule). The teacher knowledge outcome had baseline, Wave 3, and posttest data. The teacher assessment practice outcome had baseline and posttest data. The student involvement in assessment outcome had four waves: Waves 2–4 and posttest. The teacher data also included teacher background information.

Missing item data were treated as part of the creation of summary scores for each teacher outcome. For the Test of Assessment Knowledge, items missing a response were treated as incorrect, and if a teacher omitted more than 10 percent of the items, the summary score treated was missing. The Teacher Assessment Work Sample was based on ratings of three teacher assessments. Scores on each assessment were the average score across the six dimensions, where

the dimension score was the average score from the two raters. Thus, teachers had a summary score for each assessment submitted. For the Survey of Student Involvement, more than 96 percent of the teachers completing the survey responded to all 14 items. When only one item was missing, the mean response to the other 13 items was used as the summary score. When more than one item response was missing, the summary score was treated as missing.

Missing teacher summary score data were handled using the expectation maximization algorithm and multiple imputations. Data for all missing teacher summary scores were imputed with a single imputer's model that resulted in 40 imputed datasets. Imputations were run separately for the intervention and control teachers. Although teachers' entry time into the study (either original sample or late entry) was hypothesized as a possible moderator variable and included in the impact models as such, imputations run separately by entry failed to converge, likely because the sample size was too small. Imputed teacher data were used only in the estimation of CASL impacts on the teacher outcomes presented in table 5.2 of Chapter 5. Throughout this report, teacher sample sizes presented are for observed data with the exceptions of figure 2.2 which explicitly describes sample sizes for observed data and imputed data and table H2 described below.

The imputer's model included the summary scores from all waves of the teacher outcomes. Summary scores from the intermediate waves provided predictive information for imputing missing data posttest data. The imputer's model also included eight variables from the teacher background survey: years teaching experience, years teaching at current grade level, years working at current school, years experience teaching math, years experience teaching math at current grade level, number of students enrolled in class, percentage of students in class since beginning of school year, and similarity of instruction and curriculum to other teachers at the same school and grade level. Table H2 shows the numbers of teachers whose data were imputed either because of missing item data or instrument nonresponse.

Table H2. Number of teachers for whom data were imputed

Teacher data imputed	Intervention teachers	Control teachers
Baseline Test of Assessment Knowledge		
Missing more than 10 percent of items or no response	44	76
Posttest Test of Assessment Knowledge		
Missing more than 10 percent of items or no response	64	62
Posttest Teacher Assessment Work Sample		
No response	84	75
Posttest Survey of Student Involvement		
Missing more than one item	10	18
No response	82	77

Source: Teacher outcome instruments.

Appendix I. Variance components estimates and intraclass correlations

The variance from the two-level null model, a model with no covariates, can be partitioned into two random effects: σ^2, which is the variance of the residual or individual-level component from level 1 (e_{ij}), and τ_{00}, which is the variance of the intercept or school-level residual component from level 2 (u_{0j}).

Variance components were estimated from a model with no covariates, the null model, to estimate the proportion of variance in the outcome that was between schools. These estimates were produced to confirm that multilevel modeling was warranted and to check the assumption regarding clustering made in the power analyses. The intraclass correlation, the measure of the proportion of total variance that is between schools, is calculated by dividing the residual-level variance by the total variance, where the total variance is the residual variance plus the school-level variance.

Results from the null model fit to the student achievement data yielded $\sigma^2 = 1{,}003.23$ and $\tau_{00} = 5{,}199.46$ (table I1). The estimated intraclass correlation (ρ) between any two students in the same school, therefore, was .16, larger than what had been assumed for the power analyses of .15.

Table I1. Variance components and intraclass correlation for student achievement

Outcome	Estimate
Level 1 (student) variance	5,199.46
Level 2 (school) variance	1,003.23
Total variance	6,202.69
Intraclass correlation	.16

Source: 2009 Colorado Student Assessment Program data.

Null models also were fit to the student motivation data and the teacher outcome data. The intraclass correlation for student motivation was .03 for both the 2008 and the 2009 data (table I2). Intraclass correlations for the three teacher outcomes were .12 for the Test of Assessment Knowledge, .06 for the Teacher Assessment Work Sample, and .11 for the Teacher Report of Student Involvement.

Table I2. Variance components and intraclass correlations for student motivation and teacher outcomes

Outcome	Estimate
Student motivation Wave 3	
Level 1 (Student)	0.22
Level 2 (School)	0.01
Total	0.23
Intraclass correlation	.03
Student motivation posttest	
Level 1 (Student)	0.22
Level 2 (School)	0.01
Total	0.23
Intraclass correlation	.03
Test of Assessment Knowledge	
Level 1 (Teacher)	53.96
Level 2 (School)	7.29
Total	61.25
Intraclass correlation	.12
Teacher Assessment Work Sample	
Level 1 (Teacher)	0.15
Level 2 (School)	1.61
Total	1.75
Intraclass correlation	.92
Teacher Report of Student Involvement	
Level 1 (Teacher)	0.04
Level 2 (School)	0.01
Total	0.04
Intraclass correlation	0.11

Source: Student motivation data and teacher outcome data.

Appendix J. Raw means and standard deviations

This appendix provides raw means and raw standard deviations for each outcome included in the impact analysis. Raw means are means that are not adjusted for covariates or for clustering of students (or teachers) within schools. The raw means presented in the tables below are derived from the 40 imputed datasets used in each of the respective impact analyses.

Table J1. Raw means and standard deviations for student mathematics achievement

Measure	Intervention group Schools = 33 Students = 4,420	Control group Schools = 34 Students = 5,176
Pretest score		
Mean	453.02	458.91
Standard deviation	88.77	88.29
Posttest score		
Mean	500.09	505.54
Standard deviation	78.21	80.77

Source: 2007, 2008, and 2009 Colorado Student Assessment Program data.

Table J2. Raw means and standard deviations for student motivation

Measure	Intervention group	Control group
Wave 3 unadjusted score		
Number of schools	28	27
Number of students	1,179	2,579
Mean	3.28	3.27
Standard deviation	0.47	0.48
Posttest unadjusted score		
Number of schools	24	32
Number of students	2,016	3,170
Mean	3.33	3.32
Standard deviation	0.47	0.48

Source: Student Survey of Motivation.

Table J3. Raw means and standard deviations for teacher outcomes

Measure	Intervention group Schools = 33 Teachers = 178	Control group Schools = 34 Teachers = 231
Pretest score		
Mean	36.55	35.86
Standard deviation	6.05	6.28
Test of Assessment Knowledge Posttest		
Mean	41.81	38.27
Standard deviation	8.72	6.55
Teacher Assessment Work Sample		
Mean	1.62	1.60
Standard deviation	0.41	0.38
Teacher Report of Student Involvement		
Mean	0.38	0.35
Standard deviation	0.21	0.20

Source: Teacher outcome data.

Appendix K. Complete mixed model results

Table K1. Mixed model results for student achievement baseline comparison

Parameter	Estimate	Standard error	Test statistic	*p*-value
Intervention	–7.16	7.36	–0.97	.33
Grade	11.58	1.71	6.76	<.01
Block 1	453.49	11.23	40.39	<.01
Block 2	482.31	11.55	41.78	<.01
Block 3	392.98	15.23	25.8	<.01
Block 4	434.63	15.23	28.53	<.01
Block 5	479.17	21.41	22.38	<.01
Block 6	459.10	9.40	48.82	<.01
Block 7	455.48	9.48	48.07	<.01
Block 8	477.56	9.05	52.76	<.01
Block 9	466.32	15.41	30.26	<.01
Variance components				
Level 1 (student)	6,488.33	103.84	62.49	<.01
Level 2 (school)	835.83	168.96	4.95	<.01

Source: 2007, 2008, and 2009 Colorado Student Assessment Program data.

Table K2. Mixed model results for student achievement impact analysis no-covariate model

Parameter	Estimate	Standard error	Test statistic	*p*-value
Intervention	–4.93	6.66	–0.74	.46
Grade	–20.12	1.55	–13.01	<.01
Block 1	500.77	10.16	49.31	<.01
Block 2	533.81	10.45	51.08	<.01
Block 3	449.75	13.79	32.62	<.01
Block 4	483.75	13.79	35.09	<.01
Block 5	540.90	19.36	27.93	<.01
Block 6	498.00	8.50	58.57	<.01
Block 7	498.66	8.57	58.2	<.01
Block 8	517.10	8.16	63.38	<.01
Block 9	493.82	13.93	35.45	<.01
Variance components				
Level 1 (student)	5,099.39	80.24	63.55	<.01
Level 2 (school)	685.92	136.78	5.01	<.01

Source: 2007, 2008, and 2009 Colorado Student Assessment Program data.

Table K3. Mixed model results for benchmark impact analysis on student achievement

Parameter	Estimate	Standard error	Test statistic	*p*-value
Intervention	0.58	3.47	0.17	.87
Grade	–28.23	1.08	–26.05	<.01
Student pretest	0.70	0.01	104.27	<.01
Cluster pretest	0.77	0.06	12.52	<.01
Block 1	502.88	5.22	96.4	<.01
Block 2	513.60	5.54	92.75	<.01
Block 3	498.72	8.04	62.06	<.01
Block 4	500.49	7.14	70.08	<.01
Block 5	523.21	10.00	52.34	<.01
Block 6	495.80	4.39	112.91	<.01
Block 7	499.28	4.45	112.28	<.01
Block 8	500.86	4.45	112.49	<.01
Block 9	486.07	7.21	67.46	<.01
Variance components				
Level 1 (student)	1,920.93	37.71	50.94	<.01
Level 2 (school)	172.18	38.01	4.53	<.01

Source: 2007, 2008, and 2009 Colorado Student Assessment Program data.

Table K4. Mixed model results for student achievement impact analysis Cohort 1 subtest

Parameter	Estimate	Standard error	Test statistic	*p*-value
Intervention	–4.53	3.80	–1.19	.24
Student pretest	0.70	0.01	85.57	<.01
Cluster pretest	0.78	0.07	11.26	<.01
Block 1	518.09	5.77	89.73	<.01
Block 2	526.67	6.05	87.01	<.01
Block 3	512.75	8.79	58.34	<.01
Block 4	517.15	7.83	66.02	<.01
Block 5	542.43	11.02	49.24	<.01
Block 6	507.61	4.81	105.47	<.01
Block 7	518.74	4.87	106.45	<.01
Block 8	520.84	4.94	105.34	<.01
Block 9	508.61	7.83	64.96	<.01
Variance components				
Level 1 (student)	1,964.93	54.77	35.88	<.01
Level 2 (school)	202.65	47.29	4.29	<.01

Source: 2007, 2008, and 2009 Colorado Student Assessment Program data.

Table K5. Mixed model results for student achievement impact analysis Cohort 3 subtest

Parameter	Estimate	Standard error	Test statistic	*p*-value
Intervention	5.59	4.22	–2.68	.19
Student pretest	0.71	0.01	80.32	<.01
Cluster pretest	0.80	0.07	10.76	<.01
Block 1	485.69	6.30	77.12	<.01
Block 2	498.90	6.64	75.1	<.01
Block 3	483.28	9.64	50.15	<.01
Block 4	482.09	8.56	56.34	<.01
Block 5	503.75	12.06	41.77	<.01
Block 6	483.39	5.32	90.85	<.01
Block 7	482.22	5.41	89.18	<.01
Block 8	480.22	5.37	89.49	<.01
Block 9	458.28	8.76	52.32	<.01
Variance components				
Level 1 (student)	1,770.39	46.21	38.31	<.01
Level 2 (school)	241.56	56.23	4.3	<.01

Source: 2007, 2008, and 2009 Colorado Student Assessment Program data.

Table K6. Mixed model results for the student achievement maximum likelihood estimation method sensitivity analysis

Parameter	Estimate	Standard error	Test statistic	*p*-value
Intervention	0.58	3.16	0.18	.85
Grade	–28.23	1.08	–26.06	<.01
Student pretest	0.70	0.01	104.27	<.01
Cluster pretest	0.78	0.06	13.8	<.01
Block 1	502.90	4.75	105.94	<.01
Block 2	513.53	5.01	102.52	<.01
Block 3	498.88	7.29	68.46	<.01
Block 4	500.55	6.47	77.38	<.01
Block 5	523.19	9.08	57.63	<.01
Block 6	495.82	4.00	124	<.01
Block 7	499.30	4.06	123.12	<.01
Block 8	501.00	4.07	123.02	<.01
Block 9	486.07	6.56	74.13	<.01
Variance components				
Level 1 (student)	1,920.68	37.71	50.94	<.01
Level 2 (school)	138.83	28.86	4.81	<.01

Source: 2007, 2008, and 2009 Colorado Student Assessment Program data.

Table K7. Mixed model results for the student achievement minimum variance quadratic unbiased estimation method sensitivity analysis

Parameter	Estimate	Standard error	Test statistic	*p*-value
Intervention	0.59	3.01	0.19	.85
Grade	–28.23	1.09	–25.99	<.01
Student pretest	0.70	0.01	103.94	<.01
Cluster pretest	0.78	0.05	14.51	<.01
Block 1	502.91	4.52	111.21	<.01
Block 2	513.49	4.75	108	<.01
Block 3	498.97	6.93	72.04	<.01
Block 4	500.59	6.15	81.46	<.01
Block 5	523.18	8.64	60.56	<.01
Block 6	495.83	3.81	130.11	<.01
Block 7	499.32	3.87	129.08	<.01
Block 8	501.09	3.89	128.75	<.01
Block 9	486.07	6.25	77.81	<.01
Variance components				
Level 1 (student)	1,938.43	38.03	50.97	<.01
Level 2 (school)	123.84	22.67	5.46	<.01

Source: 2007, 2008, and 2009 Colorado Student Assessment Program data.

Table K8. Mixed model results for the student achievement case deletion treatment of missing data sensitivity analysis

Parameter	Estimate	Standard error	Test statistic	*p*-value
Intervention	0.51	4.38	0.12	.91
Grade	–27.60	1.08	–25.44	<.01
Student pretest	0.70	0.01	102.48	<.01
Cluster pretest	0.68	0.07	9.29	<.01
Block 1	505.27	6.61	76.46	<.01
Block 2	521.70	7.07	73.82	<.01
Block 3	499.55	10.19	49.02	<.01
Block 4	501.71	9.15	54.83	<.01
Block 5	533.01	12.65	42.12	<.01
Block 6	496.38	5.55	89.4	<.01
Block 7	502.69	5.61	89.56	<.01
Block 8	505.01	5.58	90.52	<.01
Block 9	481.98	9.30	51.84	<.01
Variance components				
Level 1 (student)	1,760.06	31.69	55.54	<.01
Level 2 (school)	288.21	60.04	4.8	<.01

Source: 2007, 2008, and 2009 Colorado Student Assessment Program data.

Table K9. Mixed model results for impact analysis on Wave 3 student motivation survey

Parameter	Estimate	Standard error	Test statistic	*p*-value
Intervention	0.01	0.03	0.26	.79
Grade	0.04	0.01	4.61	<.01
Block 1	3.33	0.04	77.37	<.01
Block 2	3.26	0.04	91.84	<.01
Block 3	3.36	0.06	55.01	<.01
Block 4	3.33	0.04	77.65	<.01
Block 5	3.23	0.08	38.73	<.01
Block 6	3.24	0.04	82.03	<.01
Block 7	3.27	0.03	108.66	<.01
Block 8	3.21	0.03	95.04	<.01
Block 9	3.38	0.05	68.27	<.01
Variance components				
Level 1 (student)	0.22	0.01	42.81	<.01
Level 2 (school)	0.00	0.00	2.27	.02

Source: Wave 3 Survey of Student Motivation.

Table K10. Mixed model results for impact analysis on the Posttest Student Motivation Survey

Parameter	Estimate	Standard error	Test statistic	*p*-value
Intervention	0.01	0.03	0.48	.63
Grade	0.04	0.01	5.6	<.01
Block 1	3.26	0.04	77.12	<.01
Block 2	3.35	0.03	96.15	<.01
Block 3	3.34	0.05	63.5	<.01
Block 4	3.28	0.05	69.86	<.01
Block 5	3.30	0.07	49.65	<.01
Block 6	3.34	0.03	101.61	<.01
Block 7	3.30	0.03	98.74	<.01
Block 8	3.27	0.03	95.21	<.01
Block 9	3.43	0.05	69.78	<.01
Variance components				
Level 1 (student)	0.22	0.00	50.67	<.01
Level 2 (school)	0.01	0.00	3.01	<.01

Source: Posttest Survey of Student Motivation.

Table K11. Mixed model results from Teacher Test of Assessment Knowledge baseline comparison

Parameter	Estimate	Standard error	Test statistic	*p*-value
Intervention	0.58	0.69	0.84	.40
Late entry	–0.45	1.14	–0.39	.70
Years teaching experience	0.04	0.13	0.31	.76
Years teaching math	–0.05	0.13	–0.37	.71
Block 1	36.49	1.17	31.08	<.01
Block 2	36.19	0.95	37.97	<.01
Block 3	34.82	1.43	24.42	<.01
Block 4	37.55	1.24	30.3	<.01
Block 5	36.43	2.08	17.56	<.01
Block 6	34.83	0.98	35.7	<.01
Block 7	36.74	0.86	42.48	<.01
Block 8	35.33	0.91	38.77	<.01
Block 9	34.75	1.26	27.55	<.01
Variance components				
Level 1 (teacher)	37.90	4.45	8.51	<.01
Level 2 (school)	0.56	1.60	0.35	.73

Source: Teacher background information and baseline Test of Assessment Knowledge.

Table K12. Mixed model results from impact analysis on Teacher Test of Assessment Knowledge

Parameter	Estimate	Standard error	Test statistic	*p*-value
Intervention	2.78	0.99	2.81	.01
Late entry	–1.68	1.00	–1.68	.09
Teacher pretest	0.53	0.09	5.79	<.01
Cluster pretest	0.61	0.23	2.63	.01
Years teaching experience	–0.08	0.12	–0.71	.48
Years teaching math	0.16	0.13	1.26	.22
Block 1	39.68	1.45	27.41	<.01
Block 2	39.13	1.21	32.43	<.01
Block 3	35.67	1.86	19.17	<.01
Block 4	38.41	1.66	23.11	<.01
Block 5	43.27	2.60	16.61	<.01
Block 6	37.74	1.29	29.22	<.01
Block 7	38.61	1.08	35.75	<.01
Block 8	38.55	1.21	31.94	<.01
Block 9	38.39	1.69	22.73	<.01
Variance components				
Level 1 (teacher)	42.27	4.81	8.79	<.01
Level 2 (school)	3.19	2.71	1.18	.24

Source: Teacher background information, baseline Test of Assessment Knowledge, and Posttest Test of Assessment Knowledge.

Table K13. Mixed model results for the impact Teacher Assessment Work Sample

Parameter	Estimate	Standard error	Test statistic	*p*-value
Intervention	0.01	0.06	0.22	.82
Late entry	–0.01	0.06	–0.19	.85
Teacher pretest	0.01	0.00	1.39	.17
Cluster pretest	0.01	0.01	0.59	.56
Years teaching experience	–0.01	0.01	–1.33	.18
Years teaching math	0.01	0.01	1.42	.16
Block 1	1.64	0.09	18.57	<.01
Block 2	1.53	0.07	20.77	<.01
Block 3	1.59	0.11	14.15	<.01
Block 4	1.53	0.10	15.67	<.01
Block 5	1.84	0.15	12.06	<.01
Block 6	1.61	0.07	22.25	<.01
Block 7	1.61	0.06	25.49	<.01
Block 8	1.56	0.07	22.54	<.01
Block 9	1.69	0.10	17.16	<.01
Variance components				
Level 1 (teacher)	0.14	0.02	7.71	<.01
Level 2 (school)	0.01	0.01	1.13	.26

Source: Teacher background information, baseline Test of Assessment Knowledge, and Posttest Teacher Assessment Work Sample.

Table K14. Mixed model results for the Teacher Report of Student Involvement

Parameter	Estimate	Standard error	Test statistic	*p*-value
Intervention	0.05	0.03	1.65	.10
Entry	0.06	0.03	2.21	.03
Teacher pretest	–0.01	0.01	–4.11	<.01
Cluster pretest	–0.02	0.01	–2.25	.03
Years teaching experience	–0.01	0.01	–0.85	.40
Years teaching math	0.01	0.01	1.03	.31
Block 1	0.39	0.04	8.68	<.01
Block 2	0.36	0.04	9.29	<.01
Block 3	0.31	0.05	5.20	<.01
Block 4	0.31	0.05	6.11	.02
Block 5	0.18	0.08	2.34	<.01
Block 6	0.32	0.04	8.18	<.01
Block 7	0.37	0.03	11.31	<.01
Block 8	0.32	0.04	8.86	<.01
Block 9	0.33	0.05	6.39	<.01
Variance components				
Level 1 (teacher)	0.03	0.01	8.77	<.01
Level 2 (school)	0.00	0.00	1.52	.13

Source: Teacher background information, baseline Test of Assessment Knowledge, and Posttest Teacher Report of Student Involvement.

References

Airasian, P. W., & Jones, A. M. (1993). The teacher as applied measurer: Realities of classroom measurement and assessment. *Applied Measurement in Education, 6,* 241–254.

Arter, J. (2001). Learning teams for classroom assessment literacy. *National Association of Secondary School Principals Bulletin, 85,* 53–65.

Arter, J., & Busick, K. U. (2001). *Practice with student-involved assessment.* Portland, OR: Assessment Training Institute.

Arter, J., & Chappuis, J. (2006). *Creating and recognizing quality rubrics.* Portland, OR: Assessment Training Institute.

Aschbacher, P. R. (1999). *Developing indicators of classroom practice to monitor and support school reform* (CSE Technical Report 315). Los Angeles: University of California, Los Angeles. National Center for Research on Evaluation, Standards, and Student Testing.

Bandura, A. (1994). Self-efficacy. In V. S. Ramachaudran (Ed.), *Encyclopedia of human behavior* (Vol. 4, pp. 71–81). New York: Academic Press.

Bandura, A. (1997). *Self-efficacy: The exercise of control.* New York: W. H. Freeman.

Bangert-Drowns, R. L., Kulik, C. L. C., Kulik, J. A., & Mogran, M. T. (1991). The instructional effect of feedback in test-like events. *Review of Education Research, 61,* 213–238.

Bennett, R. E. (2009). *Formative Assessment: Can the Claims for Effectiveness Be Substantiated?* Princeton, New Jersey: Educational Testing Services.

Bergan, J. R., Sladeczek, I. E., Schwarz, R. D., & Smith, A. N. (1991). Effects of a measurement and planning system on kindergarteners' cognitive development and educational programming. *American Educational Research Journal, 28,* 683–714.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning: Putting it into practice.* London: Open University Press.

Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–74.

Black, P., & Wiliam, D. (1998b). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 139–144.

Bloom, H. S. (2008). The core analytics of randomized experiments for social research. In P. Alasuutari, L. Bickman, & J. Brannen (Eds.), *The Sage handbook of social research methods* (pp. 115–133). Los Angeles: Sage.

Bloom, H. S., Bos, J. M., & 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., & Black, A. R. (2005). *Using covariates to improve precision: Empirical guidance for studies that randomize schools to measure the impacts of educational interventions.* (MDRC Working Papers on Research Methodology) Access ERIC: FullText (070 Information Analyses; 110 Numerical/Quantitative Data; 143 Reports--Research). New York: Manpower Demonstration Research Corporation.

Borko, H., Stecher, B. M., Alonzo, A. C., Moncure, S., & McClam, S. (2005). Artifact packages for characterizing classroom practice: A pilot study. *Educational Assessment, 10*(2), 73–104.

Boud, D. (1986). *Implementing student self-assessment* (HERSDA Green Guide 5). Kensington, New South Wales: Higher Education Research and Development Society of Australasia.

Chappuis, J. (2007). *Learning team facilitator handbook: A resource for collaborative study of Classroom Assessment for Student Learning.* Portland, OR: Educational Testing Service.

Chappuis, J., & Chappuis, S. (2002) *Understanding school assessment: A parent and community guide to help students learn.* Portland, OR: Assessment Training Institute.

Chappuis, S., Stiggins, R., Arter, J., & Chappuis, J. (2006). *Assessment for learning: An action guide for school leaders.* Portland, OR: Educational Testing Service.

Charter, P. (1984). *Marking and assessment in English.* London: Methuen.

Cizek, G. J. (2010). An introduction to formative assessment: History, characteristics, and challenges. In H. L. Andrade & G. J. Cizek (Eds.), *Handbook of formative assessment* (pp. 3–17). New York: Routledge.

Clare, L. (2000). *Using teachers' assignments as an indicator of classroom practice* (CSE Technical Report 532). Los Angeles: University of California, Los Angeles. National Center for Research on Evaluation, Standards, and Student Testing.

Clare, L., Valdés, R., Pascal, J., & Steinberg, J. R. (2001). *Teachers' assignments as indicators of instructional quality in elementary schools* (CSE Technical Report 545). Los Angeles: University of California, Los Angeles. National Center for Research on Evaluation, Standards, and Student Testing.

Connell, J. P., Spencer, M. B., & Aber, J. L. (1994). Educational risk and resilience in African-American youth: Context, self, action, and outcomes in school. *Child Development, 65,* 493–506.

CTB/McGraw-Hill. (2009). *Colorado Student Assessment Program technical report 2009.* Monterey, CA: Author.

Eccles, J. S., Wigfield, A., Flanagan, C. A., Miller, C., Reuman, D. A., & Yee, D. (1989). Self-concepts, domain values, and self-esteem: Relations and changes at early adolescence. *Journal of Personality, 57,* 283–310.

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods, 12*(2), 121–138.

Finn, J. D. (1993). *School engagement and students at risk.* Washington, DC: National Center for Education Statistics.

Gallup Organization. (2005). *Narrative report of the 2005 principal and superintendent study.* Omaha, NE: Author.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research.* Thousand Oaks, CA: Sage.

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., & Elek-Fisk, E. (2003). Methods for handling missing data. In J. A. Schinka & W. F. Velicer (Eds.). *Research Methods in Psychology* (pp. 87–114). New York: John Wiley & Sons.

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007) How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science, 8,* 206-213.

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

Hedges, L. V. (2009). Effect sizes in nested designs. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *Handbook of research synthesis and meta-analysis, 2nd Edition.* New York: Russell Sage Foundation.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* San Diego, CA: Academic Press.

Institute for Research Reform in Education. (1998). Research Assessment Package for Schools (RAPS) manual. Retrieved March 21, 2006, from http://www.irre.org/publications/pdfs/RAPS_manual_entire_1998.pdf

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119,* 254–284.

Lindemann, E. (1982). *A rhetoric for writing teachers.* New York: Oxford University Press.

Liu, X., Spybrook, J., Congdon, R., Martinez, A., & Raudenbush, S. (2006). *Optimal design for multi-level and longitudinal research* (Version 1.77): HLM Software.

Marso, R. N., & Pigge, F. L. (1993). Teachers' testing knowledge, skills, and practice. In S. L. Wise (Ed.), *Teacher training in measurement and assessment skills* (pp. 129–185). Lincoln, NE: Burros Institute of Mental Measurements, University of Nebraska-Lincoln.

Marzano, R., Gaddy, B., & Dean, C. (2000). *What works in classroom instruction.* Aurora, CO: Mid-Continent Research for Education and Learning.

Matsumura, L. C., Patthey-Chavez, G. G., Valdes, R., & Garnier, H. E. (2002). Teacher feedback, writing assignment quality, and third-grade students' revision in lower- and higher-achieving urban schools. *Elementary School Journal, 103*(1), 3–25.

Matsumura, L. C., Garnier, H. E., Pascal, J., & Valdes, R. (2002). Measuring instructional quality in accountability systems: Classroom assignments and student achievement. *Educational Assessment, 8*(3), 207–229.

Matsumura, L. C., Garnier, H. E., Slater, S. C., & Boston, M. D. (2008). Toward measuring instructional interactions "at scale." *Educational Assessment, 13,* 267–300.

Mertler, C. A. (2009). Teachers' assessment knowledge and their perceptions of the impact of classroom assessment profissional development. *Improving Schools, 12*(2), 101–113.

Mertler, C. A., & Campbell, C. (2005). *Measuring teachers' knowledge and application of classroom assessment concepts: Development of the Assessment Literacy Inventory.* Paper presented at the annual meeting of the American Education Research Association, Montreal, Québec.

Mid Continent Regional Advisory Committee. (2005). A report to the U.S. Department of Education: Educational challenges and technical assistance needs of the Mid-Continent Region. Author.

Midgley, C., Maehr, M. L., Hruda, L. Z., Anderman, E., Anderman, L., Freeman, K. E., Gheen, M., Kaplan, A., Kumar, R., Middleton, M. J., Nelson, J., Roeser, R., & Urdan, T. (2000). *Manual for the Patterns of Adaptive Learning Scales (PALS).* Ann Arbor, MI: University of Michigan.

Moerbeek, M. (2004). The consequence of ignoring a level of nesting in multilevel analysis. *Multivariate Behavioral Research, 39*, 129-149.

Murray, D. M., Hannan, P. J., & Baker, W. L. (1996). A Monte Carlo study of alternative responses to intraclass correlation in community trials: Is it ever possible to avoid Cornfield's penalties? *Evaluation Review, 20*(3), 313–337.

No Child Left Behind Act of 2001. (2002). Pub L. No. 107–110, 115 Stat. 1425.

O'Connor, K. (2007). *A repair kit for grading: 15 fixes for broken grades.* Portland, OR: Educational Testing Service.

O'Sullivan, R. G., & Chalnik, M. K. (1991). Measurement-related course work requirements for teacher certification and recertification. *Educational Measurement: Issues and Practice, 10*(1), 17–23.

O'Sullivan, R. G., & Johnson, R. J. (1993). *Using performance assessment to measure teachers' competence in classroom assessment.* Paper presented at the annual meeting of the American Education Research Association, Atlanta, GA.

Pajares, F. (1996). Self-efficacy beliefs in academic settings. *Review of Educational Research, 66*(4), 543–578.

Plake, B. S., & Impara, J. C. (1997). Teacher assessment literacy: What do teachers know about assessment? In G. D. Phye (Ed.), *Handbook of classroom assessment: Learning, adjustment, and achievement* (pp. 55–68). San Diego, CA: Academic Press.

Plake, B. S., Impara, J. C., & Fager, J. J. (1993). Assessment competencies of teachers: A national survey. *Educational Measurement: Issues and Practice, 12*(4), 10–12, 39.

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

Ramaprasad, A. (1983). On the definition of feedback. *Behavioral Science, 28,* 4–13.

Raudenbush, S. & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Raudenbush, S., Spybrook, J., Liu, X., & Congdon, R. (2006). Optimal design for longitudinal and multilevel research: Documentation for the "Optimal Design" software. Retrieved June 1, 2006, from http://sitemaker.umich.edu/group-based/files/odmanual-20060517-v156.pdf

Ross, J. A., Hogaboam-Gray, A., & Rolheiser, C. (2002). Student self-evaluation in grade 5-6 mathematics effects on problem-solving achievement. *Educational Assessment, 8*(1), 43–59.

Rowan, B., Camburn, E., and Correnti, R. (2004). Using teacher logs to measure the enacted curriculum in large-scale surveys: A study of literacy teaching in 3rd grade classrooms. *Elementary School Journal*, 105, 75-102.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist, 55*(1), 68–78.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18,* 145–165.

Sadler, P. M., & Good, E. (2006). The impact of self- and peer-grading on student learning. *Educational Assessment, 11*(1), 1–31.

Schaffer, W. D. (1993). Assessment literacy for teachers. *Theory in Practice, 32*(2), 118–126.

Schneider, M. C., & Randel, B. (2010). Research on characteristics of effective professional development programs for enhancing educators' skills in formative assessment. In H. L. Andrade & G. J. Cizek (Eds.), *Handbook of formative assessment.* New York: Routledge.

Schochet, P. Z. (2008a). *Technical methods report: Guidelines for multiple testing in impact evaluations* (NCEE 2008-4018). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Schochet, P. Z. (2008b). Statistical power for random assignment evaluations of education programs. *Journal of Educational and Behavioral Statistics, 33,* 62–87.

Schunk, D. (1991). Self-efficacy and academic motivation. *Educational Psychologist, 26,* 207–231.

Skinner, E. A., & Belmont, M. J. (1993). Motivation in the classroom: Reciprocal effects of teacher behavior and student engagement across the school year. *Journal of Educational Psychology, 85*(4), 571–581.

Skinner, E. A., Wellborn, J. G., & Connell, J. P. (1990). What it takes to do well in school and whether I've got it: A process model of perceived control and children's engagement and achievement in school. *Journal of Educational Psychology, 82*(1), 23–32.

Stiggins, R. J., Arter, J. A., Chappuis, J., & Chappuis, S. (2004). *Classroom assessment for student learning: Doing it right – using it well.* Portland, OR: Assessment Training Institute.

Taras, M. (2003). To feedback or not to feedback in student self-assessment. *Assessment & Evaluation in Higher Education, 28*(5), 449–565.

Tunstall, P., & Gipps, C. (1996). Teacher feedback to young children in formative assessment: A typology. *British Educational Research Journal, 22*(4), 389–404.

What Works Clearinghouse (2008). *Procedures and Standards Handbook (Version 2.0).* Retrieved March 10, 2009, from, http://ies.ed.gov/ncee/wwc/pdf/wwc_procedures_v2_standards_handbook.pdf

White, B. Y., & Frederiksen, J. T. (1998). Inquiry, modeling, and metacognition: Making science accessible to all students. *Cognition and Instruction, 16,* 3–118.

Wigfield, A. (1994). Expectancy-value theory of achievement motivation: A developmental perspective. *Educational Psychology Review, 6,* 49–78.

Wiliam, D., Lee, C., Harrison, C., & Black, P. (2004). Teachers developing assessment for learning: Impact on student achievement. *Assessment in Education Principles Policy and Practice, 11*(1), 49–65.

Wolfe, E., & Jarvinen, D. (2002). *Standards-aligned classroom initiative: Year 2 evaluation report.* Springfield, Illinois: State Board of Education.

Wolfe, E., & Jarvinen, D. (2003). *Standards-aligned classroom initiative: Year 3 evaluation report.* Springfield, Illinois: State Board of Education.

Yoon, K. S., Duncan, T., Lee, S. W. Y., Scarloss, B., & Shapley, K. (2007). *Reviewing the evidence on how teacher professional development affects student achievement* (Issues & Answers Report, REL 2007- 033). Washington, DC: U.S. Department of Education, Institute of Educations Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southwest. Retrieved November 6, 2008, from http://ies.ed.gov/ncee/edlabs

York-Barr, J., Sommers, W. A., Ghere, G. S., & Montie, J. (2001). *Reflective practice to improve schools: An action guide for educators.* Thousand Oaks, CA: Corwin Press.

Yuan, Y. C. (n.d.). *Multiple imputation for missing data: Concepts and new developments* (Version 9.0). Rockville, MD: SAS Institute.

Zhang, Z., & Burry-Stock, J. A. (2003). Classroom assessment practices and teachers' self-perceived assessment skills. *Applied Measurement in Education, 16*(4), 323–342.


DEPARTMENT OF EDUCATION
UNITED STATES OF AMERICA