

Preliminary Meta-Analysis for Local Systemic Change Student Outcome Studies

Xiaodong Zhang
Joy Frechtling

December 2005

The purpose of this analysis is to examine six student outcome studies of projects from local systemic change (LSC) cohorts 5 and 6. In addition to providing early evidence of the project impacts on student achievement, the analysis will be useful in refining the meta-analysis framework proposed for evaluating all student outcome studies from cohorts 5-8, which are to be completed in 2007.

Our analysis begins with a structured overview of the studies, providing a formal summary of their main features and findings (Mosteller, Nave, and Miech, 2004). Second, we present effect size estimates about the effectiveness of the LSC projects. Finally, we provide a thematic analysis of programmatic and evaluation features across the studies. The thematic analysis is intended to illustrate the range of methodological approaches used to assess project impact.

Structured Overview

Among the projects being examined, cohorts 5 and 6 are equally represented. Five of the evaluations cover projects focusing on mathematics and one, a project focusing on science. Four projects target the elementary and middle school grades; the other two, middle to high school grades. Although the major feature of the LSC program is to provide curriculum-embedded professional development, only two evaluations clearly identified the curricula. Table 1 presents a brief summary of the six projects.

Table 1.—Summary of project characteristics

Project name/ID	Cohort	Content	Grade level	Curriculum
COMMSTEP (9819288)	5	Math	7-12	Connected Math, Mathscape, Mathematics in Context, Seeing and Thinking Mathematically
BEAMM (9819468)	5	Math	K-8	Not specified
DEMCI (9819592).....	5	Math	5-12	Connected Math, Mathematics in Context
PRIME (9911754)	6	Math	K-8	Not specified
TREASURmath (9911849)	6	Math	3-8	Not specified
Project Inquiry (9986869)	6	Science	3-8	Not specified

In the section below, the project intervention and related evaluation questions are summarized, and the methodologies are reviewed in terms of approaches in measuring treatment and outcomes, sampling, design, and analysis.¹ The findings as well as the strengths and weaknesses of each study are discussed.

¹ Measuring treatment and outcomes: description of strategies and instruments used to quantify intervention and intended outcomes. Sampling: description of sampling techniques, sample size, response rate, and unit of analysis. Design: description of the research design. Analysis: description of analytical techniques in drawing descriptive, correlation, or causal conclusions.

Colorado Mathematics Middle School Teacher Enhancement Project (COMMSTEP)

Purpose

COMMSTEP supported 500 mathematics teachers of grades 7-12 in 42 schools. Conducted in years 2 and 3 of the project, the study was designed to examine the impact that COMMSTEP professional development had on student achievement and teacher practice.

Methodologies

Measuring intervention and outcomes. The intervention was measured by differences between the treatment and comparison groups. Drawing on established instruments from the Horizon LSC survey and the Longitudinal Evaluation of School Change and Performance, the evaluation team developed a survey to measure seven separate categories of teacher practices and support. The reliability of survey scales was examined. Student achievement was assessed using the Colorado Student Assessment Program (CSAP).

Sampling. All teachers that participated in COMMSTEP were asked to fill out the survey. A comparison sample comprising teachers from seven geographically and demographically similar districts and 60 schools was then recruited. (A math-related instructional gift was provided to survey participants as an incentive.) The sample sizes were 82 non-COMMSTEP teachers representing 44 schools and 148 COMMSTEP teachers representing 36 schools. To ensure comparability between the two samples, schools were further matched on the percentage of students receiving free and reduced-price lunch. The final sample had 37 non-COMMSTEP schools and 35 COMMSTEP schools. The study did not specify the sample size and unit of analysis for student achievement. It appears that the student data were averaged across grade levels and analyzed at the school level using percentage of students who were proficient and advanced.

Design. The study used a quasi-experimental matched group design. However, the level of matching occurred at the school level, which was the unit of analysis for teacher practice.

Analysis. Sample characteristics and descriptive statistics for the teacher survey were presented. The differences between treatment and comparison teachers and student achievement were examined, with both significance and effect size statistics reported. It appears that an ANCOVA model was used to assess the relationship between hours of project participation and student achievement using 2001 CSAP scores as the covariate. Finally, path analyses were conducted to look at the direct and indirect effects of school participation on teacher collegiality, teacher-centered practices, student-centered practices, and average CSAP scores.

Findings

Descriptive data showed that COMMSTEP teachers were implementing more student-oriented math activities and were more likely to engage in professional collegial practices than the comparison teachers. However, these differences did not result in differences in student achievement. In addition, the number of hours of professional development was not related to student performance. The path models showed significant relationships between school participation and teacher collegiality and their use of student-centered practices, which significantly predicted CSAP scores.

Strengths and Weakness

The strength of the study lies its approach to drawing the comparison group, use of logic similar to the propensity score approach, and use of multiple analysis strategies. One limitation of the study is the small sample size, which limited the number of variables used in the path models. Another problem is the potential sample bias. With an average of three teachers per treatment schools and even fewer for the comparison schools, making connection between the responses of the limited sample to the average school performance in student achievement is somewhat problematic.

Broadening Mathematics Access to Mathematics in Maine (BEAMM)

Purpose

BEAMM is a K-8 math curriculum implementation project for 500 teachers in seven districts. The primary evaluation question was “Did the BEAMM initiative help improve 4th and 8th grade students’ mathematics performance as compared to non-BEAMM schools?” In addition, the study sought to examine the relationship between student improvement and professional development hours and teachers’ investigative math practices.

Methodologies

Measuring intervention and outcomes. The invention was measured by the differences between treatment and comparison groups. In addition, a dosage model was used to compare the differences between teachers who had over 100 hours of professional development and those with less than 20 hours. Horizon’s LSC survey was used to measure instructional practices. The instrument used to measure student achievement was the Maine Education Assessment (MEA). Validity and reliability of the MEA were reported.

Sampling. The teacher sample was 300 randomly selected teachers. The student sample included all 4th and 8th grade students. However, the unit of analysis for student data was the district,² thus severely limiting the sample size.

Design. For the student achievement analysis, a quasi-experimental design was used with seven treatment districts matched with comparison districts based on several demographic variables. However, comparison districts on average performed higher than the BEAMM districts in the pretest. Teacher data were analyzed using a one-group pre/post design.

Analysis. A *t*-test was used to detect the differences in student performance at 4th and 8th grades between treatment and comparison districts as well as between treatment districts and the state overall.³ The average rates of change in performance levels were presented in graphics. For teacher data, *t*-tests were performed on nine composite scores to look at differences between the 2004 and 1999 responses. In addition, composite scores were correlated to teacher professional development hours on a yearly basis from 2002 to 2004. Finally, a chi-square test of independence was used to test the difference in the

² The proposal had indicated that school would be the unit of analysis.

³ Only two of the four categories, meets standards and does not meets standards, were compared. The other two categories, exceed standards and partially meets standards, were not used.

distribution of composite scores among teachers with low numbers of professional development hours to those who had received the targeted 100 hours.

Findings

In general, the percentage of students meeting standards increased in both treatment and comparison districts. Despite a lower average at pretest, 8th grade students in treatment districts had a significantly higher percentage of students meeting the standards at posttest than those in the comparison districts. Few pre- and posttest differences were found for the 4th graders. The average percentage of yearly change was higher in treatment districts than comparison districts in both 4th and 8th grades. Treatment teachers scored significantly higher in six of eight scales in 2004 compared to 1999. However, the correlations between teacher composite scores and professional development hours were low and insignificant. The chi-square test revealed that teachers who had more than 100 hours of professional development scored significantly higher in two composite areas than did those with less than 20 hours.

Strengths and Weaknesses

The study drew on multiple years of data collection. However, the student achievement analysis is problematic in terms of its unit of analysis. The small sample of districts involved posed a serious problem for the significance test. The study did a better job in analyzing the teacher survey using both the composite score approach and a chi-square test. Furthermore, the study did not completely address the two original secondary questions. The issue addressed was the relationship between professional development and teacher practice, rather than the relationship of those two areas with student achievement.

Delaware Exemplary Mathematics 6-12 Curriculum Implementation (DEMCI)

Purpose

The project provided professional development in 16 school districts. A central evaluation goal was to determine if the quality of the math instruction students received influenced their scores on the mathematics portion of the state assessment.

Methodologies

Measuring intervention and outcomes. Unlike other studies that measured the project treatment in terms of inputs (i.e., difference between treatment and comparison groups, number of years of project intervention) or outputs (i.e., hours of professional development), this study measured the intervention in terms of an outcome, namely, quality of teacher instruction. The instrument was the Innovation Configuration component of the Concern-Based Adoption Model developed by Hall and Hord. Participating teachers were observed and rated by trained raters. Each teacher received a global rating of high, moderate, or poor based on the rating of 17 traits. Both validity and reliability of the instrument were examined. Student performance was assessed with the mathematics section of the Delaware Student Testing Program, and the results were presented in both scale scores and corresponding five-category performance levels.

Sampling. During 2002–03, a sample of 39 teachers and all of their 2,102 8th grade students participated in the study. The demographics of these students and their previous math performance at the 5th grade were similar to those in the state.

Design. The study used a one-group pre/posttest design.

Analysis. Teacher data were analyzed using descriptive and one-way ANOVA (quality of instruction), while two-way ANOVA (quality of instruction, 5th grade performance) was used for student data.

Findings

Significant differences were found in nine of 17 traits among the means of high, moderate, and poor quality instruction. In addition, seven traits were significantly different between high and poor quality instruction. Student data showed that both the quality of instruction and prior math achievement had significant effects on 8th grade student performance. The patterns of student achievement at different performance level were further examined.

Strengths and Weaknesses

Overall, this is a relatively rigorous evaluation. The study could be further strengthened by looking at the relationship between quality of instruction and the amount of professional development to solidify the claims that instruction is indeed affected by the professional development from the project. Of course, the evidence would have been stronger if it were based on data from more than one year.

PRIME Mathematics Project

Purpose

PRIME is a math intervention targeting grades K–8 teachers. The evaluation looked at the project impacts on teachers’ beliefs about math instruction, collaboration with other teachers, and student achievement.

Methodologies

Measuring intervention and outcomes. The level of project intervention was measured by the difference between posttests (2002) and pretests (1999). In addition, the extent of professional development was calculated by the proportion of teachers participating in the professional development (above or below 75 percent) at the school level, and by the number of years a student had a PRIME teacher at the student level. Teachers’ beliefs were measured using the Horizon LSC survey. A self-developed scale based on items from the Horizon survey was used to measure teacher collaboration. Finally, student achievement was assessed using the Illinois Standards Achievement Test for 3rd and 5th grade students from 2000 through 2003.

Sampling. It was unclear how many teachers were surveyed. The student data came from 26 schools in one district with 826 students. The unit of analysis was the school in the trend analysis and the

relationship with participation rate, and the student in the relationship between student achievement and number of years having a PRIME teacher.

Design. The design of the study is difficult to characterize. The teacher data analysis used essentially a pre/posttest design. The student data were used in a trend analysis, an unmatched group comparison with the state average, and a dosage model to examine the relationship between school participation rate and the percentage of students at different achievement levels.

Analysis. A *t*-test was used to analyze the teacher data. Descriptive statistics and ANOVA were used for the student data.

Findings

The results showed significant improvement in the level of teachers' belief on the majority of the items in seven categories. However, no significant differences were found for teacher collaboration. Student achievement at the 3rd grade was significantly higher than the state average; 5th grade achievement was comparable to the state average. Due to extremely small sample sizes, the relationships between a school's student performance level and teacher participation rate were mostly not significant, except that schools whose participation rate was above 75 percent had a significantly smaller percentage of grade 5 students at level 1 (academic warning) than schools with participation below 75 percent. Nevertheless, the number of years a student has had a PRIME teacher had a significant effect on the achievement of students from low socioeconomic status families.

Strengths and Weaknesses

The study used multiple years of data collection. It involved many layers of subanalyses, some of which were not presented with enough detail and clarity. A major flaw of the study was found in the section regarding the relationship between the rate of school professional development participation and student achievement. The measure of school participation rate was crude, and it did not specify what counts as "participation." Using school as the unit of analysis in this case resulted in extremely small sample sizes, and thus it is not surprising that most of the findings were statistically insignificant.

TREASURmath Project

Purpose

The project provided curriculum-embedded professional development to math teachers for grades 3-8. The relatively simple evaluation examined the relationship between levels of professional development and student achievement.

Methodologies

Measuring intervention and outcomes. The level of project treatment was characterized as a dichotomous variable with "1" indicating "meeting goals of receiving 130 hours of professional development" and "0," "not meeting goals." SAT-9 scale scores (total and subsets) were used to measure student achievement.

Sampling. The study sample involved all grade 3-8 teachers (N=83) and their students (N=2,300) who have had at least two years of student data to produce the growth measures. The unit of analysis was the student.

Design. The design of the study used a value-added model. Students served as their own control group in that each test was a pretest to the subsequent test. The study also compared the treatment group student achievement with the national norm, which in essence was a quasi-experimental, unmatched design.

Analysis. The study used an ANCOVA model, controlling for teacher experience and grade assignment.

Findings

The results showed that both primary and middle grade students performed better than expected based on national experience. In addition, students whose teachers met the professional development goals had higher growth in overall scores than those whose teachers did not meet those goals. Performance growth of middle school students was greater than that of primary students. The same was true on the problem-solving and the procedure subtests, except the impact of teacher professional development on the latter was not significant.

Strengths and Weaknesses

Although somewhat limited in scope, the methodologies and results were presented clearly. However, the analysis did not address the relationship between implementation and student achievement, as it was originally proposed.

Project Inquiry

Purpose

Project Inquiry was the only science project reviewed. It focused on staff development for teachers of grades 3-8 in two districts. Conducted in 2001-02, the evaluation investigated the relationship between components of teacher instruction and student achievement.

Methodologies

Measuring intervention and outcomes. The project evaluation was based on the differences between 2002 and 2001 data. In addition, the number of hours of teacher professional development was categorized as a five-level variable (0, 1-9, 10-19, 20-39, and 40 more). Two instruments were used to measure student achievement for the 5th graders: the first was the LSC Science Assessment developed by Horizon, drawing on 52 multiple-choice items from NAEP and TIMSS; the second was the PASS Science Assessment, developed by WestEd, that contains multiple-choice, open-ended, and constructed-response items. Data on teacher instructional practices came from both self-reported surveys and classroom observation. Teacher questionnaires accompanied both LSC and PASS assessments. In addition, the Classroom Observation Protocol for Project Inquiry was used to gather observation data.

Sampling. The sampling plan for the evaluation was equally complex. All 225 teachers completed the LSC surveys, and the LSC pre/posttest scores from their 3,712 students were analyzed. For the PASS assessment and survey, 99 teachers were randomly selected, and 10 students from each of their classes were then randomly selected. Ninety-seven teachers were observed. The units of analysis included both teacher and student.

Design. A varying level of dosage, pre- and posttest design was used to relate the amount and quality of science instruction to science achievement.

Analysis. Descriptive statistics for the five composite scores of the LSC survey were reported both in the aggregate and the disaggregate by the level of teacher professional development hours. Descriptive statistics for classroom observation were also reported both as overall ratings and by items. Descriptive statistics for student pre- and posttest results were presented by five science subject areas. The relationships between LSC assessment scores and the teacher survey responses were analyzed using what appears to be hierarchical linear modeling in a stepwise fashion, with both composite scores and item responses serving as independent variables, and selected variables as covariates. A similar approach was used to examine the relationship between classroom observation and the results from the LSC and PASS assessments.

Findings

The results showed that 5th grade students made achievement gains from fall to spring. Both survey and classroom observation data suggested that teachers were moving toward a more rigorous, relevant, and inquiry-oriented science curriculum. However, there appeared to be some discrepancy between teachers' self-reported and observed behavior. While many teachers reported often engaging students in higher level thinking, nearly half of the teachers observed asked few or no higher level questions, and the composite scores from the teacher survey showed a few significant relationships with science achievement. In contrast, the overall ratings of the lessons were significantly related to both total scores and most subtest scores for both the LSC and PASS assessments. Finally, in spite of its overall low level, the amount of professional development teachers received had significant positive effects on units/lessons taught and inquiry materials used for teachers as well as student achievement.

Strengths and Weaknesses

This is by far the most comprehensive and well-conducted evaluation of the six, reflecting many best practice principles in evaluation. For example, multiple instruments were used to measure teacher and student outcomes; different sampling strategies were employed appropriately; the analysis strategies were more advanced; and finally, all of the procedures and results were clearly articulated. The only drawback of this evaluation is that it was based on one year of data collected in the middle of the intervention. Evidence might be stronger if it involved multiple years of data collection.

Statistical Synthesis

The limited evidence from six evaluations suggests that by and large, both LSC projects and the extent of professional development had significantly positive impacts on students and teachers (table 2). In addition, the quality of teacher instruction was found to have positive associations with student

achievement. Collectively, these studies give evidence about how the treatment can be characterized and how its outcomes can be attributed back to it.

Table 2.—Impact of projects on students and teachers

Project name/ID	Project		Professional development		Teacher practice
	Teacher practice/attitude	Student achievement	Teacher practice	Student achievement	Student achievement
COMMSTEP (9819288)	+	+/O		O	
BEAMM (9819468)		+	+	O/+	
DEMCI (9819592).....					+
PRIME (9911754)	+/O	+		O/+	
TREASURmath (9911849)				+	
Project Inquiry (9986869)	+	+	+	+	+

+ denotes significant positive findings; O suggests the finding is not statistically significant. Cases with multiple notations indicate that the findings draw on more than one data sources and designs.

A major goal of our analysis is to measure the impacts on student achievement using effect size estimates. A major challenge faced by evaluators of large-scale evaluation is the diverse designs and lack of common measures of the outcomes. Effect size is often suggested as a way to resolve these problems by quantifying the results on a common scale, and to avoid the problems resulting from a small sample.

Table 3 reports the statistics used for the effect size calculation, sources of these statistics from each study, and the resulting effect size estimates. Here we focus on the effect sizes for student achievement by presenting one effect size estimate for each study.⁴ Except for one study that reported the effect size, the effect size estimates for the five other studies have been calculated by Westat, using formulas provided by Lipsey and Wilson (2001). The strategies for calculating these effect sizes vary from study to study due to differences in data provided: two use mean differences and standard deviation or confidence interval; four use F statistics and sample size or degree of freedom.

The estimates are presented as “d” effect size. Overall, the average effect size of six studies is 0.06, which can be characterized as small by Cohen’s standards and is not unexpected for large-scale programs (Borman, et al., 2003). While five effect sizes are positive, one is negative. In that case, the negative sign was affected by a larger negative estimate for school participation than that for student exposure to teacher receiving professional development.⁵ Strictly speaking, the effect sizes from different designs or measuring different aspects are not entirely comparable. Due to small sample size, we take a more liberal approach here in comparing and averaging different effect size estimates.

⁴ If multiple effect sizes were available for one study, we took the average instead of using the proposed HLM model for the full meta-analysis.

⁵ In this case, the average effect size for school participation rate is -.23, and the effect size for student exposure is .05. The school participation rate is not statistically significant due to extremely small sample sizes.

Table 3.—Elements of the effect size estimation

Project name/ID	Statistics used for effect size estimation						Location		Effect size	
	Mean difference	Standard deviation	Confidence interval	F statistics	Sample size	Degree of freedom	Reported effect size	Page #		Table/graph #
COMMSTEP (9819288).....	X	X					X	11	T4	0.06
BEAMM (9819468)	X		X					6-7	T2, G1-2	0.07
DEMCI (9819592) ..				X		X		12	T9	0.08
PRIME (9911754)...	X	X		X		X		7, 11	T2,4,7	-0.09
TREASURmath (9911849).....				X		X		7	NS	0.15
Project Inquiry (9986869).....				X	X			17,20, 23	T7H, 8F, 9F	0.15

NS = not specified.

Thematic Synthesis

All studies share the core theme of looking at the effect of curriculum-embedded professional development on student learning, but a closer look at the evaluations show some interesting differences in 1) specific relationship examined, 2) procedures for quantifying treatment/intervention for evaluation purposes, 3) outcome measures, 4) sampling and analysis units, 5) design, and 6) analyses. In this section, we present a thematic synthesis with a cross-study discussion of the programmatic and evaluation features of the six studies. The thematic analysis is intended to illustrate the range of methodological approaches used to assess project impact. As more evaluations are received, we plan to continue to track and catalogue these differences. At the end of our work, we will more closely assess how these differences contribute to the pattern of outcomes that emerge.

Specific Relationships Examined

The main charge of the student outcome studies is to find out whether the program works to improve student learning. To do so, we must examine the correlation or causal relationships between the treatment/intervention and student outcomes, primarily by using quantitative approaches. How treatment is defined and quantified can vary, as can the auxiliary questions to be addressed. These additional questions can expand the scope of the studies considerably, and addressing them often requires the use of qualitative strategies.

Table 4 presents a summary of specific relationships examined by the six studies. It shows that the majority of the studies directly examined the effect of the project on student achievement and teacher practices. In addition, many evaluations looked at the relationships between the extent of teacher professional development and student achievement and teacher practices. A few studies also assessed the relationship between teacher practices and student achievement.

Table 4.—Specific relationships examined

Project name/ID	Project effect on:		Extent of professional development on:		Effect of teacher practices on student achievement
	Teacher practices	Student achievement	Teacher practices	Student achievement	
COMMSTEP (9819288).....	X	X		X	X
BEAMM (9819468)	X	X	X	X	
DEMCI (9819592).....					X
PRIME (9911754)	X	X		X	
TREASURmath (9911849)				X	
Project Inquiry (9986869).....	X	X	X	X	X

Quantifying Treatment

All evaluations are essentially looking at the relationship between the treatment (independent variable) and its outcomes (dependent variable). In most cases, however, while the measure for program outcomes is clearly stated, the measure for treatment is not. This discrepancy can be problematic, as some have argued that getting an accurate measure of the program intervention is more important for validity than a measure of the outcomes (Langbein, in press).

Quantifying the intervention is sometimes dealt with as a design issue. For example, an experimental or a quasi-experimental design measures the intervention by the presence or absence of the treatment either with one group (pre-post design) or multiple groups (treatment-comparison/control design). A nonexperimental design often characterizes the treatment by the variation of exposure to the intervention (i.e., dosage model) among participants. The following illustrates various ways of approaching the definition.

- **Holistic.** The treatment is considered holistically as a “program.” Its model uses a “1” for treatment and “0” for the lack of treatment. The advantages of this approach are that it is simple and that it is concerned with the complete system; the disadvantages are exactly the same, because in reality teachers in the treatment group might not receive the same amount of treatment, while those in the comparison group often receive some elements of the treatment from other sources. The approach is often criticized as creating a “black box.”
- **Proxy.** One may choose to use certain project quantifiable outputs as proxy measures. In the LSC studies, many used amount of professional development a teacher received for this measure. This approach has the advantage of being easily quantifiable and can be applied to nonexperimental design where only data on treatment group are collected; the drawbacks are that it measures limited aspects of the treatment.
- **Fidelity.** A more nuanced approach is to actually measure the implementation level/fidelity using composite (total), factor, or IRT score based on classroom observation and survey responses. The strength is that it offers evidence about different elements of the treatment. Another advantage is the flexibility of composite/factor scores because they can be either analyzed holistically or related to individual item response or observation. The drawbacks are often more technical with regard to the validity and reliability of the instruments and scoring techniques involved.

A closer look at the six studies reveals differences in terms of how project intervention was quantified—often implicitly (table 5).

- Four evaluations used the holistic approach. Specifically, two projects used the differences between pre- and posttest for the treatment group only, while two used the differences between pre- and posttest in both treatment and comparison groups to reflect the treatment.
- Five used proxy measures such as extent of professional development. To characterize the extent of professional development, four projects used hours as the measure. While treating “hour” as a continuous variable is preferred statistically, only one project did so. Other projects often treated it as a categorical variable because of the ease in data collection. For example, one project used a more nuanced five-category measure, one compared the differences between high (PD>100) and low (PD<20) participation based on the distribution, and still another used a dichotomous variable distinguishing teachers who met the goal of receiving 130 hours and those who did not. The fourth project chose to measure professional development from both school and student perspectives.

The school participation rate was calculated by dividing the number of participants over the total number of teachers. This is an easy but rather crude measure mainly because it fails to differentiate high level participants from the low level ones. Meanwhile, a student exposure to professional development was measured by the number of years a student has had teachers with project-provided professional development. This measure could be useful if the amount of teacher professional development is known.

- None of the six studies included a measure of the implementation level of professional development. The indexing/scoring approach, however, was used in two studies to compile composite scores to measure teacher instruction quality or practices, which in themselves compare an outcome of the project and an interim variable to student achievement.

Table 5.—Quantifying treatment

Project name/ID	Presence/absence of treatment (holistic)	Extent of professional development (proxy)	Implementation level (fidelity)
COMMSTEP (9819288)	Diff=T-C (yr2003-yr2002)	Continuous	No
BEAMM (9819468)	Diff=T-C (yr2003-yr1999)	Diff=H(≥ 100 hr)-L(≤ 20 hr)	No
DEMCI (9819592)	No	No	Instruction quality (observation scores)
PRIME (9911754)	Diff=Post(yr2002)-Pre(yr1999)	<ul style="list-style-type: none"> • School participation rate = number participant/total number of teacher • Student exposure=number of years with PRIME teachers 	No
TREASURmath (9911849)	No	Diff=Meet Goal(≥ 130)-Not Meet Goal (<130)	No
Project Inquiry (9986869)	Diff=Post(yr2002)-Pre(yr2001)	Diff=(0hr, 1-9hr, 10-19hr, 20-39hr, 40 more hr)	Instructional practice (survey composite scores)

Measuring Outcomes

There are also important differences in the types of measures selected to assess both student and teacher outcomes (table 6).

To measure student achievement, four studies used state assessments. Given the requirement of No Child Left Behind, we expect such instruments will increasingly be used. The results of the state assessment were often reported in two ways. An individual performance status was measured by the proficiency or performance level, and school and/or district performance was often quantified as the percentage of students in each performance level. One project used a norm-referenced assessment. The remaining project used two external but complementary instruments to measure science achievement.

Five of the six studies directly measured teacher outcomes. Of those five, four examined instructional practice and attitudes using the Horizon teacher survey. Responses were analyzed by item and/or as composite scores. One used the PASS survey to triangulate the results from the Horizon survey. In addition, two projects used classroom observation to collect observed instead of self-reported outcomes. The one study that used both surveys and observation found some discrepancies between two results.

Table 6.—Measuring outcomes

Project name/ID	Student achievement	Teacher practice/attitude
COMMSTEP (9819288)	Colorado Student Assessment Program (% proficient/advanced)	<ul style="list-style-type: none"> Self-developed based on Horizon teacher survey and Longitudinal Evaluation of School Change and Performance (9 composite scores)
BEAMM (9819468)	Maine Educational Assessment (% meet standards, did not meet standards)	<ul style="list-style-type: none"> Horizon teacher survey (9 composite scores)
DEMCI (9819592)	Delaware Student Testing Program (scale score, 5-category performance level)	Innovative Configuration (high, moderate, low)
PRIME (9911754)	Illinois Standards Achievement test (% at/above standards)	<ul style="list-style-type: none"> Horizon teacher survey (7 clusters) Self-developed scales (collaboration, problem-solving skills)
TREASURmath (9911849)	SAT-9 (scale score)	No
Project Inquiry (9986869)	<ul style="list-style-type: none"> Horizon (% correct) PASS (% correct) 	<ul style="list-style-type: none"> Horizon teacher survey (items, 5 composite scores) Horizon observation (items, overall rating) PASS (items, overall rating)

Sampling and Analysis Units

Table 7 summarizes the sampling approaches involved in the six reviewed studies. We found related student outcomes from students in grades 3 to 8. Three studies looked at the trend of student achievement by comparing the current year of student performance with that of the previous year. In these studies, the samples often included all the students in selected grades, but the unit of analysis was

aggregated to the school or even the district level. As a result, the sample sizes for two of these studies were found to be inadequate. In other studies looking at the relationship between student achievement and their teacher’s professional development or practices, the samples included all students taught by the sampled teachers. The unit of analysis—the classroom—was generally adequate. In one study using a HLM model, the units of analysis included both classrooms and students.

Table 7.—Sampling strategies and analysis units

Project name/ID	Sampling strategies	Unit of analysis	Size
COMMSTEP (9819288)	Recruited from all 6-8th grade teachers	Teacher, School	T=200 (?) Sch=72
BEAMM (9819468)	<ul style="list-style-type: none"> • Random sample of teachers yearly • All 4th, 8th grade students in treatment and comparison districts 	Teacher, District	T=300 D=14
DEMCI (9819592)	Sample of eligible 8th grade teachers—representative	Teacher, Student	T=39 S=2102
PRIME (9911754)	<ul style="list-style-type: none"> • Teacher (Not specified) • Student (all 5th graders in cohort 3 2001–03 in 1 district) 	Teacher, School, Student	Tch=Not specified Sch=26 Std=826
TREASURmath (9911849)	All 3-8th grade teachers and students	Teacher, Student	T=83 S=2,300
Project Inquiry (9986869)	<ul style="list-style-type: none"> • Horizon (all 5th teachers, students) • PASS (99 teachers, each with 10 students in 1 district) • Observation (97 teachers) 	Teacher, Student	Horizon (T=225, S=3712) PASS (T=99, S=994)

Among studies that examined the teacher outcomes, three attempted to draw all teachers in the population, although one was less successful than the others due to low response rates. Two other evaluations adopted sampling strategies: one used random sampling, and the other drew a small sample based on availability and later tested the comparability of the sample with the population. The unit of analysis in these studies is the teacher.

Design

The majority of the six evaluations addressed multiple questions regarding students and teachers, with different designs being used to address respective questions. We will focus on the design for the primary question of student impact. Since none of the studies used an experimental design, the following discussion reflects different approaches in quasi-experimental and nonexperimental designs (table 8).

- Four projects used a one-group, treatment-only design, three of which involved pre- and posttest and one a value-added model. The threats to the one-group design, with the exception of the value-added model, are that of history and maturation; that is, one cannot be certain that the observed differences can be attributed to the intervention and not something else.
- A two-group, matched-comparison design was used by two other studies. The first one matched the treatment teachers with comparison teachers based on school characteristics. For the second, the matching occurred at the district level based on demographic variables. In order to further improve the matching, the study used prior student achievement as covariates in the model.

- Two of these studies used a dosage model to address the extent of professional development question with cross-sectional posttest only data.
- Time is another dimension. Half of the evaluations used data collected in two points in time, while the other half involved multiple years of collection. Only one project tracked individual students longitudinally.

Table 8. Design types

Project name/ID	Quasi-experimental			Nonexperimental		Longitudinal/ cohort
	One-group	Two-group		Dosage	Covariate- adjusted	
	Pre/post	Matched	Unmatched			
COMMSTEP (9819288).....		X				X
BEAMM (9819468)	X	X	X	X	X	X
DEMCI (9819592).....	X				X	
PRIME (9911754)	X			X		
TREASURmath (9911849)			X	X	X	X
Project Inquiry (9986869).....	X			X		

NOTE: Boldface indicates strategies were used for the primary question.

Analyses

The projects employed a range of data analysis techniques to evaluate student impacts. Table 9 shows that all of the studies used standard descriptive statistics to understand the characteristics of the sample and/or provide detailed and substantive information from the surveys or observations. In addition, inferential statistics such as significance tests (*t*, ANOVA, chi-square, F) were used to examine the correlation or causal claims about the program effect. Four evaluations employed at least one advanced techniques such as ANCOVA, multiple regression, path model, value-added model, and hierarchical linear model to strengthen the validity and reliability of the results.

Table 9.—Analysis techniques

Project name/ID	Descriptive	Trend analysis	Correlation	<i>t</i> -test	F-test	Chi-square	ANOVA	Path analys. /SEM	ANCOVA/ multiple regression	HLM
COMMSTEP (9819288).....	X			X				X	X	
BEAMM (9819468).....	X	X	X	X		X				
DEMCI (9819592).....	X						X			
PRIME (9911754)	X	X		X					X	
TREASURmath (9911849).....	X	X							X	
Project Inquiry (9986869).....	X				X					X

NOTE: Boldface indicates strategies were used for addressing the primary question.

Implications for Future Study

The analysis provides an initial examination of the six LSC cohorts 5-8 student outcome studies. Collectively, these studies show that the LSC projects had statistically significant impacts on student

achievement, but the magnitude of the impacts appeared to be small. In addition, they provide evidence that the amount of LSC professional development teachers receive has positive effect on teacher quality of instruction and might also affect student learning in a positive direction. There is also limited evidence that teacher instruction quality and practice are positively related to student achievement.

While the overall quality of the six studies is good and represents a marked improvement from the cohorts 1-4 evaluations previously reviewed (Zhang and Wang, 2003), there remain several limitations. Problems include small sample sizes, difficulties in implementing treatment/comparison designs, studies that were too short in duration to capture full program impact, and lack of clarity in describing approaches and procedures. In many ways, the limitations reflect the tensions between scientific rigor and the reality of limited financial and technical resources. We hope that by presenting each study in detail, the reader can obtain a better understanding of the range of the possibilities in each area, issues and challenges associated with each choice, and potential solutions and trade-offs.

This preliminary analysis will help us to further refine the meta-analysis coding protocols for characterizing project, evaluation and effect size descriptors. The full-scale meta-analysis to follow should thus be able to provide the following benefits:

- Calculate the effect size estimates for the LSC program. Due to small sample size, we averaged the effect size estimates within and across studies. In the full analysis of 20 studies, we will use techniques such as bias correction, homogeneity correction, and use of an analytical model of HLM to provide more rigorous effect size estimates. In addition, we hope to distinguish different kinds of effect size estimates such as those characterizing project, amount of professional development, and teacher practices on students.
- Present a thematic synthesis with a cross-study discussion of the programmatic and evaluation features to show interesting differences in 1) specific relationship examined, 2) procedures for quantifying treatment/intervention for evaluation purposes, 3) outcome measures, 4) sampling and analysis units, 5) design, and 6) analyses. We will highlight notable methodological innovations and offer prescriptive recommendations on how each aspect could be strengthened.

References

- Borman, G., Hewes, G. M., Overman, L. T., and Brown, S. (2003). Comprehensive school reform and achievement: a meta-analysis. *Review of Educational Research*, 73(2): 125-230.
- Mosteller, F., Nave, B., and Miech, E. J. (2004). Why we need a structured abstract in education research. *Educational Researcher*, 33 (1): 29-34.
- Langbein, L. I. (in press). *Statistical guide to program evaluation*.
- Lipsey, M. W., and Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage Publications.
- Zhang, X., and Wang, L. (2003). *National Science Foundation local systemic change: Evaluation of data collection and analysis for cohorts 1-4 project on student achievement* (prepared under contract to the National Science Foundation). Rockville, MD: Westat.