Local Systemic Change *through* Teacher Enhancement

A Summary of Project Efforts to Examine the Impact of the LSC on Student Achievement

by

Eric Banilower

October 2000

Prepared for: The National Science Foundation
4201 Wilson Boulevard
Arlington, VA 22230

Prepared by: Horizon Research, Inc.
326 Cloister Court
Chapel Hill, NC 27514-2296

Introduction

In February 2000, HRI surveyed the PIs of the LSC projects to ascertain whether they had undertaken any studies examining the impact of the LSC on student achievement. Forty-seven of the 68 projects responded. Of these, 12 indicated that they had no student achievement studies. HRI then contacted the remaining 35 projects for information about their study design, instrumentation, and results. Twenty-seven of the projects participated in the interviews between April and June 2000. The interviews revealed twelve projects had completed or nearly finished studies, four had begun studies, and six were in the planning stage of their studies. The other five projects interviewed did not have student achievement studies. A summary of the data collection process is shown in Figure 1.

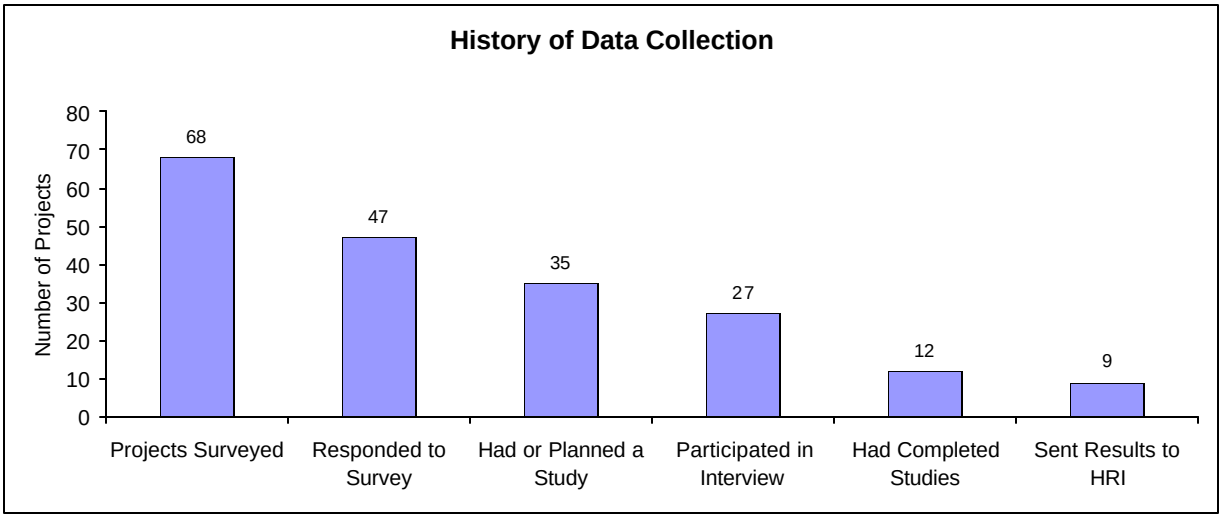

Figure 1

This report analyzes individually the nine completed, or nearly completed, studies HRI was able to obtain, and then attempts to draw some conclusions across all of them. It is important to note that many of the studies reported only group means and did not statistically test group differences. Without information regarding the variance of group scores (i.e., standard errors or standard deviations), it was impossible for HRI to statistically test these differences or to estimate the magnitude of any differences for these projects. When possible, information on effect sizes[1] and the results of statistical tests are included.[2]

One of the key issues to consider when analyzing these studies is their internal validity. In other words, how strong a case do the authors build that any impacts (i.e., increases in student achievement) are attributable to the treatment variable (i.e., participation in the LSC)? Two of the main factors contributing to a study's validity are how well the study's methodology controls for extraneous variables (e.g., initial ability level or school tracking policies) and the level of bias in sample selection (e.g., only teachers of advanced students are in the experimental group). A

[1] When comparing percents, the effect size is calculated using the difference between the arcsine transformation of the percents of the two groups. For means, the effect size is calculated as the difference between the group means, divided by the standard deviation of the population. Following standard conventions, effect sizes of .2 are considered small effects, .5 medium effects, and .8 large effects (Jacob Cohen, *Statistical Power Analysis for the Behavior Sciences*, Hillsdale, NJ: Lawrence Erlbaum Associates, 1988).

[2] Statistically significant differences ($p \leq .05$) are noted with an asterisk.

solid study should be designed to rule out plausible rival hypotheses that could explain any differences found equally well as the study's research hypothesis. To help the reader weigh the results, this report point outs any major threats to internal validity in each of the studies.

Mathematics Studies

HRI was able to obtain reports or summaries of results from five mathematics projects, including one project which sent results from studies done independently by five participating schools. Overall, the quality of the studies is mixed; while most of the mathematics studies had notable threats to internal validity, a couple took steps to reduce these threats and strengthen their case that the LSC is attributable for gains in student achievement. One of the most common weaknesses of these studies was not controlling for initial differences between treatment and control groups. The exceptions were the Project 4 study and school #5 in Project 3.

In general, the studies appear to show positive impacts of the LSC on students' mathematics achievement. However, results need to be interpreted with caution since in most cases it is difficult to make the case that the impact is due primarily to the LSC and not to other, unmeasured, interventions or policies.

Project 1 (K-12 Mathematics)

At this time, the project has compared project wide results for 4^{th}, 8^{th}, and 10^{th} graders from 1999 to 1998 using performance assessment items developed by the Balanced Assessment Project. Over 1000 students per grade level were tested each year. Sixteen items were repeated on both years' assessments: six 4^{th} grade items, six 8^{th} grade items, and four 10^{th} grade items. The study found significant differences on nine of the sixteen items. As can be seen in Table 1, students in 1999 scored higher than students in 1998 on six of the items (three at 4^{th} grade, two at 8^{th} grade, and one at 10^{th} grade) and scored lower on 3 items (all at the 8^{th} grade level). The author points out that one should not read too much into these results as they are small changes, that different students were tested each year, and there was no control for initial differences in students' ability levels.

Table 1
1998-1999 Comparison of Student Performance Scores

Item	Grade	Max Score	1998 Mean	1999 Mean	Difference	Effect Size
Halve It	4	15	4.82	6.05	1.23*	.31
Block Towers	4	5	1.79	2.34	0.55*	.26
Toothpick Squares	4	5	2.51	2.92	0.41*	.24
Tim's Number	4	15	4.08	4.05	-0.03	
Favorite Sports	4	5	2.00	1.97	-0.03	
Pears and Bananas	4	5	2.22	2.08	-0.14	
Toothpick Squares	8	5	1.65	2.94	1.29*	.87
Leisure Center	8	15	5.26	5.72	0.46*	.18
Take a Cube	8	5	2.81	2.33	-0.48*	-.25
Pam's Number	8	15	7.93	5.91	-2.02*	-.44
Building Units	8	5	3.37	2.75	-0.62*	-.52
Metro	8	5	2.12	2.18	0.06	
Calendar Patterns	10	15	3.81	4.42	0.61*	.20
Number Grids	10	15	3.78	4.05	0.27	
Swimming Race	10	5	1.82	1.75	-0.07	
Bottle	10	5	1.14	0.89	-0.25	

Project 2 (K-8 Mathematics)

The project looked at the percent of students at or above the national norm on the ITBS, comparing 1999 data to each school's baseline year (the year before they became involved in the LSC). Results show increases for most of the schools involved with the LSC (see Table 2). However, there are no comparable data shown for non-LSC schools making it difficult to attribute these increases to the LSC. It is possible that ITBS scores were higher over this time period across the entire city due to factors unrelated to the LSC (e.g., familiarity with the assessment, or district retention policies).

Table 2
Percent of Students at or above National Norm on the ITBS

School	Baseline	1999	Difference	Effect Size
21	18.6	52.0	33.4*	0.72
29	31.9	65.1	33.2*	0.68
24	9.1	27.7	18.6*	0.50
26	51.9	73.4	21.5*	0.45
23	21.8	42.4	20.6*	0.45
9	41.2	62.7	21.5*	0.43
20	18.3	37.1	18.8*	0.43
15	17.9	37.0	19.1*	0.43
6	48.6	68.9	20.3*	0.42
30	18.7	36.1	17.4*	0.39
22	18.4	34.8	16.4*	0.38
5	56.0	73.5	17.5*	0.37
12	14.7	30.0	15.3*	0.37
3	12.0	25.1	13.1*	0.34
1	43.2	59.7	16.5*	0.33
2	51.4	66.5	15.1*	0.31
10	32.1	47.2	15.1*	0.31
17	29.9	44.9	15.0*	0.31
8	15.1	27.8	12.7*	0.31
18	14.9	27.4	12.5*	0.31
16	38.8	53.4	14.6*	0.29
4	38.0	51.8	13.8*	0.28
25	21.1	32.1	11.0*	0.25
13	15.3	22.7	7.4*	0.19
14	29.2	37.8	8.6*	0.18
27	16.0	23.3	7.3	
28	34.1	40.0	5.9	
19	96.6	100.0	3.4	
11	13.3	16.2	2.9	
7	65.6	67.0	1.4	

Project 3 (6-12 Mathematics)

Serving schools in several states in its geographic region, the project finds it difficult to collect common data from the schools it serves. Thus, the project has encouraged each school in their LSC to undertake its own study. The project provided results from five of the thirty-three participating schools[3]; each compared students in classes using LSC designated materials (IMP) to students in traditional mathematics classes. The schools used a variety of instruments including the mathematics portion of the Scholastic Achievement Test (SAT), a state mandated mathematics assessment, the New Standards Reference Exam (NSRE), and the Terra Nova. One school also used a self-developed problem-solving test.

The first school categorized students into four groups, advanced IMP, regular IMP, advanced traditional, and regular traditional, though there was no information given as to how the advanced/regular distinction was made. While this distinction may have been intended to control for initial differences in student ability levels, not enough information is provided to judge

[3] HRI does not know whether these were the only schools conducting studies or if the project chose to send only these results to HRI.

whether this goal was accomplished. The school found that, at both levels, students in IMP classes had higher mathematics SAT scores than students in non-IMP classes (see Table 3).

Table 3
Mathematics SAT Scores by Class Type

	IMP	Non-IMP
Advanced	635	595
Regular	505	409

The study also shows that IMP students had higher verbal SAT scores than non-IMP students (see Table 4). While the school may have provided these data to show impact on verbal SAT (given the writing intensive nature of IMP), it raises the question about initial differences between the two groups of students, as there was only a marginal control for initial differences in student abilities. No significance testing was reported.

Table 4
Verbal SAT Scores by Class Type

	IMP	Non-IMP
Advanced	640	580
Regular	510	445

The second school compared IMP and traditional students using the mathematics portion of the SAT and the state mathematics assessment. As can be seen in Table 5, students in IMP classes outperformed students in basic and standard level mathematics courses, but not those in honors courses (the IMP students were roughly 75% standard level students and 12.5 % each of basic and honors, though no information was provided as to how this categorization was made). They also showed that IMP juniors outperformed non-IMP juniors on a school-developed problem solving test (an average score of 3.4 out of 5 for IMP students compared to 2.5 for non-IMP students). No significance testing was reported.

Table 5
Student Performance by Track

Student Track	Mathematics SAT	Percent Proficient or Advanced on State Assessment
Basic	438	2
Standard	554	29
Honors	657	94
IMP	577	56

The next school administered the Terra Nova test to all 9th graders in the spring of 1996 (the end of the first year of IMP implementation). The trend was for IMP students to score higher than non-IMP students in all nine strands, with the differences statistically significant in five of the nine. There were no controls for initial differences between the two groups of students.

Table 6
Terra Nova Results by Class Type

Strand	Non-IMP	IMP	Difference
Data Analysis, Statistics, and Probability	53.63	64.57	10.94*
Number and Number Relations	61.19	70.86	9.67*
Measurement	37.03	46.57	9.54*
Geometry and Spatial Sense	38.42	47.64	9.22*
Computation and Estimation	41.36	49.76	8.40*
Patterns, Functions, and Algebra	28.70	34.43	5.73
Problem Solving and Reasoning	50.41	55.05	4.64
Writing Strategies	63.25	66.24	2.99
Communication	63.55	66.12	2.57

The fourth school compared mathematics SAT scores of IMP and non-IMP students, grouping the students in two levels – honors and regular. IMP students scored higher at both levels than did non-IMP students (see Table 7). No significance testing was reported and not enough information is provided to judge whether IMP and non-IMP students were initially equivalent.

Table 7
Mathematics SAT Scores by Class Type

	IMP	Non-IMP
Honors	633	603
Regular	479	417

The final school compared IMP students to non-IMP students using scores from the NSRE administered at the 10th grade. It is important to note that, with three exceptions for highly motivated students, the IMP course was offered only to students who had a raw SAT score of 40 or higher (though they could choose a traditional mathematics class). To compensate for initial differences between IMP students (high achievers and self-selected) and non-IMP students, the school constructed three matched-pair samples to use as comparison groups. They then used an ANOVA to show that there were no significant differences on the 8th grade SAT scores among the four groups, allowing them to combine the three control groups into one large control group.

The results of the school's comparisons of IMP and non-IMP students can be seen in Table 8. On the 10th grade NSRE, the 33 IMP students scored significantly higher than the 99 students in the control group on the 10th grade Math Scale Score (a difference of .57 standard deviations). They also found that the IMP students scored higher than the control group on the 10th grade SAT-9 Scale Score (an effect of .54 standard deviations). There were no differences among the groups on the Math Concepts scale, but IMP students scored higher on the Math Skills and Math Problem Solving scales (effect sizes of .75 and .56 respectively).

Table 8
Scores on the NSRE by Class Type

Scale	IMP	Non-IMP	Difference	Effect Size
NSRE 10th Grade Math Scale Score	150.7	145.1	5.6*	.57
NSRE 10th Grade SAT-9 Scale Score	738.8	718.2	20.6*	.54
Percent of Students Meeting or Exceeding Standard in Math Skills (NSRE)	97.0	72.7	14.3*	.75
Percent of Students Meeting or Exceeding Standard in Math Problem Solving (NSRE)	39.4	15.2	24.2*	.56
Percent of Students Meeting or Exceeding Standard in Math Concepts (NSRE)	36.4	24.2	12.2	

Project 4 (K-12 Mathematics)

This study examined the results of the school district's implementation of a standards-based educational system using student achievement scores. The district used two national assessments for this purpose. The first was the New Standards Mathematics Reference Examination for grade 4 that contains sub-scales for skills, concepts, and problem solving. The second assessment used was the Iowa Test of Basic Skills (ITBS), chosen to verify that students using the new reform-oriented curriculum did not suffer in basic skills.

The district first administered the New Standards exam in 1996. Comparing achievement scores from 1996 to 1998, the district showed a significant increase in the percent of students meeting or exceeding the standard in all three areas: skills, concepts, and problem solving. Further, data from the ITBS show a small but significant increase in student achievement, lending evidence to the claim that students experiencing a reform-oriented curriculum do not do worse in basic skills, and may in fact do better than students experiencing a traditional curriculum.

Taking the analysis a step further, the project then compared schools based upon whether they were rated as strong or weak implementers of the mathematics program. Ratings of individual teachers were made by teacher leaders at each site and were based upon 1st through 4th grade teachers' use of the curriculum (Everyday Mathematics) as intended. The three sites classified as weak schools were those where all but one or two teachers in grades 1 through 4 were rated as weak implementers. In order to be classified as a strong implementing school, all 3rd and 4th grade teachers had to be strong implementers (8 schools met this criterion). To protect against initial differences between strong and weak schools, the strong schools were further split into two groups. One contained three strong schools with similar demographics to the weak schools (number of students, percent free/reduced lunch, percent African-American, etc.) and the other contained the remaining strong schools.

Students in both groups of strong implementing schools outperformed those from weak implementing schools on all three sub-scales of the New Standards exam and on the ITBS. Further, while a sizable achievement gap existed between white and African-American students at all schools, both groups in strong implementing schools outperformed their counterparts at weak implementing schools.

Project 5 (6-12 Mathematics)

This project looked at the percent of students passing the state's 8th grade mathematics assessment, comparing average passing rates for districts within the project to the state as a whole. As can be seen in Table 9, districts in the LSC averaged a 10% increase in their pass rates (from 36% to 46%) one year after the implementation of the LSC compared to a 7% increase state-wide (from 59% in 1998 to 66% in 1999). Unfortunately, the study does not build a case as to why or how the LSC is responsible for the larger gains in LSC district. As the LSC districts started with a pass rate well below the state average, it could be argued that the increase is to some extent due to regression to the mean.

Table 9
Percent of Students Passing the State Mathematics Assessment

	1998	1999	Difference
LSC Districts	36	46	10
State Average	59	66	7

Science Studies

HRI was able to obtain the results of studies from four science projects. While there is quite a bit of variability in the quality of the study designs, in general, the science studies did a better job of controlling for threats to internal validity. As with the mathematics studies, the results of the science studies are generally positive.

Project 6 (K-8 Science)

With no state assessment in science, individual districts are given the prerogative to choose when and how to measure student achievement in science. Although they use a variety of instruments, most districts do administer a science assessment at the 4th and 8th grade level. These include the ITBS, CTBS, and SAT/OLSAT. Thus, the project was able to collect student achievement trend data from sixteen districts it serves. However, as nine of the districts are very small, rural districts, the project aggregated their data into one composite district, leaving eight districts (seven of the original sixteen and the one composite district) to be analyzed. Of these, four have baseline data (student test scores from the year prior to LSC implementation) giving them four data points. Three districts have three years of data, and one district has two years of data.

From project records, the average number of professional development hours per school as well as the level of kit-usage (low, medium, and high) for each district were computed. Then, district-wide scores were examined (visually) to see if any patterns emerged. The data, shown in Table 10, reveal no clear trends in the relationship between PD hours and changes in student achievement scores or between kit usage and increased achievement. However, given the inadequacy of the data available (district means of student scores and measures of participation in the LSC), it is highly unlikely that any changes, good or bad, could be detected, much less an argument be made that the LSC was responsible for those changes.

Table 10
School Test Scores and Level of Participation in LSC

School	Grade	Number of Years in Project	Baseline Score	1999 Score	Difference	Average Hours of LSC PD	Level of Kit Use
G	4	5	43.4	54.5	11.1	85	High
E	4	5	67.0	78.0	11.0	107	High
A	4	3	56.4	61.3	4.9	128	High
C	4	5	70.0	74.0	4.0	129	High
D	4	5	61.0	65.0	4.0	70	Low
H	4	2	62.0	62.0	0.0	61	Medium
B	4	3	60.0	59.0	-1.0	125	Medium
F	4	3	65.0	61.1	-3.9	127	Medium
M	8	5	73.0	87.0	14.0	81	Low
L	8	5	57.0	67.0	10.0	73	Low
J	8	3	61.0	62.0	1.0	93	Low
I	8	3	70.1	70.6	0.5	128	Low
O	8	5	52.4	51.4	-1.0	69	Medium
K	8	5	84.0	80.0	-4.0	130	Medium
N	8	3	62.1	56.9	-5.2	16	Low
P	8	2	68.0	60.0	-8.0	53	Low

Project 7 (K-8 Science/Mathematics)

The study, one of the strongest submitted to HRI, made comparisons of students' test scores on the SAT-9 open-ended assessment, grouping students by the number of years they had a LSC trained teacher. The analysis was done separately for two grade levels, 5th and 7th grades with over 1000 students participating at each grade level. Additionally, students' reading test scores (3rd and 4th grade respectively for the two analyses) were used to control for initial differences in student abilities.

As can be seen in Table 11, students at the 5th grade level who had a LSC trained teacher for one or two years outperformed students who had never had a LSC trained teacher by about 3.5 points (.17 standard deviations).

Table 11
Predicted NCE Scores for 5th Grade Students
by Number of Years Teachers had LSC Professional Development

Years	Predicted NCE Score
0	47.59
1	50.35
2	51.28

Table 12 shows that 7th grade students who had LSC trained teachers for two or three years scored about 3 percentage points (.14 standard deviations) higher than 7th graders who had a LSC teacher for one year or less.

Table 12
Predicted NCE Scores for 7th Grade Students
by Number of Years Teachers had LSC Professional Development

Years	Predicted NCE Score
0	57.19
1	57.82
2	60.18
3	59.05

Project 8 (K-8 Science)

This study compared results on the SAT-9 Science Open-ended assessment for two matched-pairs of schools (two schools that had participated in LSC professional development and two schools that had not participated). It is unclear on which variables the control schools were selected. Slightly more than 100 4th grade students were tested at each school; twice as many 5th graders were tested. Participation was defined as having a school-wide average of at least three kit trainings per teacher. Overall, there were few differences detected between students at the LSC treated schools and the non-treated schools, with the two exceptions being at the 4th grade level. One was that 4th grade students at one treated school outperformed the students of the matching school (see Table 13). The other exception was that 4th grade students at both treated schools outperformed the untreated schools' students on the Problem-Solving and Decision-Making sub-scale (see Table 14). There were no differences on the other five sub-scales.

Table 13
4th Grade SAT-9 Scale Scores

School Pair	Treated School	Untreated School	Difference	Effect Size
1	589.14	576.56	12.58*	.38
2	591.52	595.98	-4.46	

Table 14
4th Grade Problem Solving Scale Scores[4]

All Treated 4th Graders	All Untreated 4th Graders	Difference	Effect Size
1.44	1.24	.20*	.22

While this study used a matched sample to control for initial differences in student ability, the relatively small sample sizes reduce the study's chances of detecting differences between the control and experimental groups with this design. Further information regarding how control schools were selected and how initial equivalency of students was determined would strengthen this study.

Project 9 (K-8 Science)

The school district administered the Stanford Achievement Test – 9th edition, Form T to all 4th and 6th grade students. They analyzed only the scores of students who had been enrolled in the district for the past four years (around 630 at each grade level), allowing them to compare students who had and had not been exposed to their LSC science program. Mean percentile

[4] Data for the other sub-scales were not included in the report sent to HRI.

rankings were presented for each grade level. Neither standard deviations nor standard errors were included in the report and no statistical tests were used to compare group means.

The data appear to show, at both grade levels, that students of teachers who participated in the LSC professional development and used the LSC designated instructional materials during the 1998-99 school year scored higher on the SAT-9 than did students of teachers who did not participate. The data were then further disaggregated by the number of years (from zero to four) students had teachers who participated in the district's LSC. The data appear to show a stair-step increase in student performance on the SAT-9. As can be seen in Table 15, the mean score increases with each additional year of having a LSC trained teacher.

Table 15
SAT-9 National Percentile Rankings
by Years of Student Participation in LSC Science Program

Years	Grade 4	Grade 6
0	21	27
1	32	32
2	38	42
3	47	50
4	53	64

The project also examined pass rates on the 6th grade writing proficiency test with the hypothesis that the writing-intensive nature of the science program would improve students' writing abilities. Student scores were disaggregated in the same two ways as with the science scores, with similar results found (see Table 16). Students of teachers who participated in the LSC during the 1998-99 school year had a higher passing rate than students of teachers not participating. Further, a similar stair-step pattern emerges, as with the science data, when the data are broken down by the number of years the students had a LSC participating teacher, although there is no difference in pass rates for students in the 3 and 4 years groups. HRI was able to run significance tests on these differences[5] and found that students of participating teachers did pass the writing proficiency test at a higher rate than did students of non-participating teachers.

Table 16
Grade 6 Writing Proficiency Pass Rate
by Years of Student Participation in LSC Science Program

Years	Percent Passing
0	23
1	68
2	71
3	90
4	89

While these data appear promising, there are some dangers to drawing conclusions regarding the impact of the LSC in the district. First, with no standard deviations reported, it is impossible to

[5] To statistically test for differences in percent passing the writing test, all that is required is the number in each group and the percent passing, both of which were provided in the report sent to HRI. To test for differences in the mean national percentile rankings, the standard deviation or standard error of the mean is required. This information was not included in the project's report.

judge the magnitude of the differences. Second, questions remain regarding how teachers were selected to participate in the LSC's professional development. Were schools targeted on a cohort basis? If so, were the original targeted schools high performing than schools targeted in later years? Or were participating teachers volunteers, and perhaps more enthusiastic about teaching science? Finally, while the results of the writing proficiency test are reported to show a crossover effect, such assessments are commonly used as ability measures to control for initial differences. Hence, the trend of higher SAT-9 scores with increased years of participation could be explained by initial differences in student ability levels as measured by the writing proficiency test.

Conclusions and Recommendations

Given the limited information provided in many of these studies, one must interpret their results with extreme care. Table 17 summarizes the results of these studies and provides a rough measure of each study's internal validity. At first glance, it appears that in both mathematics and science, the LSCs are having a positive impact on student achievement. All of the mathematics and three of the four science projects show increases in student performance. However, many of the studies do not present enough information to build a convincing case that the LSC was responsible for improved student achievement. Because of this, it impossible to judge with any certainty whether the results from these studies are real or spurious, due to factors other than the LSC or perhaps simply artifacts of the study's methodology. The most common threats to internal validity in these studies were:

- Lack of a control group – for example, the study reported gain scores for schools in the LSC, but not for schools outside of the LSC.
- Failure to account for initial differences between control and experimental groups – while the study may have reported that LSC students scored higher than non-LSC students, it was unclear as to whether the two groups started at the same achievement level.
- Sample selection bias – the study did not address how teachers were selected for participation in LSC training and whether this may have affected the study's results.

Table 17
Results of Student Achievement Studies

Project	Direction/ Magnitude	Internal Validity[6]
Mathematics		
Project 1	⇧	Indeterminate
Project 2	⇧	Indeterminate
Project 3		
School #1	⇧	Indeterminate
School #2	⇧	Indeterminate
School #3	⇧	Indeterminate
School #4	⇧	Indeterminate
School #5	⇧⇧	Strong
Project 4	⇧⇧	Solid
Project 5	⇧	Indeterminate
Science		
Project 6	⇔	Indeterminate
Project 7	⇧	Strong
Project 8	⇧	Solid
Project 9	⇧	Solid[7]

While it is impossible to generalize to the LSC program as a whole given the small number of studies made available to HRI, it is encouraging that all five studies rated as solid or strong found positive impacts on student achievement. Each of these five studies makes a defensible case that the gains are attributable to their LSC. As more of the individual LSCs undertake and complete studies of quality comparable to these, the stronger the case can be made as to the LSC program's impact on students.

Given that 8 out of the 13 studies, either through omission of data or poor research design, did not present enough evidence to make their results credible, NSF may want to consider offering the LSCs additional support for conducting studies of student outcomes. For projects that have research and evaluation experts on staff, a set of criteria or guidelines that communicate NSF's information needs should be sufficient. Other projects will require some form of technical assistance, ranging from small doses to help refine research plans to extensive assistance in design research studies and analyzing data. NSF may want to consider offering a conference to help projects articulate their information needs, raise their awareness of key issues in research design, and become savvier consumers of technical assistance.

[6] A strong study controls for most threats to internal validity and provides enough evidence for the results to be compelling. A solid study controls for many threats to internal validity, and although some flaws in methodology or analysis remain, the results are credible. Studies categorized as "indeterminate" did not provide enough information to make a persuasive argument as to the credibility of the results.

[7] Project 9 provided HRI with a preliminary write-up of their results. While the study design appears solid, HRI does not have enough information to make a fully-informed judgement on the study's validity.