

# National Science Foundation Local Systemic Change

Evaluation of Data Collection and Analysis for  
Cohort 1-4 Projects on Student Achievement

Prepared by:

Xiaodong Zhang  
Lawrence Wang

November 2002

Prepared for:

Horizon Research, Inc.  
326 Cloister Court  
Chapel Hill, NC 27514

Prepared by:

**WESTAT**  
1650 Research Boulevard  
Rockville, Maryland 20850-3195

NOTE: The views, findings, conclusions and recommendations expressed in this report are those of the authors and do not necessarily represent the official views, opinions, or policy of the National Science Foundation.

# Table of Contents

Chapter		Page
	Acknowledgments .....	v
	Executive Summary .....	vii
1	Introduction.....	1
2	Literature Review .....	3
3	Review of Studies of LSC Impact on Student Outcomes .....	5
	3.1 Teacher Enhancement for Student Success (California, Cohort 2: Science/Math K-8).....	7
	3.2 Elementary Science Education Partners (Georgia, Cohort 2: Science K-8) .....	8
	3.3 Partnership for Systemic Change (New Jersey, Cohort 2: Science/Math K-8).....	9
	3.4 Asset Inc. Teacher Enhancement (Pennsylvania, Cohort 2: Science K-8).....	10
	3.5 Mathematics Renaissance K-12 (California, Cohort 3: Math K-12) .	11
	3.6 Science Connections (Maryland, Cohort 3: Science K-8) .....	11
	3.7 Minneapolis Public Schools Systemic Change in Science Initiative (Minnesota, Cohort 3: Science K-8) .....	12
	3.8 Pittsburgh Reform in Mathematics Education (Pennsylvania, Cohort 3: Math K-12) .....	13
	3.9 Reconceptualizing Mathematics Teaching and Learning Through Professional Development (New York, Cohort 4: Math K-8) .....	14
	3.10 Greater Philadelphia Secondary Mathematics Project (Pennsylvania, Cohort 4: Math 6-12) .....	15
	3.11 Austin Collaborative for Mathematics Education (Texas; Cohort 4: Math 6-12) .....	16
	3.12 Valle Imperial-Project in Science (California, Cohort 4: Science K-8).....	16
4	Analysis of Inquiry-Based Science in Seattle Public Schools (Cohort 2) .....	19
	4.1 Background.....	19
	4.2 Data.....	19
	4.3 Research Design and Analysis.....	20
	4.4 Findings .....	22

## Table of Contents (continued)

<b>Chapter</b>		<b>Page</b>
5	Conclusions and Recommendations .....	25
	5.1 What Have We Learned About LSC Impact on Student Outcomes? .	25
	5.2 What Can We Learn About the Common Issues? .....	26
	5.3 What Strategies Can We Suggest to Improve Future Program Evaluations?.....	27
	References.....	31

### List of Appendices

<b>Appendix</b>		
A	Rating Criteria for Evaluation of Studies.....	A-1
B	Statistical Procedures and Analyses for Seattle Public Schools (SPS) .....	B-1

### List of Tables

<b>Table</b>		<b>Page</b>
1	Overview of the reviewed LSC projects .....	6
2	Summary of the reviewed LSC student achievement studies in data collection and analysis.....	6
3	Summary of the reviewed LSC studies using other student outcomes in data collection and analysis .....	7
4	Ratings of the reviewed studies .....	17
5	Variable list in the SPS data.....	21

### List of Figures

<b>Figures</b>		<b>Page</b>
1	Logic model for professional development .....	3
2	Changes of ITBS science scores in SPS (1996-2002) .....	21
3	Gross versus net income .....	29

## ACKNOWLEDGEMENTS

---

We would like to acknowledge the invaluable assistance from our colleagues in the preparation of this report. Joy Frechtling, Associate Director of Education Studies, has provided the intellectual guidance throughout all stages of the project. Joan Michie was responsible for the earlier phase of the study and collection of the reports. For the analysis for the Seattle Public Schools, Philip Fletcher advised us on the study design. Bill Scheig facilitated data file transformation and offered technical support. Carol Litman was our technical editor, and Sylvie Warren provided desktop publishing service.

We want to thank all of the LSC principal investigators and their designates for taking the time to provide us with their insights on the programs. We owe enormous debts to projects that shared their evaluation reports, and especially staff of the Seattle Public Schools for providing us with the opportunity to conduct a study based on their data. In particular, Elaine Woo, the principal investigator of the Seattle program, has offered remarkable insights that helped us understand the scope and challenges of the program, build our models, and interpret our findings. Jody Schultz provided us with the data and timely feedback on our questions about the data and test instruments.

# EXECUTIVE SUMMARY

---

This report presents findings of project efforts in data collection and analysis on student achievement under National Science Foundation's (NSF) Local Systemic Change Program (LSC). Commissioned by Horizon Research and conducted by Westat, the evaluation is designed to examine the status of project evaluations in Cohorts 1-4 undertaken before new requirements regarding outcome evaluation established.

Westat has performed thorough evaluation of 12 studies conducted by LSCs projects. In addition, we examined the impact of school and teacher participation in the LSC program on student achievement, based on data from Seattle Public Schools. The results from the reviewed analyses and our own study present evidence that LSC programs have generated positive impact on student outcomes.

- Ten of the 12 studies reviewed suggest that the LSC programs lead to better student outcomes in the target areas.
- Studies disagree on the substantive effect of the program. While some studies report substantial program impact on student outcomes, others find that the program has contributed to a modest increase in student learning, not up to the desired level. In addition to school participation, a few studies show that the length of teacher participation in professional development activities has a direct effect on student outcomes.
- Studies also draw different conclusions in terms of the impact on subpopulations. Whereas some detect a significant program effect on minority students, others argue that the impact is insignificant.
- It appears that programs that promote standards-based teaching often result not only in improvement in the target subject area, but also have a spillover effect in other areas such as reading, especially in sites with a high percentage of minority students.

- The impact analysis is highly sensitive to the outcome measures. It appears that the program is more likely to show effects measured by criterion-referenced tests, especially those that are well-aligned with the program goals, than norm-referenced tests.
- Finally, getting implementation right is a prerequisite to outcome evaluation. This need is often overlooked in impact analyses. A few projects have recognized the importance of documenting implementation and have demonstrated that well-implemented programs will lead to better student learning.

All of the studies are analyzed and rated by Westat based on their adequacy with regard to the following elements: instrument selection, design, data collection, data analysis, and reporting. The majority of studies are rated as satisfactory, with an average score of 3.7 out of 5 points. Five studies are considered to be good quality, and they may serve as examples for future evaluation efforts. The overall quality of the existing LSC project evaluations is found to be better than similar studies in professional development, identified in a previous synthesis by Westat (Frechtling et al., 1995).

In addition, the analysis underscores common features and issues among LSC projects.

- Among 57 projects, 30 percent (19) have conducted analyses/reports regarding program impact on student outcomes.
- Twelve of these studies met what we feel are at least minimal standards for scientific rigor, employing clearly specified procedures of data collection and analysis according to the established principles.
- Most studies analyzed individual student-level data. Student outcomes from grades 3-10 were assessed most frequently, with the exception of the 6th grade.

- A variety of instruments were used to measure student achievement such as commercial standardized tests, state assessment tests, as well as measurements for student attitudes, motivation, and engagement. Many of the studies took the effort to demonstrate the program effect on student assessments that are well aligned with the LSC goals.
- Studies involved either quasi-experimental design or nonexperimental design. No randomized experimental design was conducted.<sup>i</sup> Existing studies used a variety of designs such as pre- and post-test comparisons, post-test comparisons, and post-

test time series analysis, many of which dealt successfully with the constraints of data and resources.

- Studies were most likely to use t-tests and/or ANOVA as inferential statistical procedures and only a few analyses involved more sophisticated methods.

The study concludes by emphasizing the needs for better data collection and analysis in the future. Specific strategies are recommended in study design and analysis to cope with data availability, analytical capacities, and resource constraints.

---

<sup>i</sup>A “randomized experiment” is defined as an experiment where subjects in the treatment group and control group are randomly - assigned. A “quasi-experiment” does not have random assignment. However, comparison group is created to match the characteristics of the treatment group. A “non-experiment” does not involve a control/comparison group. Rather, statistical control is used to isolate the confounding factors.

# 1. INTRODUCTION

---

During the 1990s, NSF funded a number of large-scale initiatives designed to improve science, mathematics, and technology teaching and learning at the K-12 grade levels. These efforts aimed to align all aspects of the educational system in support of curriculum and performance standards.

The Local Systemic Change (LSC) program was created in 1995 to improve teaching and learning in these areas by focusing on professional development of teachers within schools or school districts. The underlying rationale was that professional development would improve teachers' skills, confidence, and knowledge, thereby developing the capacity of schools to deliver quality instruction that leads to better student outcomes. Projects were expected to designate instructional materials and then to provide extensive professional development to help teachers strengthen their content knowledge and become skilled in the use of the specified instructional strategies. In addition, projects were designed to provide support for teachers as they implemented the instructional materials in their classrooms. These projects varied in the relative emphasis on content knowledge, the required hours of professional development, and the extent to which they provided professional development districtwide rather than at the school site. All projects were funded for 5 years.

Until a few years ago, the LSCs had been assessed primarily in terms of the quality of professional development and its effects on teaching. As a result, there was little concrete evidence among earlier projects (Cohorts 1-4) that LSC efforts had improved student learning. This is not unique to LSC projects. Kennedy (1998) found that only a few studies on teacher professional development included measures of student learning. However, as the nation embraces the notion of accountability in educational reforms, NSF and other federal agencies are becoming increasingly concerned about student achievement. In fact, the absence of good evaluation has led some policymakers to

express skepticism about the value of these programs (Fox, 1998). Stakeholders at all levels want to know about student outcomes. At the local level, parents and the community closely watch how well public schools are meeting students' needs; at the state level, decisionmakers want to know how well school systems are operating; and at the federal level, policymakers closely monitor the nation's accomplishment in education. Building support for reform effects rest strongly on showing that investments pay off in improving what students know and can do (Frechtling, 2002).

It is within this policy context that federal agencies have placed greater emphasis on collecting and analyzing data on student learning. NSF has identified student impact as one of the critical indicators of the success of its programs and therefore asked projects to provide such evidence. Individual projects are now being asked to assess the effectiveness of the program on students. In other words, when teachers are provided extensive professional development around the use of high-quality instructional materials, do their students learn more? If the LSC does not lead to improved student learning, it will be difficult to make the case that the program should be continued. If, on the other hand, different studies using a variety of outcome measures demonstrate the effectiveness of LSC program on student achievement, there will be good reasons to replicate its major feature.

In 2002, Horizon Research, Inc, NSF's contractor to evaluate the LSC program, asked Westat to examine the status of evaluations undertaken by all Cohort 1-4 projects before new requirements regarding outcome evaluation were established. The analysis was to provide answers to the following questions:

- What can we learn from these studies about the impacts of LSC programs?
- What can we learn about the common issues faced by such studies?

- What strategies can we suggest to improve future program evaluations?

This report contains five chapters. Following the introduction, Chapter 2 is a review of the current literature and studies on teacher professional development, with special attention to the approaches advocated by the LSC program. Chapter 3 summarizes the status of data on student outcomes among LSC projects (Cohorts 1-4). It also presents a detailed review of 12 impact analyses conducted by the LSC projects. Following a brief summary of research findings,

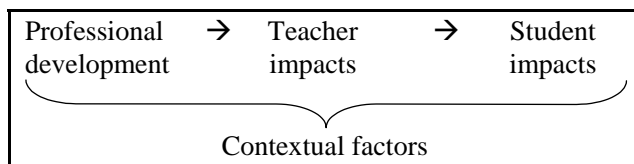
each study is examined regarding its performance in instrument selection, design, sampling, analysis, and reporting. In order to provide technical assistance and demonstrate methods of evaluation, Westat helped the Seattle Public Schools to conduct an impact analysis of its LSC program, and that effort is described in Chapter 4. Based on the results from the previous chapters, Chapter 5 summarizes findings on the LSC projects, discusses common issues underlying the existing evaluation efforts, and recommends strategies to improve future NSF program evaluation.

## 2. LITERATURE REVIEW

---

Understanding the process by which teacher professional development affects students' outcomes (i.e., knowledge, attitudes, and behavior) is a topic of considerable interest to educational researchers and policymakers. Teachers' skills and knowledge have frequently been identified as among the most critical determiners of student learning. Darling-Hammond and Ball (1998) found that teacher quality accounts for about 40 percent of the variation in student achievement. Darling-Hammond and Rustique-Forrester (1997) argued that monies spent in improving teachers' qualifications net greater gains in student learning than other educational expenditures. Improving teacher quality is a cornerstone of President Bush's No Child Left Behind legislation, and many attempts to increase the supply of high-quality teachers have focused on professional development. Most existing studies of professional development are based on the following logic model:

**Figure 1.**  
**Logic model for professional development**



Considerable attention has been given to the nature of the professional development experience itself. The main debate focuses on the relative importance of content and pedagogy and how to balance them. Professional development has been presumed to benefit teachers in attitude and leadership, as well as content knowledge and pedagogical practices. Teacher participation, in turn, is expected to affect students in terms of subject matter knowledge, attitudes, course participation, and career choices. Contextual factors, i.e., characteristics of students and teachers who are involved and the organizational and reform context in which change is taken place,

have received the least attention of any components of the model.

There is a growing consensus among experts that professional development is the cornerstone of educational reform and instructional improvement (Elmore, 2002). Professional development is expected to serve multiple purposes: (1) as a tool for introducing teachers to new practices or information; (2) as a means for reviewing or refreshing existing skills; and (3) as a way to fill the gap left by previous educational experiences. It is especially important for math and science teachers, as many of them, especially those teaching in the earlier grades, come to the classroom with relatively little training in these disciplines. In addition, professional development has become a key component of today's efforts toward educational improvement and systemic reform. Policymakers have consistently made professional development a critical part of the change process.

A comprehensive review of studies on professional development by Frechtling and associates (1995) including a total of 50 studies, suggests a somewhat mixed result regarding the impact of professional development on student outcomes. The overall quality of existing research was found to be unsatisfactory, with less than half of the studies reviewed employing any rigorous and scientifically based methods. Nevertheless, these studies provided suggestive evidence that professional development improves student learning.

After examining results from a national sample of over 1,000 mathematics and science teachers, Garet et al. (2001) identified three features of professional development activities that have significant effects on teachers' self-reported increases in knowledge and skills and subsequent changes in classroom practice: (1) focus on content knowledge, (2) opportunity for active learning, and (3) coherence with other learning activities. Cohen and Hill (1998) found that

students benefit most when teachers' learning opportunities are grounded in the curriculum and extended in time. Weathersby and Harkreader (1999) identified 26 outstanding staff development programs in the core content areas of language arts, mathematics, science, social studies, and interdisciplinary programs for the National Staff Development Council. Their study found that most of those programs often last for several weeks with followup activities throughout the year, rather than being a one-time workshop. Virtually all of the programs emphasized a mixture of increasing teachers' content knowledge and pedagogical skills. Resnick and Harwell (1998) suggested that the quality of professional development in schools is positively related to achievement of students.

Among the approaches advocated by NSF, low-achieving elementary students who were placed in

problem-solving or peer-collaboration situations were found to achieve higher math scores and report higher levels of motivation than students who received neither interventions (Ginsburg-Block and Fantuzzo, 1998). Several other studies have demonstrated the value of peer tutoring and collaboration (Fantuzzo, King, and Heller 1992; Greenwood, Carta, and Hall 1988) as well as the benefits of contextualizing instruction in real-world problems (Verschaffel and De Corte 1997). Cohen and Hill (1998) noted that the frequent use of these approaches is positively related to scores on the California Learning Assessment System Math Test at the school level after controlling for demographic data. Mayer (1998) found a small positive relationship between a similar set of practices and student scores on a standardized multiple-choice test.

### 3. REVIEW OF STUDIES OF LSC IMPACT ON STUDENT OUTCOMES

---

In 2002, Westat contacted the principal investigators of all LSC projects in Cohorts 1-4 to inquire about the status of data collection and evaluation on student outcomes. Although those projects were not required to conduct such analyses, many projects chose to address this issue.

By August 31, 2002, 56 of the 57 projects responded. Twenty-six percent (14) indicated that they did not have data on student outcomes in the target subjects, with the data lacking more frequently in science than mathematics. Thirty percent (19) of the projects had conducted some form of analysis and reporting based on student outcome data, while the remaining 44 percent had not analyzed the data, were in the process of analysis, or were planning an analysis.

Although student learning was identified as an important goal in many proposals, the projects in Cohorts 1-4 had been assessed primarily in terms of the quality of professional development and its effect on teaching. Therefore, the 19 studies on student outcomes were entirely local initiatives. While many other projects indicated that they were interested in analysis of student outcomes, they have been unable to do so for various logistical, administrative and philosophical reasons.

- They had no achievement scores in science.
- Students' scores could not be linked to the amount of teacher professional development.
- Although projects had data on student achievement and teacher professional development, administrators were not sure how much time would be involved trying to link them.
- Data could not be disaggregated by teachers because of union concerns.

- The implementation model changed in the course of the project.
- State tests changed several times over the course of implementation.
- The test was not reflective of what had been done and was not aligned with new curriculum.
- The project had many problems with district support—just implementing it was a challenge.

Westat conducted a thorough analysis of the existing 19 studies from LSC projects. The following report provides a summary of 12 studies that involved at least minimally rigorous procedures in data collection and analysis. The remaining seven studies generally report data without using any statistical methods. Table 1 provides a brief overview of the reviewed studies.

Based on the established principles of program evaluation, the 12 studies were reviewed according to their performance in 5 procedures: instrument selection, design, data collection, analysis, and reporting.

- Criteria to judge the adequacy of an instrument included not only its reliability, but also its alignment with the LSC goals.
- The analysis explored different design issues. While a pre- and post-test design with comparison group was often the preferred approach, we also looked for innovative and useful ways to cope with data limitations that prevent an evaluator from using such approach. In addition, we evaluated the extent to which mixed-methods were incorporated.
- Sampling techniques were analyzed primarily to determine whether or not there was evidence of sampling bias in data collection.

- Studies were judged by the level of appropriateness and sophistication in using statistical procedures in quantitative analysis.
- Reporting was evaluated in terms of how studies were presented and how technical and practical concerns were addressed.

The detailed rating criteria are explained in Appendix A. The rating involves not only whether the study used a particular procedure, but also

judgment of how effectively the procedure was used and whether its choice was appropriate for the status of data. Table 2 and Table 3 present summaries of data collection and analysis efforts for the 12 projects. Table 2 summarizes student achievement studies whereas Table 3 presents studies that using other student outcomes such as student attitude, engagement assessment as well as course enrollment and success rates. Three studies examine impact on both student achievement and other outcomes and are listed in both tables.

**Table 1.**  
**Overview of the reviewed LSC projects**

Project name	Acronym	Cohort	State	Subject
Teacher Enhancement for Student Success .....	TESS	2	CA	S/M (K-8)
Elementary Science Education Partners .....	ESEP	2	GA	S (K-8)
Partnership for Systemic Change.....	PSC	2	NJ	S/M (K-8)
Asset Inc. Teacher Enhancement.....	ATE	2	PA	S (K-8)
Mathematics Renaissance K-12.....	MRS	3	CA	M (K-12)
Science Connections .....	SC	3	MD	S (K-8)
Minneapolis Public Schools Systemic Change in Science Initiative .....	Mosaic	3	MN	S (K-8)
PRIME: Pittsburg Reform in Mathematics Education .....	PRME	3	PA	M (K-12)
Reconceptualizing Mathematics Teaching and Learning Through Professional Development.....	RMTLTPD	4	NY	M (K-8)
Greater Philadelphia Secondary Mathematics Project .....	GPSMP	4	PA	M (6-12)
Austin Collaborative for Mathematics Education .....	ACME	4	TX	M (K-8)
Valle Imperial - Project in Science .....	VIPS	4	CA	S (K-8)

**Table 2.**  
**Summary of the reviewed LSC student achievement studies in data collection and analysis**

Project	Level	Sample	Instrument	Alignment	Grade	Design	Method	Finding	Effect size*
TESS	District	NA	SAT-9.,CTSA, SEAS	Yes	K-8	Post-test Non-exp	t-test	Somewhat Positive	NA
PSC	Student	NA	SAT-9, state tests (ESPA, GEPA)	Yes	5, 7th	Post-test Non-exp	t-test	No	NA
ATE	Student District	1,542	MTIMSS	Yes	3,4,7th	Post-test non-exp	t-test	Positive	NA
MRS	Student	NA	MARS	Yes	4,8,10th	Post-test	t-test	Positive	0.07
SC	Student	2,251	CTA	Yes	5,7th	Pre-post test Quasi-exp	ANCOVA	Positive	0.35
Mosaic	Student	20,000	SAT-9,CTB5, CSIAC, Met, state tests	Possible	3,5,7th	Posttest Quasi-exp	Regression	Positive	0.14
PRME	Student	NA	State test (NSMRE), ITBS	Yes	4th	Post-test Non-exp	t-test	Positive	NA
RMTLTPD	Student	1,536	CTBS	No	3-8th	Post-test Non-exp	HLM	No	NA
GPSMP	Student	NA	PSAT	No	NA	Post-test Quasi-exp	t-test	Positive	NA
ACME	Student	550	TAAS, ITBS	Yes	3-8th	Post-test Non-exp	t-test	Positive	NA
VIPS	Student	1,262	SAT-9 (S/M) District proficiency tests	No	4,6th	Post-test non-exp	t-test, ANOVA	Positive	0.5

Note: No studies include effect size. The effect size is calculated based on available statistics using the formula by Wolf (1987).

**Table 3.****Summary of the reviewed LSC studies using other student outcomes in data collection and analysis**

Project	Level	Sample	Instrument	Alignment	Grade	Design	Method	Finding	Effect size*
TESS	District	NA	Course enrollment, success rates	Yes	K-8	Post-test Non-exp	t-test	somewhat Positive	NA
ESEP	Student	955	mATSI	Possible	5th	Pre-post-test Non-exp	MANOVA, ANOVA	Positive	0.25
SC	Student	2,251	Engagement (LE), motivation (MOPO) assessments	Yes	5,7th	Pre-post test Quasi-exp	ANCOVA	Positive	0.35
GPSMP	Student	NA	Attitude survey, Ivy League Exist test	No	NA	Post-test Quasi-exp	t-test	Positive	NA

Note: No studies include effect size. The effect size is calculated based on available statistics using the formula by Wolf (1987).

It is important to note that these reports often were developed to serve purposes other than studies of program impact on student outcomes. However, our ratings focus only on the impact analysis and do not necessarily reflect the quality of the overall reports. For example, several studies were pulled from what were essentially implementation evaluations. A few studies written for journal publication focused on specific aspects of the program (i.e., impact on student attitudes), while others were as project evaluations. Therefore, using one set of criteria to judge the quality of these studies can be quite arbitrary, partial, and somewhat unfair.

The following section presents detailed review of the 12 studies. Each review begins with a summary of ratings based on a scale of 1 to 5. After introducing the main evaluation issue, a brief summary of evaluation follows. Performance on each of the five procedures is then detailed and rated according to the protocol in Appendix A.

### 3.1 Teacher Enhancement for Student Success (California, Cohort 2: Science/Math K-8)

Rating summary (2.6): instrument (4), design (2), sampling (2), analysis (2), reporting (3)

#### Evaluation Summary

The program was designed to reform curricula based on National Council of Teachers of Mathematics (NCTM) and National Research

Council (NRC) content standards and provide teacher professional development. The study examined changes in student outcomes over time at the district level based on multiple criteria. It concluded that the program appeared to generate mixed, but somewhat positive outcomes.

- Comparing student enrollment in core courses in 1993-94 with that in 1998-2001, the evaluator found that the percentage enrolled in selected high school courses increased across all ethnic groups. In contrast, there were few changes in enrollment in advanced coursework.
- Comparing success rates in high school core courses between 1995-96 and 1999-2001, the study showed increases in geometry and algebra II for African Americans, but drops in algebra I.<sup>1</sup>
- The correlation between the length of professional development received by teachers and SAT-9 scores of their students was found to be significant.
- The district performance in the California Content Standards Assessment (CTSA) increased from 1999 to 2001, but did not keep pace with the state overall improvement.
- Whereas the number of students in all ethnic groups taking and receiving 3+ in advanced placement exam improved when comparing 1994 to 2000 and 2001, the number taking University of California A-G course

<sup>1</sup> This might be an indication that more students were taking algebra I in earlier grades.

requirement declined in 2000-2001 from the previous year.

- Half of the grades tested showed improvement in 2001 on the Science Embedded Assessment System (SEAS) exam, which was created to allow elementary schools teachers to evaluate students in specific curricular units.

### Rating of Procedures

- *Instrument (4)*: The strength of the study lies in its use of multiple criteria to analyze program impacts. Among these measures, SEAS seems to be most closely aligned with professional development goals.
- *Design (2)*: The study can be characterized as a nonexperimental design, involving a post-test treatment group only. With the exception of SAT-9, most of the cited program impacts are related to student measures that involve neither comparison groups nor baseline data. The internal validity is somewhat questionable.
- *Sampling (2)*: Although the study used district-level data, sampling can still be problematic because it was not clear whether all of the students took the identified tests.
- *Analysis (2)*: In addition to looking at overall program impact, the study examined program impacts by ethnic groups and in different content areas. However, many conclusions, except those drawn from SAT scores, were drawn by direct comparison of scores without using any statistical procedure.
- *Reporting (3)*: Since the report was designed as an implementation evaluation, it did not go into enough detail to address the sampling and design issues as one would like to.

### 3.2 Elementary Science Education Partners (Georgia, Cohort 2: Science K-8)

Rating summary (3.8): instrument (3), design (3), sampling (5), analysis (4), reporting (4)
---

### Evaluation Summary

The program was based on the assumptions that hands-on, inquiry methods as well as standards-based professional development would benefit all students, especially children of color and girls. The study investigated the program effect on African American students' attitudes toward science at the 5th grade. It concluded that the program was moderately successful in producing positive attitudes toward science.

- Program students showed positive and significant differences on two of five attitude scales (anxiety and desire), but had less positive views about the value of science to society.
- The differences were found most frequently in schools where the principal was supportive of science reform.

### Rating of Procedures

- *Instrument (3)*: The study offers an interesting example of using alternative measures other than test scores to examine program impacts. The evaluators used a modified Attitude Toward Science Inventory (mATSI)<sup>2</sup> to measure changes in students' attitudes. Although the alignment issue was not discussed directly, one may infer that a positive attitude toward science learning is at least one of the program goals.
- *Design (3)*: The evaluators used a pretest-posttest design by comparing 477 students in 7 schools in 2000 with 480 students in 7 schools in 1996. However, they failed to examine whether or not the pre-test and post-test groups were comparable. This approach is problematic. It would be important to compare two groups in terms of variables such as demographics and achievement level to ensure that they are comparable. If not, the differences should be accounted for

---

<sup>2</sup> mATSI is a 25-item, Likert-type instrument made up of five scales: perceptions of the science teacher, anxiety toward science, value of science to society, self-confidence in science, and desire to do science.

statistically or a sub sample taking both the pre- and posttest would be examined. Focus groups were used to understand the school influence on students' attitudes.

- *Sampling (5)*: Sampling of the data was purposive, focusing on schools with primarily African American students. The eligible schools were then selected randomly.
- *Analysis (4)*: The data analysis involved a three-way MANOVA to determine possible differences between schools and gender, as well as a one-way ANOVA on each of the attitudinal scales.
- *Reporting (4)*: While the program's impact on student attitudes is in the desired direction, the conclusions about school impact that were drawn from the focus groups were not well explained.

### 3.3 Partnership for Systemic Change (New Jersey, Cohort 2: Science/Math K-8)<sup>3</sup>

Rating summary (3.6): instrument (4), design (3), sampling (3), analysis (3), reporting (5)

#### Evaluation Summary

The program aimed to provide 100 hours of high-quality professional development in science and mathematics to about 8,000 teachers from four partner districts. Using a variety of instruments, the study attempted to investigate the program impact by comparing the student outcomes over the years, as well as with nontreatment districts. The overall the student performance data did not show significant gains in partnership districts. CPRE argued that the findings did not mean there was no program impact. Instead, the researchers

discussed the serious limitations of the instruments.

- CPRE used SAT-9 open-response science items to examine changes in scores for grade 5 and 7 students since 1996 and found that scores for regular students (excluding special education students) declined slightly in the 5th grade and showed mixed patterns for the 7th graders.
- Data from the first year of the Partnership Performance Assessment (PPA) were collected as a baseline for future analyses. In general, students at 3rd and 7th grade performed well on the assessment. Tests of differences between subgroups found little differences by gender. Asian and white students scored slightly better than African American and Hispanic students. The evaluators also looked at the results on specific tasks in order to detect patterns of students' relative strengths and weakness in specific content areas.<sup>4</sup>
- No correlation was found between SAT-9 and PPA scores.
- The district results on statewide assessment were compared with other jurisdictions in 1999 and 2000. The district performed at the average level in Elementary School Performance Assessment (ESPA) (4th grade), and slightly better than average in Grade 8 Elementary Performance Assessment (GEPA) (8th grade).

#### Rating of Procedures

- *Instrument (4)*: The study used the following instruments to measure student outcomes: SAT-9 open-response science items, criterion-referenced state assessment (ESPA, GEPA),

---

<sup>3</sup> The Merck Institute for Science Education, which also provided funding for the district's professional development, commissioned the Consortium for Policy Research in Education (CPRE) to prepare its evaluation and annual report. This review considers CPRE's 6th-year report, which is primarily an implementation evaluation with some attention to program impacts.

---

<sup>4</sup> For example, on the 7th grade test, students generally handled the process questions with little trouble, but were less familiar with the science content embedded in the assessment. Ninety percent of the students successful described the experiment, but only 57 percent explicitly controlled for a specific aspect of the experiment. Students had little problem recording their experimental data and drawing a conclusion: more than 90 percent of the students were able to complete these components of the assessment.

and PPA, which was developed by Educational Testing Service (ETS) as a module-based assessment for the project. However, since PPA was only administered for one year, scores were only collected as baseline data for future analysis. The report noted:

“SAT-9 represents additional testing. Because no stakes are associated with the test, there seems to be low motivation to take it seriously. It may not be well aligned with state standards. State tests may be aligned with standards, but may not be aligned with the science modules used by the districts and short-run difference may be hard to detect.” (Consortium for Policy Research in Education, 2001, 63-64)

- *Design (3)*: Designed as a nonexperimental post-test time series analysis, the validity of the study is jeopardized without evidence from any comparison groups or prior performance. The strength of the study, however, rests in its use of qualitative methods such as case study and interviews to help increase understanding of the data.
- *Sampling (3)*: Little was explained about how the sample was selected, although it was described as being representative when comparing with the student population in the district.
- *Analysis (3)*: The evaluators used t-tests to indicate the difference of means between years of treatment. However, such statistical inferential procedure does not account for confounding variables and consequently, the internal validity is somewhat problematic. More importantly, since the LSC grant was used to strengthen other professional development services provided through the Merck program, it is hard to separate or disaggregate the effects of the LSC per se.
- *Reporting (5)*: The study was well written and reported.

### 3.4 Asset Inc. Teacher Enhancement (Pennsylvania, Cohort 2: Science K-8)

Rating summary (4): instrument (5), design (3), sampling (5), analysis (3), reporting (4)
---

#### Evaluation Summary

The Asset program was built around extensive professional development and the use of standards-based curriculum materials. The evaluators were interested in determining how program students compared nationally and internationally, and to what extent the length of teacher involvement in professional development made a difference in student performance measured by a modified TIMSS. The study found that continuing professional development and implementing standard-based curriculum had positive effects on student learning.

- Examining 1,542 students in the 3-4th and 7th grades, the study found that the mean scores for the 3-4th grade students were significantly higher than international and national scores whereas scores for the 7th graders performed at the same level as their national and international peers.
- When comparing 828 students who had been exposed to 4-5 years of the program with 714 who were in the program for 1-2 years, evaluators reported that the first group scored half a standard deviation higher than the second group. The same was true when data were compared across years for males and females, although there was no difference by gender at either of the time periods.

#### Rating of Procedures

- *Instrument (5)*: This study used TIMSS data in two different ways. For comparison of performance with national and international students, the overall TIMSS was used. To look more closely at the impact of the curriculum, certain test questions were selected from TIMSS to address science concepts and process skills emphasized in the modules used

in Asset districts (called Asset TIMSS). This is a good example of creating a subscale measure to align with the goals of LSC from the existing data.

- *Design (3)*: The comparison with international and national data, though informative, cannot be used as evidence for program impact, because there were no baseline data. However, this comparison might satisfy the information needs of local stakeholders. The second analysis used a post-test only design. Conceptual comparison groups were created by separating students whose teachers were exposed to different level of treatment to look at how degree of intervention affects outcomes. In addition, researchers used teacher questionnaires and class observations to detect factors contributing to students' science learning.
- *Sampling (5)*: Random selection was used in sampling, and the initial status of comparison and treatment groups was examined by comparing students' SES and school average scores.
- *Analysis (3)*: A t-test was performed to determine whether differences among groups were significant.
- *Reporting (4)*: Overall, the study was well presented and addressed not only concerns from NSF but also those from local stakeholders.

### 3.5 Mathematics Renaissance K-12 (California, Cohort 3: Math K-12)

Rating summary (3.8): instrument (5), design (3), sampling (5), analysis (2), reporting (4)

#### Evaluation Summary

Mathematics Renaissance (MRS) was design to help school districts to develop a connected and articulated mathematics program for students that provokes questioning and collaboration and encouraging learning concepts for understanding.

The study analyzed the item-by-item results<sup>5</sup> of students' performance from 9 California participating school districts in grades 4, 8, and 10 in Mathematics Application and Reasoning Skills, a performance assessment similar to NAEP and TIMSS according to the evaluators. The findings were also compared to earlier results from 1998-2000, using mean test. Evaluators concluded that while some improvement was evident each year, most students participating in the assessment had not yet achieved the desired levels of performance.

#### Rating of Procedures

- *Instrument (5)*: MARS was designed to help determine the nature of student learning that resulted from the program. It is well aligned with the LSC program.
- *Design (3)*: Using a post-test time series design, the study presented an alternative way to examine student outcome data that provide information on how students perform in different content areas. In addition, teachers were surveyed after participation in professional development programs.
- *Sampling (5)*: Since all of the students in selected grades were tested, there was little sampling bias.
- *Analysis (2)*: The analysis used exclusively descriptive statistics, coupled with t-tests to compare with results from previous years.
- *Reporting (4)*: The study presented an alternative approach to communicate the student assessment results to the stakeholders.

### 3.6 Science Connections (Maryland, Cohort 3: Science K-8)

Rating summary (4.6): instrument (5), design (4), sampling (5), analysis (4), reporting (5)

---

<sup>5</sup> The N under each item being examined varied from 72 to 1905.

## Evaluation Summary

The study provided evidence that Chemistry That Applies (CTA), a program-promoted teaching method, led to higher outcomes than typical curricular options in the district.

- Although overall achievement was low, the post-test mean was significantly higher than that of the pretest.
- The motivation and engagement measures reported higher scores on two out of five scales.
- An analysis by subpopulation using ANCOVA found significant gains by the lowest performing students. More students described as African Americans and Hispanics and classified as recipients of free and reduced-price lunches moved from the level of no understanding to a higher level of understanding than their peers in the comparison group. CTA appeared to have provided an advantage for students who had previously received ESL services.
- The CTA results were also positive on these outcomes with higher scores for basic learning engagement and mastery orientations toward earning science than in comparison group.

## Rating of Procedures

- *Instrument (5)*: Instruments included “conservation of matter” achievement assessment as well as measures of students’ engagement (Basic Learning Engagement and Advanced Learning Engagement) and motivation (Mastery Orientation and Performance Orientation). The instruments appear to align with the program goals.
- *Design (4)*: The study examined 2,251 8th graders in 10 schools (half treatment and half nontreatment). The design used treatment and comparison groups with pre-and post-tests. Additional qualitative evidence might further strengthen the conclusions.

- *Sampling (5)*: Since schools within each pair were randomly assigned to the treatment group and the comparison group, there was little selection bias. This was further verified by comparing the students’ demographics between two groups.
- *Analysis (4)*: ANOVA and ANCOVA were performed in data analysis, which explored not only the overall program impact, but also the effect on at-risk groups.
- *Reporting (5)*: The results of the study were well presented.

### 3.7 Minneapolis Public Schools Systemic Change in Science Initiative<sup>6</sup> (Minnesota, Cohort 3: Science K-8)

Rating summary (4.4): instrument (4), design (4), sampling (5), analysis (4), reporting (5)
---

## Evaluation Summary

The project studied the relationship between teachers’ use of various instructional practices and students’ achievement in science and mathematics. The evaluation involves data collection and analysis from multiple sites. Overall, it reported small but statistically significant positive correlation between student achievement and instructional practices that were consistent with LSC initiatives.

- The study found a substantial variation in teaching practices within schools regardless of the degree of participation in the program. Teachers’ use of LSC-advocated practices appeared to be positively related to student achievement at most sites, but the effects were quite small and rarely reached statistical significance.

---

<sup>6</sup> The study, known as Mosaic Project, was commissioned by NSF and conducted by Rand. The report was based on data from multiple sites that included the Minneapolis project.

- Use of traditional practices, by contrast, was often negatively related to student achievement, particularly in math.
- The largest positive relationship was found in a site with a standardized coefficient of 0.09. According to the evaluators, this means that at this site, the average students of a teacher who used appropriate reform practices monthly is predicted to score at about the 48th percentile on the test, while students of a teacher using reform practices weekly would score about 54th percentile.
- The pooled analysis revealed statistically significant positive relationships between teachers' use of LSC-advocated practices and achievement on both norm-referenced and criterion-referenced tests in math and science. However, the effects of professional development were much smaller than that of other variables such as SES.

### Rating of Procedures

- *Instrument (4)*: As a multiple-site evaluation, the study used a variety of instruments to measure student achievement, which enhances the validity of results. Measures included state tests, SAT-9, Comprehensive Test of Basic Skills (CTBS), California Initiatives Assessment Collaborative (CSIAC), and Metropolitan Achievement Tests (Met) (open-end, multiple choice). However, there was little indication about the extent to which and how these instruments were aligned with the program goals.
- *Design (4)*: The design was essentially a cross-sectional quasi-experiment in 6 sites (2 states, 4 systems), each of which was asked to selected 10 schools where reforms had been implemented and 10 where reforms had not yet been implemented.<sup>7</sup> The sample comprised 20,000 students in grades 3, 5 and 7. Teacher

questionnaires were used to measure instructional practice on a 5-point scale.

- *Sampling (5)*: The procedure for selecting comparison groups is interesting and practical. Schools and districts were asked to identify comparison groups based on their knowledge rather than strict demographic data.
- *Analysis (4)*: The researchers used linear regression analysis to control for student background characteristics and previous test scores. Separate analyses by subject areas (math/science), types of questions (open-ended/multiple-choice), and teaching practices (traditional/LSC) were performed. In addition, a pooled analysis was performed, assuming homogeneity across various sites. Controlling for confounding factors in the model further enhanced the validity of the findings, as did combining pooled analysis with separate subset analyses. However, since the identity of individual project was not revealed, it was difficult to interpret the overall success of the six projects as evidence of success for the Minneapolis project.
- *Reporting (5)*: The report was well written.

### 3.8 Pittsburgh Reform in Mathematics Education (Pennsylvania, Cohort 3: Math K-12)

Rating summary (2.8): instrument (4), design (3), sampling (2), analysis (2), reporting (3)

### Evaluation Summary

Since 1992, Pittsburgh Public Schools has adopted a standards-based system that reforms standards, assessments, accountability, curriculum, and professional development. The study looked at how the program design affected student outcomes and the effect of implementation on achievement gains. It examined three years of data from the New Standards Mathematics Reference Exam and ITBS for 4th graders. The evaluators concluded that the program appeared to produce an overall rise in math achievement in the district.

---

<sup>7</sup> The Mosaic project was conducted in 11 sites. Four sites are states (Connecticut, Louisiana, Massachusetts, and South Carolina) while seven are school districts or combinations of districts (Columbus, OH; Detroit, MI; El Paso, Socorro, and Ysleta, TX; Fresno, CA; San Francisco, CA; Philadelphia, PA; and Minneapolis, MN).

- The results from state assessments indicated significant improvement in skills and concepts categories, as well as a sharp drop in the number of students at the lowest score level.
- The results from ITBS showed little change.
- Rating schools according to the level of implementation, the study found significant difference between strong and weak implementation schools in both state assessments and ITBS.
- African Americans in strong implementation schools performed far better than those in weak implementation schools. However, the study also pointed out that teachers had been more effective in strong implementation schools prior to reform.

### Rating of Procedures

- *Instrument (4)*: This is another example of using both norm-referenced and standards-based assessments as student outcome measures.
- *Design (3)*: Although all schools were exposed to treatment, the study conceptually incorporated quasi-experimental elements by clustering schools according to implementation level. This is a useful way to deal with lack of baseline data while effectively addressing the implementation question.
- *Sampling (2)*: The sampling technique was unclear. It appeared that all of the students were sampled, but the relatively small number of observations suggested otherwise.
- *Analysis (2)*: The student outcomes among different groups were plotted and compared without using any statistical procedures. Student demographics and prior academic achievement were compared, but no controls were employed. Impacts on African American students were examined.
- *Reporting (3)*: While background, design, and discussion of the study were well explained, the section on how conclusions were reached needed further elaboration.

### 3.9 Reconceptualizing Mathematics Teaching and Learning Through Professional Development (New York, Cohort 4: Math K-8)

Rating summary (4.2): instrument (3), design (4), sampling (4), analysis (5), reporting (5)

### Evaluation Summary

The program was designed to improve teaching and student learning through professional development activities that target training in subject matter and the ability to enact the acquired knowledge and skills in the classroom. However, the evaluators did not find significant program effect on student outcomes:

- While classroom achievement in math varied, none of the classroom level predictors explained that variance.
- Teacher participation in professional development did not appear to have any significant influence on student achievement in math nor did it help reduce differences in achievement patterns of students from different SES backgrounds.

### Rating of Procedures

- *Instrument (3)*: Comprehensive Test of Basic Skills (CTBS), a norm-referenced test, was used as the instrument for student outcomes. There was little evidence provided on the extent to which the instrument was aligned with the LSC goals.
- *Design (4)*: The study was designed as a non-experimental post-test time series analysis with no comparison group or pre-test data. A teacher survey was used for additional data collection efforts.
- *Sampling (4)*: The study examined three years of data for 1536 students in grades 3-8. In an extensive discussion on data collection, the researchers pointed out such problems as missing data and lack of variations in some

classrooms, which might contribute to sampling bias.

- *Analysis (5)*: HLM was used for statistically modeling the effects of the hierarchical structure on student outcomes in order to avoid the problems of aggregation bias and mis-estimation of standard errors of parameter estimates. The analysis underscored the importance of using a more sophisticated statistical procedure than t-tests or ANOVA. On the surface, there was some indication that the program had been successful. Between 1988 and 1998, the percentage of students achieving at or above grade level rose from 66 percent to 82 percent in math, which may prove statistically significant through t-tests. However, the simple comparison does not tease out the confounding factors. For example, the district was said to maintain high standards for hiring teachers and principals and to target its resources at those students who need most attention. The findings of the study, however, suggested otherwise when after accounting for the confounding variables.
- *Reporting (5)*: The study was well reported.

### 3.10 Greater Philadelphia Secondary Mathematics Project (Pennsylvania, Cohort 4: Math 6-12)

Rating summary (2.4): instrument (4), design (3), sampling (1), analysis (2), reporting (2)
---

#### Evaluation Summary

Interactive Mathematics Program (IMP) is a new learning model that organizes instruction around applied problems that include content from algebra, geometry, and statistics in each year of the course sequence as opposed to the traditional sequence in which students learn each content separately for 1 year. Since LSC grant was used to fund an existing local initiative, the study analyzed both earlier and later data using matched samples. The evaluators concluded that program students consistently outperformed similar students who

were taught using a pre-reform standards curriculum and lecture style.

- In the Student Attitudinal Surveys in 1994, students overwhelmingly preferred IMP.
- IMP students had consistent higher passing and attendance rates in 1994-96. They outperformed students taught by traditional methods on the PSAT.
- In 1996-98, IMP students outperformed traditional students in a majority of the multiple-choice categories and had higher math-related subscores.
- In the Ivy League University Math Exit Test, IMP students completed more than 50 percent of the questions correctly; nonprogram students answered less than 25 percent correctly.

#### Rating of Procedures

- *Instrument (4)*: As did the Teacher Enhancement for Student Success (see 3.1), this study examined program impact by multiple criteria. However, none of the instruments appeared to be aligned with the LSC program. Although the researchers were well aware of the misuse of standardized test scores in comparing schools and curricular programs, they chose to analyze standardized test scores to deal with the “political reality.”
- *Design (3)*: The study appeared to be a post-test quasi-experiment, comparing outcomes from students received IMP with those who were taught by traditional methods.
- *Sampling (1)*: Since schools could choose to participate in IMP and there was evidence that many treatment schools were suburban, the study was prone to selection bias. However, no attempt was made to compare student demographics to rule out or adjust for potential bias. Using comparison group without considering the sampling issues could be quite misleading.
- *Analysis (2)*: Other than one conclusion that was based on ANOVA analysis, most of the

conclusions were derived by direct comparison using no statistical techniques.

- *Reporting (2)*: Since only a summary was provided, the study lacked a clear description of the research procedure.

### 3.11 Austin Collaborative for Mathematics Education (Texas; Cohort 4: Math 6-12)

Rating summary (3.8): instrument (4), design (3), sampling (5), analysis (3), reporting (4)

#### Evaluation Summary

The Austin Collaborative for Mathematics Education (ACME) projects was design to build the instructional capacity of math teachers by providing a minimum of 120 hours of professional development through summer institutes and follow-up sessions. Using scores from 550 students from grades 3 to 8 in both the Texas Assessment of Academic Skills (TAAS) and ITBS, the study attempted to examine the impact of implementation on student achievement based on observation ratings for 30 teachers.

- Students' math achievement as measured by TAAS was higher in classrooms with strong program implementation than those with weak and moderate implementation.
- There was no clear difference between student performance for teachers rated moderate and weak in implementation.
- There was no program effect on student achievement measured by ITBS.

#### Rating of Procedures

- *Instrument (5)*: Both in design and method, the study was similar to that the Minneapolis Public Schools Systemic Change in Science Initiatives (see 3.7). The evaluators used norm-referenced and criterion-referenced test results as outcome measures. While TAAS was primarily designed to meet the state

standards, there was evidence of alignment with the LSC objectives.

- *Design (3)*: The researchers conceptually assigned comparison groups according to the quality of implementation.
- *Sampling (5)*: Since teachers under study were randomly selected, their students were probably representative to the population.
- *Analysis (3)*: The study used ANOVA and Chi-square to test differences between groups and subgroups of interest.
- *Reporting (4)*: The study was well reported.

### 3.12 Valle Imperial-Project in Science (California, Cohort 4: Science K-8)

Rating summary (4): instrument (4), design (3), sampling (4), analysis (4), reporting (5)

#### Evaluation Summary

The program supports a constructivist approach in science learning through the use of kit-based instruction. The study summarized the results of a 4-year project in science education. The data were collected to measure student achievement in science, reading, writing, and mathematics and were analyzed relative to the number of years students participated in the program. Results indicated that achievement increased in relation to the length of student participation in the project.

- For results in science, positive correlation was found between years in program and respective mean scores for all five categories.
- Consistent differences were found between limited English proficient and English proficient students.
- The program had spill-over effect on mathematics learning.
- In writing and reading assessments, students' pass rates increased proportionally in relation to the number of years of participation.

## Rating of Procedures

- *Instrument (4)*: One of the distinctive features of the study was that it measured student achievement not only in science, but also in other subjects such as mathematics, reading, and writing. Since the district is located in the poorest county in California, where the majority of students are Latino, the program aimed for dual goals: (1) to develop cognitive knowledge of science content; and (2) to enhance writing skills in English. This might offer an example to study the spill-over effect not necessarily contained in the target content area, although the issue of spurious causation or simultaneity should be addressed. In terms of instrument selection, SAT 9 Form T was used to assess science and mathematics achievement because it was mandatory in California. Reading and writing skills were assessed by a district writing proficiency test developed locally.
- *Design (3)*: The study was a nonexperiment with post-test comparisons. Students were divided into groups based on years (0-4) of participation in the program. The study also included participant observations in classroom.
- *Sampling (4)*: The sample consisted of 628 in the 4th grade and 634 in the 6th grade. The sample was further disaggregated into five categories by proficiency levels in English: (1) limited English proficient; (2) limited/

fluent English; (3) fluent English (who speak an additional language); (4) English only; and (5) re-designated/fluent.

- *Analysis (4)*: For results in science, positive correlation was found between years in program and respective mean achievement scores for all five categories of students. In addition, ANOVA suggested consistent difference between limited English proficient and English proficient as well as consistent mean effect of years in scores for both grades. Similar results were found in mathematics. In writing and reading assessments, students' pass rates increased proportionally in relation to the number of years of participation.
- *Reporting (5)*: The report was well presented.

Table 3 presents a summary of ratings of the 12 studies on their performance in each procedure. The overall average quality of the studies was 3.7 out of 5, and five of the studies scored 4 points or above. Since the studies were preselected based on minimum criteria of rigor (see Appendix A), most of the studies do not contain serious flaws with regard to the 5 identified criteria, except for TESS (3.1), which used post-test only design and GPSMP (3.10), which did not address the sampling issue. Comparison across different procedures suggests that on average, the highest ratings were found for instrument selection and reporting.

**Table 3.**  
**Ratings of the reviewed studies**

Project	Instrument	Design	Sampling	Analysis	Reporting	Average
TESS .....	4	2	2	2	3	2.6
ESEP.....	3	3	5	4	4	3.8
PSC .....	4	3	3	3	5	3.6
ATE.....	5	3	5	3	4	4.0
MRS .....	5	3	5	2	4	3.8
SC .....	5	4	5	4	5	4.6
Mosaic.....	4	4	5	4	5	4.4
PRME.....	4	3	2	2	3	2.8
RMTTPD.....	3	4	4	5	5	4.2
GPSMP .....	4	3	1	2	2	2.4
ACME .....	4	3	5	3	4	3.8
VIPS .....	4	3	4	4	5	4.0
Average.....	4.1	3.2	3.7	3.2	4.1	3.7



## 4. ANALYSIS FOR INQUIRY-BASED SCIENCE IN SEATTLE PUBLIC SCHOOLS

---

### 4.1 Background

The Seattle Public Schools (SPS) LSC program<sup>8</sup> was designed to enhance an inquiry-based science teaching for K-5 teachers through a partnership with multiple departments and research centers in the University of Washington and several companies. SPS is a urban, multi-ethnic school district with a K-12 student population of about 50,000. Almost half of the student population are eligible for free or reduced-price lunch. The Partnership for Inquiry-Based Science, which had been piloted in two schools in 1995, was initiated in summer 1996. School participation was voluntary. NSF required the program to provide 100 hours of professional development for each eligible teacher. However, the program would stay open to all teachers once their schools have signed in. According to SPS, the program had expanded over the year until the summer of 2001 when all the rest of the 72 schools and approximately 94 percent of K-5 teachers participated in the program.<sup>9</sup> Although NSF had provided funding for 5 years, the Seattle program was underspent in 2000-2001, and therefore continues in 2001-02 on a no-cost basis.

The program offered the following activities for K-5 teachers to deliver effective science instruction to elementary students:

- Summer and fall science institutes on unit implementation, pedagogy, and content as well as classes throughout the school year;

- School-based professional development and support based on school or individual teacher needs they relate to the project;
- Materials support, analysis, and refurbishment provided by a materials supervisor, a volunteer scientist, along with other volunteers at the District Science Materials Center;
- Scientist and university support focused on the science content courses; and
- Other university support focused on family awareness and family celebration.

During our interviews with the LSC principal investigators, SPS indicated its interest in working with Westat in analyzing its LSC data. The purpose of this impact analysis is both to provide technical assistance to SPS and to demonstrate our approach to program evaluation.

### 4.2 Data

SPS has maintained a comprehensive database with student-level data on ITBS scores in science, math, reading, language, and social studies, since 1996. Scores in math and reading from Washington Assessment of Student Learning (WASL), a state-mandated criterion-referenced test, were available in reading and math from 1998.<sup>10</sup> In addition, student demographic data have been collected through the years. However, not until 2000 did the school district begin to collect data on teacher participation in LSC programs and to establish links with student data. The teacher

---

<sup>8</sup> SPS was also part of a five-district NSF LSC grant for middle schools, a separate project from the one described here.

<sup>9</sup> These data refer only to eligible teachers for science programs, excluding teachers in physical education, music, art, special education, and librarians.

---

<sup>10</sup> From 1996 and 1997, all grades 2-11 were tested with ITBS/TAP. In 1998, all grades 2-11 were tested with ITBS/TAP and grades 4 and 7 also tested with WASL. In 1999, grades 3, 5, 6, 8, and 9 were tested with ITBS/TAP and grades 4, 7, and 10 were tested with WASL. From 2000 on, ITBS/ITED has been implemented for grades 3, 5, 6, 8, and 9, and grades 4,7,10 were tested with WASL. The data on ITBS science scores pertain only to 5th graders.

data also include other teacher characteristics such as experience and education.

The overall research question is: Does the Inquiry-Based Science Program in SPS have any effect on student achievement in science? An ideal assessment strategy would include measurement of program treatment, the extent to which inquiry-based instruction has been applied in the classrooms, because this is what students eventually benefit from the program. However, SPS does not have any data on actual implementation. An alternative measure is the length of training received by teachers. The assumption is that the greater number of hours of professional development teachers receive, the more effective they become and hence the greater the impact on student achievement. This assumes that the quality of teachers' implementation of inquiry-based instruction is proportional to the quantity of professional development that they received. As previously mentioned, SPS did not collect data on hours of professional development for individual teachers until 2000. Using teacher participation as a measurement of professional development will lose observations from earlier years (1996-99). Another option is to use school participation as a proxy for professional development. Schools can be identified in different cohorts, indicating the year when they joined the program, and such an approach can take full advantage of the existing data SPS collected from 1996-2002.

Westat has created two datasets based on the raw data from SPS to address two sets of questions. The first set of questions examines how school participation in the LSC program affects student outcomes. The second set explores how the length of professional development for teachers impacts student achievement. This can be done by merging the student data with teacher data from 2000 to 2002. Because data are collected by year with the same variables, each year of data will be treated as a separate group. The researchers then stack up the

data to form two panels so that the first panel contains 7 years of data and the second panel comprises 3 years of data.

### 4.3 Research Design and Analysis

Based on the status of SPS data, the main research question can be broken down into the following questions:

- Does school participation in the program have any impact on student achievement?
- How does the length of professional development received by teachers affect student outcomes?
- How does the program affect student outcomes across different cohorts of schools and across different years?
- What effect does the program have on different subgroups of students (i.e., ethnicity, gender)?
- To what extent does the program affect student outcomes in other subjects such as reading, measured by both norm-referenced and criterion-referenced tests?
- Does the quality of implementation make a difference in program effect?

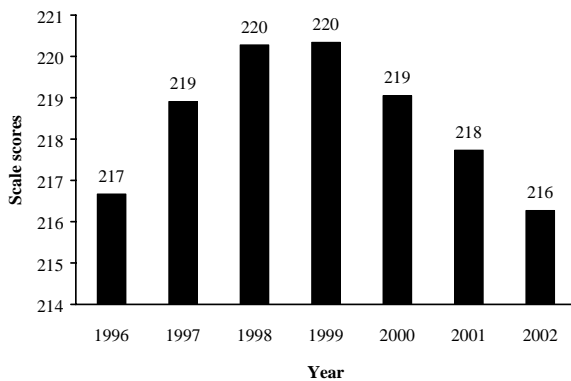
Table 4 presents descriptive statistics for the data, which give an idea of what our population is like. For example, during 1996-2002, about 58 percent of K-5 students in SPS are identified as non-white minorities, 48 percent are on free/reduced-lunch subsidy, 11 percent are bilingual, and 60 percent live with both parents. K-5 teachers, on average, have a bachelor's degree with 90 additional credits, and 12 years of experience. The demographic statistics appear to be stable over the years.

**Table 4.**  
**Descriptive of variables in the data (SPS)**

Variable	Label	Values	Mean	Std. Dev.
ITBSSC .....	Iowa scale score science	Numeric	216.47	35.55
ITBSSM.....	Iowa scale score math	Numeric	198.44	27.61
ITBSSR .....	Iowa scale score reading	Numeric	196.47	29.06
SPART.....	School participation in LSC	0=No, 1=Yes		
TPART .....	Teacher participation in LSC	Cumulative hours of training		
SEX.....	Student gender	0=male, 1=female	0.49	0.5
ETH .....	Student ethnicity	0=White, 1=Non-White	0.58	0.49
BILG.....	Student bilingual eligibility	0=No, 1=Yes	0.11	0.32
FRL.....	Student free/reduced lunch	0=No, 1=Yes	0.48	0.5
GIFT .....	Student gifted	0=No, 1=APP, 2=Spectrum	0.05	0.24
LWC .....	Student living arrangement	1=both parents, 0=other	0.06	0.49
GRD.....	Student grade	Numeric		
EDUC .....	Teacher education (degree+credits)	1=BA, 2=BA+22.5, 3=BA+45, 4=BA+45+MA, 5=BA+90, 6=BA+90+MA, 7=BA+135, 8=BA+135+MA, 9=BA+155+MA, 10=PhD	5.38	0.24
YRSEXP.....	Teacher years of experience	Total years of experience	12.82	9.25
CHT .....	Cohorts for school participation	1=1996, 2=1997, 3=1998, 4=1999, 5=2000, 6=2001		
TIME(S) .....	Years for school participation	1=1996, 2=1997, 3=1998, 4=1999, 5=2000, 6=2001, 7=2002		
TIME(T) .....	Years for teacher participation	1=2000, 2=2001, 3=2002		

Figure 2 presents changes of ITBS K-5 mean scores in science at the district level from 1996 to 2002. The message appears to be disturbing. The mean scores indicate a slight increase from 1996 to 1999 and then a steady decline until 2002 to the same level of 1996. Without the benefit of statistics, eyeballing the data suggests that the program does not have any effect on the student scores.

**Figure 2**  
**Changes of ITBS science scores in SPS: 1996-2002**



However, this broad look at findings may not present the entire picture. Indeed, simply looking at overall trends may mask important program effects. We, therefore, take a closer look at the achievement outcomes, using both general linear model (GLM) and multiple regression techniques. The purpose of using both is to demonstrate how the two strategies complement and reinforce each other. There are several advantages to using GLM. In addition to analyzing multiple factors simultaneously, we can test whether levels of one independent variable affect the dependent variable in the same way across the levels of the second variable. That is, besides detecting the existence of an overall program effect, one can identify the strongest program effect that occurred to what cohort of schools in what year. The strength of multiple regression lies in its ability to capture the net impact of the program. In other words, by controlling for other confounding factors associated with student outcomes such as students' ability and the environment, one will be able to evaluate the impact of the program itself.

The statistical procedures and analyses are detailed in the Appendix B.

## 4.4 Findings

Our analyses show that considerably more information is gained when multiple regression and GLS approaches are used. Multiple regression analysis produces consistent and more encouraging results.

- Both school and teacher participation in the LSC program have significant positive effects on K-5 student learning in science and measured by ITBS tests.
- Even when ITBS scores from 1999 to 2002 were in decline, the LSC program appears to have reversed the trend to some extent.
- The substantive significance of the program is modest. School participation in the program only results in an average increase of two points in students' ITBS science scores, and teacher participation in 100 hours of professional development merely lead to four points of increase.<sup>11</sup>
- Notwithstanding the overall impact, the program appears to be ineffective for African American students.
- Results for other subject areas are more ambiguous. In terms of possible spillover effect in reading, the program appears to have positive impact on students' reading measured by WASL. When measured by ITBS test, its effect becomes insignificant.<sup>12</sup>

---

<sup>11</sup>The moderate increases need to be viewed in light of history and relative impact of other factors that affect student achievement. In 7 years between 1996 and 2002, the average ITBS science scores had been fluctuated within 4 points of difference between 216 to 220. In this context, the seemingly modest increase is regarded as quite meaningful. In addition, the comparison of standardized coefficients suggests that the relative impact of school and teacher participation in the program is almost as large as influences from gender, ethnicity, or socioeconomic status from an individual student perspective. Given that we cannot change students' socioeconomic status, the LSC program may turn out to be one of the few tools one can use to make some meaningful changes in a short term.

<sup>12</sup>One may wonder if the type of instruction encouraged by LSC works better for application skills than recall abilities commonly tested by norm-referenced tests. If true, that might as well be the purpose of the program. Furthermore, if reading scores are any indication of the overall alignment of the ITBS measures, we might expect to see an even larger impact of student learning in science, if

The GLM analysis indicates similar results in that school participation has modest impact on student achievement. It also suggests that:

- The program impact was more apparent in the first year of a school's participation than in the later years.
- Male and white students seem to benefit more from the program than female and African American students. There is no evidence that the program narrows the gap between gender in science achievement.
- Schools in different cohorts performed differently. Schools of cohort 1997 whose test scores increased in three consecutive years demonstrated a long-term program effect while all other cohorts demonstrated a 1-year effect.
- It is interesting to note that all schools in SPS seemed to perform less proficiently after year 1999. This might be attributed to other overwhelming changes that occurred in the SPS in 1999, and the negative effects these changes may have cancelled out the modest LSC program effect. A follow-up study is needed to investigate this widespread negative effect.

Judged by the criteria that we set out to review other studies, we recognize that our study is not free of problems. Some of those issues are as follows:

- *Instrument:* We only used one instrument to measure student outcomes, and there is evidence that the ITBS science test is not well aligned with the LSC objectives.
- *Design:* Our design is that of a pre- and post-test time series analysis with no comparison group. Our qualitative approach has primarily been limited to discussions with program staff and review of program reports. We would like to uncover the "black box" of implementation by using strategies such as a survey of teachers and interviews with program staff,

---

it had been measured by test instruments that are better aligned with the LSC goals. The results from previously reviewed studies (see sections 3.8, 3.11) seem to lend credence to this expectation.

professional development staff, and teachers. Qualitative information could help us explain our findings from the quantitative analysis and uncover the conjunctual link between professional development activities (input) and student achievement (outcome).

- *Sampling:* The selection bias in program participation among different cohorts is obvious, but our models are relatively successful in mitigating the bias.

- *Analysis:* We would like to explore the possibility of using multi-level modeling to account for the possibility of a cluster effect. This strategy adopts a more robust statistical procedure to analyze nonindependent sample data. For example, in the school setting, repeated observations of a student are not independent from each other and students are not randomly assigned to a teacher.



## 5. CONCLUSIONS AND RECOMMENDATIONS

---

The analysis of existing studies as well as our own study on LSC projects present important findings on both our knowledge about the effect of these programs and approaches for future NSF programs in professional development.

### 5.1 What Have We Learned about the LSC Impact on Student Outcomes?

The study has provided consistent but modest support to the LSC program.

- In addition to our own analysis, findings from 10 of the 12 studies reviewed suggest that the LSC programs lead to better student outcomes in the target areas.
- Researchers disagree on the substantive effect of the program. While some studies report substantial program impact on student outcomes, others find that the program has contributed to a modest increase in student learning. Some indicate that even if improvement is evident, it is not up to the desired level.<sup>13</sup>
- Studies also draw different conclusions about the impact of the programs on subpopulations. While some detect significant effects on minority students, others argue that the impact is insignificant.
- There is some suggestive evidence that programs promoting standards-based teaching result in improvement not only in the target subject area, but also have a spillover effect in other area such as reading, especially in sites with a high percentage of minority students.
- The impact analysis is highly sensitive to the outcome measures. It appears that the program is more likely to show effect on criterion-

referenced tests, especially those that are well-aligned with the program goals, than on norm-referenced tests.

- Correct implementation is a prerequisite for effective outcome evaluation, a fact that is often overlooked in many impact analyses. Evaluators should be careful about condemning a program simply because we do not observe the intended outcomes. On one hand, the program might not be implemented properly. On the other hand, it may have achieved other necessary conditions that will eventually lead to the intended outcomes. A few studies recognize the importance of implementation and have demonstrated that well-implemented programs lead to better student learning.

There are questions about the validity of the results, as some analyses lacked adequate statistical controls to tease out confounding factors other than the program impact. Such factors might include students' socioeconomic status and prior achievement, as well as teacher and school characteristics. Another concern is that the reviewed studies represent 20 percent of the LSC projects, leading to questions about its external validity. In other words, how generalizable are these results to other LSC projects that did not have data or conduct analyses?

- We also have gained considerable knowledge about the technical aspects of these studies. A variety of instruments were used to measure student achievement such as standardized tests scores (i.e., SAT-9, ITBS, CTBS); state assessment tests in California, New Jersey, Maryland, Pennsylvania, and Texas; assessments developed solely for the purpose of the program; as well as measurement for students' attitudes, motivation, and engagement. Many studies used instruments that were well aligned with the LSC goals.

---

<sup>13</sup>No study reports effect size of the program. Based on available statistics, we calculated the program effect size for five studies, which range from 0.07 to 0.5 with an average of 0.26.

- Student outcomes from grades 3 to 10 were assessed most frequently with the exception of the 6th grade.
- Studies involved either quasi-experimental design or nonexperimental design. No randomized experimental design was performed. While some studies used pre- and posttest comparisons, others chose post-test analysis.
- Studies used a variety of analytical methods from univariate analysis to multivariate analysis such as t-test , ANOVA, ANCOVA, and multiple regression. One study employed multi-level analysis.

## 5.2 What Can We Learn About the Common Issues?

The study raises a set of common issues for the LSC projects. Among projects that have conducted impact analyses, the evaluation suggests that the overall quality of the efforts is moderately satisfactory. The average overall rating for 12 studies is 3.0 on a 5-point scale. Five out of 12 of the studies can be described as having relatively high quality with ratings equal to or above 4. These studies can serve as examples for future evaluation efforts. Table 3 shows the ratings of each study based on the five identified criteria. The overall quality of LSC outcome evaluation is better than similar studies found in a synthesis by Westat (Frechtling et al., 1995)

- In terms of instrument selection, many studies used multiple instruments, including some instruments that were well aligned with LSC objectives. In cases where the instruments were not aligned with program goals, the evaluators generally were able to recognize and discuss the limitations when interpreting the findings.
- Both quasi-experimental design and nonexperimental design were employed. Only one study used quasi-experimental design with pre- and post-tests, which is generally regarded as the strongest design other than randomized experimental design. Although the

lack of such design is, in many ways, a reflection of data limitations, many studies have demonstrated useful ways to cope with various data constraints, which will be discussed in the next section. It is encouraging that most studies had adopted a mixed-method approach by combining both quantitative and qualitative analysis.

- The sampling techniques leave some room for improvement, as some studies did not address how teachers/students were selected for participation in the program and whether this may have affected the results. This is especially troublesome for studies using quasi-experimental design because selection is usually a major problem.
- A few studies used univariate t-tests or ANOVA to test the difference between treatment and comparison groups or results before and after the treatment. Such procedures may have been adequate in randomized experiments where treatment and control groups are randomly assigned and conditions are strictly controlled, or in quasi-experiments where treatment and comparison groups are closely matched. But in nonexperiments or quasi-experiments where matching is not done effectively, controlling for confounding factors in the statistical models is necessary if one is interested in the net impact of the program. Therefore, multivariate analysis is necessary.
- Reporting scores highest among the five criteria, as many studies address different aspects of a research adequately. However, a few studies did not provide adequate explanation in design and sampling procedures.

For projects that had not conducted impact analysis, it is important to underscore the importance of evaluation. Evaluation is not separate from or added to a project, but rather is part of it. Evaluation provides information to help improve the project and to communicate to a variety of stakeholders. In light of many federal and local reform initiatives that target student outcomes and accountability, it is critical to

conduct impact analysis in order to allow projects to better tell their story and prove their worth.

Interviews with project evaluators suggest that data limitations present major obstacles in evaluation efforts. Although most of the states and districts routinely test their students, evaluators sometimes cannot get the access to individual student data. Some localities do not have achievement tests in the target areas such as science. Many projects could not link student scores to teacher professional development. Philosophical concerns sometimes prevent evaluators from conducting impact analysis. Without assessments well aligned to the LSC goals, evaluators were reluctant to use tests that were not believed to reflect what had been done. But measuring outcomes is absolutely necessary in any accountability system, and instead of waiting for the perfect measure, evaluators should probably use the best available measure—and make note of its pros and cons.

The advantages of using commercially produced tests include their familiarity to most of the stakeholders and their availability and variety. Most of these tests have been validated against a national sample in order to determine what children in a certain grade are expected to know. A disadvantage is that these products are not tailored to local standards, and tend not to model the kind of teaching and learning embraced by reform efforts. Another approach, and one in line with recommendations in the national science and mathematics standards, is to use open-ended items or performance assessments that involve multiple responses that can reflect real-life, complex problems. Disadvantages of this approach include the difficulty of finding an appropriate instrument and the amount of time needed to administer and score the performance items, as well as the costs associated with each (Horizon Research and Westat 2001). Many evaluations rely on locally available student achievement data, in large part because administering additional measures is expensive and often not feasible. Many principals and teachers believe that their students spend far too much time taking the tests that are required locally, and they are reluctant to volunteer for supplementary testing. Locally developed and administered tests may also be preferred because

they are presumed to be more closely aligned with local reform efforts than a measure chosen by outside evaluators would be (Klein, Hamilton, and Rand 2001).

### **5.3 What Strategies Can We Suggest to Improve Future Program Evaluations?**

The evaluation has several important implications for NSF. First, a strong case can be made for continuing support of data collection, program evaluation, and outcome-oriented assessment in future programs. As the standards movement gathers increasing momentum among the stakeholders, it is clear that student achievement has become the most valued and credible outcome. For NSF, it is important to emphasize student achievement in its future funding activities. For local school systems, addressing the public concerns for student achievement by empirically demonstrating the results is almost inevitable. Nevertheless, researchers must be careful about overemphasizing the achievement data. After evaluating the pros and cons of each measure, our study concludes that the best approach is to use multiple measurement.

Second, NSF should encourage innovative approaches that emerged from existing LSC studies. Evaluators are confronted all too frequently with situations where it is impossible to implement the “very best” evaluation design because of ethical reasons, time and resources, and cost. Nevertheless, one should strive to use the same logic of inquiry and research procedures. Evaluators must review the range of design systems in order to determine the most appropriate one for the particular evaluation. There is no single, always best design, and the choice always involves trade-off. We advocate using what Rossi and Freeman (1995) call the “good enough” rule in which the evaluator should choose the best possible design from a methodological standpoint, having taken into account the potential importance of the program and the practicality and feasibility of each design.

Finally, interviews with the principal investigators of Cohort 1-4 projects suggest a dilemma between a high level of interest in evaluation and a lack of

capacities for conducting rigorous evaluations. We recommend that future NSF programs provide technical assistance to help projects develop their own abilities to conduct evaluative work.

The studies reviewed here represent different approaches to dealing with both the technical and practical constraints in program evaluation. The following discussion will elaborate strategies that have shown promise and can be replicated in other settings.

Outcome measurement is crucial for program evaluation. Good measures should not only be reliable and relevant, but also practical. In selecting instruments, multi-measurement should be encouraged.

- Despite the repeated criticism of the standardized tests, which rely on selected responses to multiple-choice items, many projects recognized that these tests were recognized by the stakeholders and easily available. Further, they are required in most states and are used for assessing school effectiveness.
- Instruments that are well aligned with the LSC objectives are particularly effective in measuring the program outcomes. Some programs took the efforts in designing and administering additional assessments that are well aligned with LSC goals. These assessments include Partnership Performance Assessment (PPA) in New Jersey (3.3), Mathematics Application and Reasoning Skills (MARS) in California (3.5), and Chemistry that Applies (CTA) in Maryland (3.6). However, such efforts might be costly for some localities. Projects such as ATE in Pennsylvania (3.4) used an alternative approach by constructing a subscale of assessment from an existing commercial test (TIMSS), so that the modified assessment is in line with the program goals. This option is only feasible if the evaluator has the access to results for individual items on the assessment.
- Georgia (3.2) and Maryland (3.6) studies chose student attitude measures as outcomes.

- VIPS in California (3.12) looked at not only program impact on the content area, but also spillover effects in other subjects such as mathematics, reading, and writing.

The integrity and credibility of a study depend on having an appropriate and sound study design. In general, the mixed-method approach combining both quantitative and qualitative analysis, which takes advantage of the strengths and minimizes the weakness of each strategy, should be encouraged. For example, while a quantitative approach is ideal for producing broad statements and dominant patterns pertaining to large bodies of data and diverse cases, qualitative strategies can be used to understand how and under what circumstances different factors are combined to achieve certain outcomes. A combination of these two strategies can be used to complement and triangulate findings (Ragin, 1997).

In an ideal world, evaluators would like to use randomized experimentation. Unfortunately, in most applications, including NSF projects, these conditions simply cannot be created. Most of the programs are not targeted to participants in a carefully controlled and restrictive environment, but rather to those in a complex social environment that has a bearing on the success of the project. Nevertheless, many useful approaches have been adopted to increase the validity and reliability of the studies.

- If pre-test data are available and the project has a comparison group, a quasi-experimental design involving treatment and comparison groups and pre- and post-tests is often the most rigorous design.
- However, many projects do not have both conditions and yet still produce useful studies. For example, if no pre-test data are available, a study is best if it involves a comparison group. The comparison group can be created or identified in various ways either in reality or in concept:
  - It could be a group that did not receive the treatment;

- It could be a group received different level of treatment. In other words, if all teachers in the district are participating to some extent in the program, one might look at differences in the amount of training each has received (quantity) or the degree of implementation (quality) as ways to define treatment and comparison groups;
- A comparison group also can be established against an accepted standard such as outcomes of students in similar districts or grade-level-equivalent scores. For example, some studies compare the program outcomes against that of the state and the nation.
- If no comparison group is available, one may still consider conducting pre- and post-tests with the treatment group. In this type of analysis, it is advisable to include repeated measures of the outcomes over several years instead of at two points of measurement.
- In general, the “treatment group only, post-test only” design used by a few projects should be discouraged. However, if that is the only design possible, one may want to collect multiple years of data, carefully documenting other variables that might be contributing to the outcomes, presuming that any changes that have occurred over time are net effects.

Evaluators frequently brought up the problem of lack of individual student data. Although we should encourage efforts to collect individual data, higher levels of outcome data such as grade or even school can be used to study program impact. Instead of examining the program impact on individual student, the study would then address the question of how program affects the appropriate unit of analysis (e.g., grade, school).

Another concern is that student outcome data are not linked with professional development. While it is extremely desirable to link individual student outcomes to teacher participation in professional development, as our study for SPS demonstrates, one can study the difference between schools,

which are involved in the program and non-program schools, or between the pre-program years and after-program years, coding the presence or absence of treatment as dummy variables.

In addition to student outcome data, it is also important to collect data on student demographics and on program implementation. Implementation data might include measurement of level of treatment (quantity), as well as extent of implementation (quality). These data are essential to build the analytical model for analyzing program impact.

A critical distinction must be made between gross outcomes and net outcomes. Gross outcomes consist of all observed changes in a measure. They are easily measured and ordinarily consist of the differences between pre- and post-program values on some measure. Net outcomes, which are much more difficult to measure, are those results that can be reasonably attributed to the intervention, free and clear of the effects of any other causes that may be at work (Rossi and Freeman, 1995) The relationship between gross and net outcomes can be expressed in Figure 3:

**Figure 3.**  
**Gross versus Net Outcome**

$\text{Gross outcome} = \text{Effect of intervention (net outcome)} + \text{Effects of other processes (confounding factors)} + \text{Design effects}$
--

Confounding factors refer to all sorts of threats to internal validity such as history (i.e., external factor, testing, maturation, regression artifacts, attrition), selection, and contamination. Design effects, which result from the research process itself, require different strategies to deal with random and non-random errors. Discussion of confounding factors and design effects can be found in many statistics textbooks and are not the major focus of our analysis.

All of the studies being analyzed involve some sorts of statistical testing. Many rely on univariate t-tests and ANOVA. These methods are adequate for randomized experiments and quasi-

experiments where matching of treatment and comparison group is successful. However, in nonexperiment and quasi-experiment where matching is not so successful, these two methods are only able to indicate differences between treatment and comparison groups or between the years with treatment and those without rather than to the program itself. Various confounding variables such as student demographics and teacher and school characteristics need to be controlled for in the statistical model. In other words, univariate t-tests and ANOVA are adequate for gross outcome analysis. However, multivariate analysis such as ANOVA, MANOVA, multiple regression and hierarchical linear model (HLM) should be encouraged if one is interested in net impact.

In addition, it is advisable to subdivide the total group into subunits of particular interest. Such practice will allow the evaluator to detect program impact not only on the overall population, but also on various subgroups to show if there are patterns in the outcomes. Sub-groups might be divided by demographics such as gender, race/ethnicity, and

economic status, as well as level of implementation. Or if programs are found to be effective, we would also be concerned whether they address the needs of special populations. If programs are found to be ineffective, is the failure attributable to the design or implementation?

Even the strongest study will be of very little interest or value if it is not reported appropriately and well. A full report should provide an overview of the project, including the objectives and goals. It is extremely important to provide a clear description of the study including instrument, data collection and analysis process. Results should be discussed with alternative explanation provided. The study should address conclusions and implications to a variety of stakeholders.

For overview on program evaluation, please refer to *The 2002 User-friendly Handbook for Project Evaluation* (Frechtling, 2002). For techniques in impact analysis, *Revised Handbook for Studying the Effects of the LSC on Students* (Horizon Research and Westat, 2001) is a good reference.

## REFERENCES<sup>14</sup>

---

- \*Amaral, O., Garrison, L., and Klentschy, M. (2002). Helping English Learners Increase Achievement Through Inquiry-based Science Instruction. *Bilingual Research Journal*, 26(2): 213-39.
- \*Austin Independent School District. (2001). *Austin Collaborative for Mathematics Education. 1999-2000 Evaluation*. Austin, TX: Austin Independent School District.
- Banilower, E. (2000). *Local Systemic Change Through Teacher Enhancement: A Summary of Project Efforts to Examine the Impact of the LSC on Student Achievement*. Chapel Hill, NC: Horizon Research, Inc.
- \*Boaler, J., et al. *Mathematics Teaching and Learning Study: A Comparison of IMP1 and Algebra 1 at Greendale School*. Unpublished report prepared for the National Science Foundation.
- \*Briars, D., and Resnick, L. *Standards, Assessments – and What Else?* Unpublished report prepared for the National Science Foundation.
- \*Calhoun, D. (2002). *Fresno Systemic Program: Evaluation Report*. Unpublished report prepared for the National Science Foundation.
- \*Carrol, C., et al. (2001). *Mathematics Renaissance K-12: MARS/MRK-12 Student Performance Assessment Report*. Unpublished report prepared for the National Science Foundation.
- Cohen, D., and Hill, H. (1998). *State Policy and Classroom Performance: Mathematics Reform in California*. Philadelphia, PA: Consortium for Policy Research in Education.
- \*Consortium for Policy Research in Education. (2001). *Steady Work: A Report on the Seventh Year of the Merck Institute for Science Education, 1999-2000*. Philadelphia, PA: Consortium for Policy Research in Education.
- Darling-Hammond, L., and Ball, D.L. (1998). *Reading for High Standards: What Policymakers Need to Know and Be Able to Do*. Philadelphia, PA: National Commission on Teaching and America's Future. Consortium for Policy Research in Education.
- Darling-Hammond, L., and Rustique-Forrester, E. (1997). *Investing in Quality Teaching: State-Level Strategies*. Denver, CO: Education Commission of the States.
- Elmore, R.F. (2002). *Bridging the Gap between Standards and Achievement: The Imperative for Professional Development in Education*. Washington, DC: Albert Shanker Institute.
- Fantzzo, J.W., King, J.A., and Heller, L.R. (1992). Effects of Reciprocal Peer Tutoring on Mathematics and School Adjustment: A Component Analysis. *Journal of Educational Psychology*, 84: 331-39.
- Fox, J. (1998). "NSF Programs Attacked as Weak, Unclear." *Education Daily*, July 24, 1-2.

---

<sup>14</sup> Asterisks indicate studies analyzed in this report (see Table 1)

- Frechtling, J. (2002). *The 2002 User Friendly Handbook for Project Evaluation*. National Science Foundation.
- Frechtling, J., Sharp, L., Carey, N., and Vaden-Kierman, N. (1995). *Teacher Enhancement Programs: A Perspective on the Last Four Decades*. Arlington, VA: National Science Foundation: Directorate for Education and Human Resources.
- Garet, M.S., Porter, A.C., Desimone, L., Birman, B.F., and Yoon, K.S. (2001). What Makes Professional Development Effective? Results From a National Sample of Teachers. *American Education Research Journal*, 38 (4): 915-45.
- Ginsburg-Block, M.D., and Fantuzzo, J.W. (1998). An Evaluation of the Relative Effectiveness of NCTM Standards-based Interventions for Low-Achieving Urban Elementary Students. *Journal of Educational Psychology*, 90: 560-69.
- Greenwood, C.R., Carta J.J., and Hall, R.V. (1988). The Use of Peer Tutoring Strategies in Classroom Management and Educational Instruction. *School Psychology Review*, 17: 258-75.
- \*Harwell, M., D'Amico, L., Stein, M. K., and Gatti, G. (2000). Professional Development and the Achievement Gap in Community School District #2. Unpublished report prepared for the National Science Foundation.
- Horizon Research, and Westat. (2001). *Revised Handbook for Studying the Effects of the LSC on Students*.
- Kennedy, M. (1998). *Form and Substance in Inservice Teacher Education* (Research Monograph 13). Madison, WI: National Institute for Science Education.
- \*Klein, S., and Hamilton, L. (2001). *Mosaic of Evaluation of Systemic Reform Initiatives*. Santa Monica: CA: Rand.
- \*Lynch, S., and Pyke, C. (2002). Preliminary Report to the National Science Foundation (IERI) on the Quantitative Results of Examining the Effects of Highly Rated Science Curricula for Diverse Student Population. Washington DC: George Washington University.
- Mayer, D.P. (1998). Do New Teaching Standards Undermine Performance on Old Tests? *Educational Evaluation and Policy Analysis*, 20: 53-73.
- \*Raghavan, K., Cohen-Regeve, S., and Strobel, S.A. (2001). Student Outcomes in a Local Systemic Change Project. *School Science and Mathematics Journal*, 101: 417-26.
- Ragin, C. (1987). *The Comparative Method: Moving Beyond Qualitative and Quantitative Strategies*. Berkely and Los Angeles, CA: University of California Press.
- \*Redsnick, L.B., and Harwell, M. R. (1998) *Professional Development and Teaching Quality in a Standards Referenced Education System*. Unpublished report to the U.S. Department of Education, Office of Educational Research and Improvement on the UCLA-CRESST project.
- Rossi, P.H., and Freeman H.E. (1995). *Evaluation: A Systematic Approach (5th ed)*. Newbury Park: Sage, CA: Publications.

- Sanders, W.L., and Horn, S.P. (1994). The Tennessee Value-Added Assessment System (TVAAS): Mixed-Model Methodology in Educational Assessment. *Journal of Personnel Evaluation in Education*, 8:299-311.
- Seattle Public Schools. (2001). *Hands-On Science in Seattle Public Schools, K-5 Annual Overview*. Unpublished report prepared for the National Science Foundation.
- Verschaffel, L., and De Corte, E. (1997). Teaching Realistic Mathematical Modeling in the Elementary School: A Teaching Experiment with Fifth-graders. *Journal for Research in Mathematics Education*, 28: 577-601.
- Weathersby, J., and Harkreader, S. (1999). *Staff Development and Student Achievement: Making the Connection*. Paper presented at the Annual Meeting of the American Education Research Association. Montreal, Quebec.
- Wolf, F. M. (1986). *Meta-analysis: Quantitative Methods for Research Synthesis*. Beverly Hills: Sage Publications.

**Appendix A**  
**Rating Criteria**



# APPENDIX A. RATING CRITERIA

---

## Instrument

- 1 no outcome measures
- 2 single instrument (i.e. commercial test, state assessment, attitude)
- 3 single instrument with discussions of alignment
- 4 multiple assessments with discussions of alignment
- 5 measures with evidence of alignment

## Design <sup>15</sup>

- 1 no explicit discussion of design issues
- 2 post-test only with no comparison groups
- 3 post-test time series; pre- and post-test with no comparison groups
- 4 pre- and post-test time series with no comparison groups; post-test with comparison group
- 5 randomized experiment or quasi-experiment with pre-post comparison design

## Sampling

- 1 no discussion of sampling procedure
- 2 evidence of bias in sample selection without correction
- 3 discussions that recognize sampling bias
- 4 adequate procedures to mitigate sampling bias
- 5 little sampling bias

## Analysis <sup>16</sup>

- 1 direct comparison
- 2 correlation, chi-square or t-test on selected instrument
- 3 univariate analysis (t-test, ANOVA)
- 4 multivariate analysis (ANOVA, MANOVA, multiple regression)
- 5 multi-level analysis (HLM)

## Reporting

- 1 addresses less than 3 aspects of the 5 procedures
- 2 addresses 3 or more aspects of research
- 3 all aspects are addressed but not well explained
- 4 good discussion of procedures and results
- 5 excellent discussion of procedures and results with implications to different stakeholders

---

<sup>15</sup>Rating assumes that studies will employ mix-method strategy. Lack of mix-method approach will be penalized by at least 1 point.

<sup>16</sup>These only measure level of sophistication in data analysis, whereby level of appropriateness will be judged in conjunction with the status of data and study design.



**Appendix B**

**Statistical Procedures and Analyses for SPS**



# APPENDIX B. STATISTICAL PROCEDURES AND ANALYSES FOR SPS

---

This appendix provides detailed elaboration on the statistical procedures and analyses regarding the LSC program impact on SPS 5th grade student achievement in science. The analysis begins with a test of difference of means, a procedure used by many existing studies. The main procedure involves two more advanced methods, namely, general linear modeling (GLM) and multiple regression analysis to demonstrate the added-value.

## 1. School Participation (t-test)

Before performing GLM and multiple regression analysis, we first use an independent sample t-test to examine whether students' scores changed significantly after their schools joined the program. Table B-1 shows the results for the overall data from 1996-2001, as well as data by year, which indicate that post-test scores are significantly higher than pre-test scores. Specifically, the post-treatment scores for cohort 1997, 1999, 2000 are higher than pre-test scores<sup>17</sup>. The post-test scores for cohort 2001 are significantly lower than pre-test scores. The difference between pre-and post test scores for cohort 1998 is not statistically significant.

The test of difference of means suggest that the program appears to be effective in improving student learning in science. The effect size of the program is 0.13, which is below the 0.20 threshold

for minimum of educational significance. However, the t-test does not take into account the confounding factors that might also contribute to the differences. Such concerns can be addressed by using GLM and multiple regression analysis.

## 2. School Participation (GLM Analysis)

The purpose of this analysis is to determine whether school participation in the LSC program affects students' achievement scores in science. School participation in LSC was not standardized in SPS, that is, schools participated in the program in different school years with different percentages of teachers enrolled. For the participating teachers, the dosages of the training, in terms of hours, were also different. Implementation data indicating to what extent LSC strategy was applied in the classroom were scant. Due to the limitations of the data, we can barely group schools by any substantively meaningful way, except by the participation cohort. Also, every student in a school was considered as a participant if the school enrolled in the program, and their performances were compared with those students in schools not participating.

It is important to note that GLM model, unlike that of the ordinary least square model that multiple regression uses, does not require any assumption of the linear order on part of the independent variables. GLM compares school performance of

**Table B-1.**  
**Test of difference of means**

	Overall	1996	1997	1998	1999	2000	2001
Pre-test.....	213.73	0.00	216.50	220.85	213.49	212.17	211.95
Post-test.....	218.25	222.57	221.17	220.82	215.49	214.39	209.19
Sig (2-tailed).....	0.00	NA	0.00	0.99	0.07	0.03	0.02

---

<sup>17</sup>Cohort 1996 did not have pre-test data.

different cohorts with one another in order to determine if schools of various cohorts perform differently by calculating an overall statistical significance. If a significant overall effect is found, a post hoc comparison can be conducted to detect separate effect from different cohorts.

A general notation for our GLM model is illustrated as follows:

$$*^2(\text{dependent variable}) = *^2(\text{main effects}) + *^2(\text{interaction effects}) + *^2(\text{error})$$

The total variation of the scores is partitioned into three sources:

- Variation between the group means and the grand mean (main effects),
- Variation between the means of each interaction levels and the grand means (interaction effects), and
- Variation of scores within groups (or random error).

There are three main effects in our design. The first factor—cohort—indicates the year school participated in the program. There are six levels in the cohort variable (1996-2001). The second factor—time—indicates the school year when student achievement tests were administered. There are seven categories in the time factor from 1996 to 2002. The last and probably the most important factor is treatment, which indicates school participation status in the school year when the achievement test was administered.

We choose ITBS science scores as the dependent variable. Only 5th grade students were included in the analysis because 3rd and 4th graders were not required to take science tests.

In order to control the sizable differences of students’ economic status and intellectual ability among students, free-reduced lunch and gifted program are entered as covariates in the model. The purpose of introducing covariates in the model is to “equate” schools that are essentially nonequivalent, thereby increasing the precision of the analysis. An elaborate notation for our GLM model is:

### Model 1

$$*^2(\text{science score}) - *^2(\text{SES \& intellectual ability}) = *^2(\text{cohort})^{18} + *^2(\text{time}) + *^2(\text{treatment}) + *^2(\text{gender}) + *^2(\text{race}) + *^2(\text{time*cohort}) + *^2(\text{cohort* treatment}) + *^2(\text{cohort*time*treatment}) + *^2(\text{error})$$

### Model 1-1: Base Model

In the first model, we include two fixed factors (time, cohort); and two covariates (frl, gift). Table B-2 (Model 1-1) shows that after controlling students’ free-reduced lunch status and intellectual ability, schools in different cohorts performed differently (F=28.3, P=.00) and test scores changed significantly from year to year (F=9.36, P=.00). More importantly, the interaction between time and cohort is also significant (F=1.9, P=.00), indicating that the program effects across seven time points were not the same among cohorts. We conclude that different cohorts must be analyzed separately for treatment effects, and the program effects need to be investigated year by year.

---

<sup>18</sup>Denotes the amount of variance of the dependent variable that the independent variables, cohort, time...etc. accounted for.

**Table B-2.**  
**Summary table of general linear analysis of main effects**

Source	BASE (1-1)			GENDER (1-2)			RACE (1-3)			TREATMENT (1-4)		
	F <sub>(df)</sub>	Sig.	Effect Size	F <sub>(df)</sub>	Sig.	Effect Size	F <sub>(df)</sub>	Sig.	Effect Size	F <sub>(df)</sub>	Sig.	Effect Size
Main effect												
Time .....	9.3 <sub>(6, 22733)</sub>	0.00	0.00	8.9 <sub>(6, 22691)</sub>	0.00	0.00	3.3 <sub>(6, 14478)</sub>	0.00	0.00	9.7 <sub>(6, 22733)</sub>	0.00	0.00
Cohort .....	28.3 <sub>(5, 22733)</sub>	0.00	0.00	26.9 <sub>(5, 22691)</sub>	0.00	0.00	9.6 <sub>(5, 14478)</sub>	0.00	0.00	10.5 <sub>(5, 22733)</sub>	0.00	0.00
Gender .....				42.4 <sub>(1, 22691)</sub>	0.00	0.00						
Race .....							1256.5 <sub>(1, 14478)</sub>	0.00	0.10			
Treatment .....										7.2 <sub>(1, 22733)</sub>	0.00	0.00
Interaction												
Time * Cohort .....	1.9 <sub>(30, 22733)</sub>	0.00	0.00	1.9 <sub>(30, 22691)</sub>	0.00	0.00	1.8 <sub>(30, 14478)</sub>	0.00	0.00			
Treatment *												
Cohort .....												
Time * Cohort *				1.6 <sub>(35, 22691)</sub>	0.00	0.00						
Gender .....												
Time * Cohort *							2 <sub>(41, 14478)</sub>	0.00	0.00			
Race .....												
Time * Cohort *										1.8 <sub>(29, 22733)</sub>	0.00	0.00
Treatment .....												
Covariates												
SES .....	5518.2 <sub>(1, 22733)</sub>	0.00	0.20	5515.8 <sub>(1, 22691)</sub>	0.00	0.20	1356.7 <sub>(1, 14478)</sub>	0.00	0.10	5518.2 <sub>(1, 22733)</sub>	0.00	0.20
Intellectual												
Ability .....	1194.3 <sub>(1, 22733)</sub>	0.00	0.00	1197.1 <sub>(1, 22691)</sub>	0.00	0.10	671.3 <sub>(1, 14478)</sub>	0.00	0.00	1194.3 <sub>(1, 22733)</sub>	0.00	0.00

**Model 1-2. Gender Effect**

Model 1-2 investigates whether there is any interaction between program and gender effect. We modify the base model by adding gender as an additional factor and the interaction term gender\*cohort\*time. Table B-2 (Model 1-2) shows that the interactions between gender and program for different cohorts are statistically significant (F=1.6, p=.00). In other words, boys of one cohort reacted to the program differently than did the girls in the same cohort. The estimated mean scores presented in Table B-3 and the corresponding plots in Figure B-1 (1-6) help explain the gender effect. The elements of Figure B-1 suggest that the two lines representing each gender are not parallel. The strongest interaction occurred to cohort 1996 in year 1996-97 when the male group accelerated faster than the female group and in year 2001-02 when male performance improved again and female performance dropped continuously.

**Model 1-3. Race Effect**

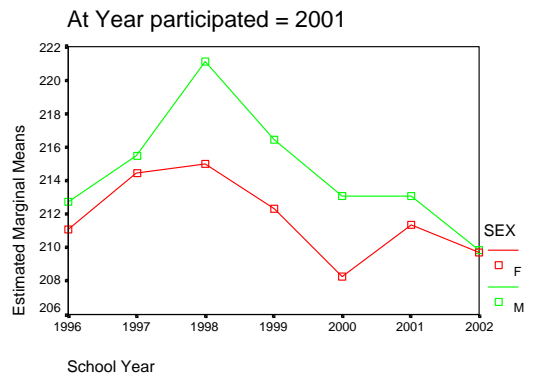
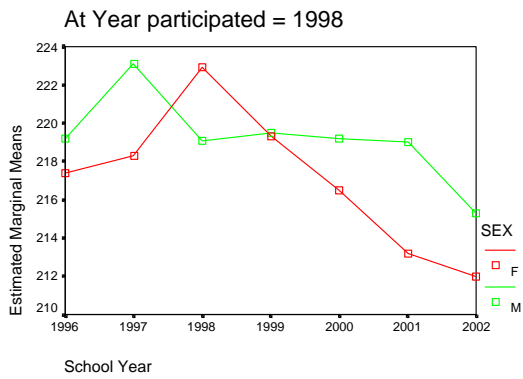
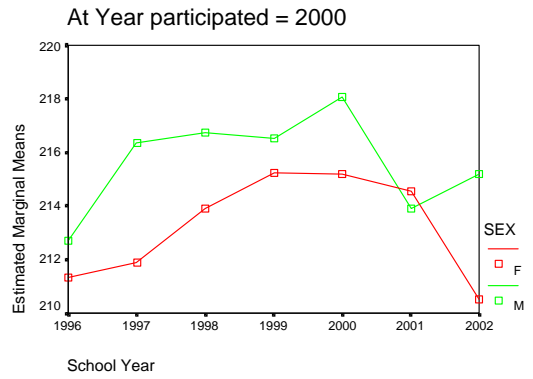
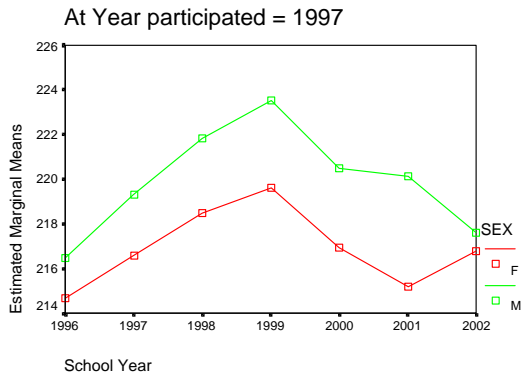
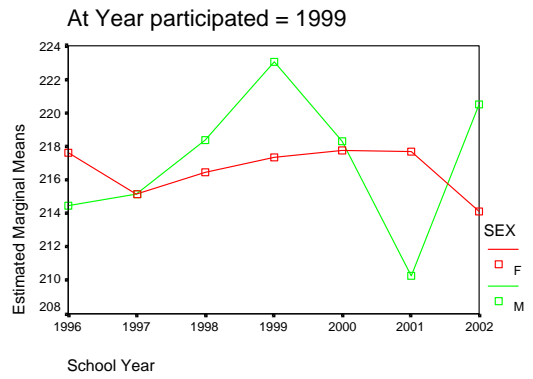
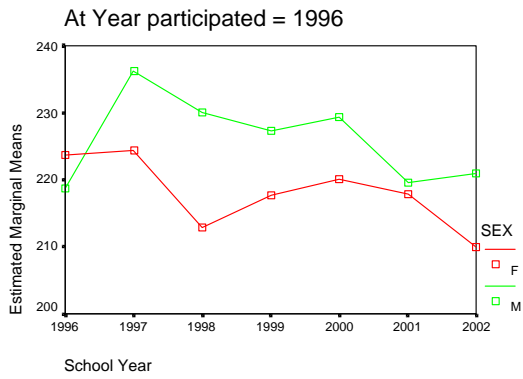
To investigate race effect, we modify the base model by entering race as a main effect and including its interaction with cohort and time. Since we would like to focus on African American students to illustrate the point, students of other ethnic groups are not included in this analysis.

The results (Table B-2, model 1-3) show that African American students of one cohort did not react to the program in the same way as the white students did in the same cohort (F=2.0, P=.00). The estimated means presented in Table B-4 and Figure B-2 (1-1) reveal that while the program affected the test scores of the white students positively in the first year of program implementation for most cohorts, the same effect did not occur to the corresponding cohorts of African American students. In fact, the program showed either trivial or negative effects on African American students of every cohort except for cohort 1998, where a positive impact was found after one year of program implementation.

**Table B-3.**  
**Estimated means of ITBS science scores by gender, cohort and time**

Year participated	School year	Female		Male	
		Mean	Standard error	Mean	Standard error
1996.....	1996	223.8	3.9	218.7	3.9
	1997	224.4	4.1	236.2	3.6
	1998	212.8	4.4	230.0	3.6
	1999	217.7	3.6	227.2	3.7
	2000	220.0	3.6	229.4	3.4
	2001	217.8	3.7	219.6	3.5
	2002	209.9	4.2	221.0	3.5
1997.....	1996	214.7	1.6	216.5	1.5
	1997	216.6	1.6	219.3	1.5
	1998	218.5	1.5	221.8	1.5
	1999	219.6	1.5	223.5	1.5
	2000	216.9	1.6	220.5	1.5
	2001	215.2	1.5	220.1	1.5
	2002	216.8	1.6	217.6	1.5
1998.....	1996	217.4	1.9	219.2	1.9
	1997	218.3	1.8	223.1	1.8
	1998	222.9	1.9	219.1	1.9
	1999	219.3	1.8	219.5	1.7
	2000	216.5	1.8	219.2	1.7
	2001	213.2	1.8	219.0	1.8
	2002	212.0	1.8	215.3	1.7
1999.....	1996	217.6	1.9	214.5	1.8
	1997	215.1	1.8	215.1	1.9
	1998	216.5	1.9	218.4	1.9
	1999	217.3	2.0	223.1	1.9
	2000	217.8	2.0	218.3	2.0
	2001	217.7	1.9	210.3	2.0
	2002	214.1	1.9	220.5	1.9
2000.....	1996	211.3	1.7	212.7	1.7
	1997	211.9	1.7	216.4	1.7
	1998	213.9	1.7	216.7	1.7
	1999	215.2	1.7	216.5	1.7
	2000	215.2	1.7	218.1	1.7
	2001	214.5	1.6	213.9	1.7
	2002	210.5	1.7	215.2	1.6
2001.....	1996	211.1	1.8	212.7	1.9
	1997	214.5	1.9	215.5	1.7
	1998	215.0	1.7	221.1	1.8
	1999	212.3	1.8	216.4	1.7
	2000	208.2	1.7	213.1	1.6
	2001	211.4	1.7	213.0	1.6
	2002	209.7	1.6	209.8	1.7

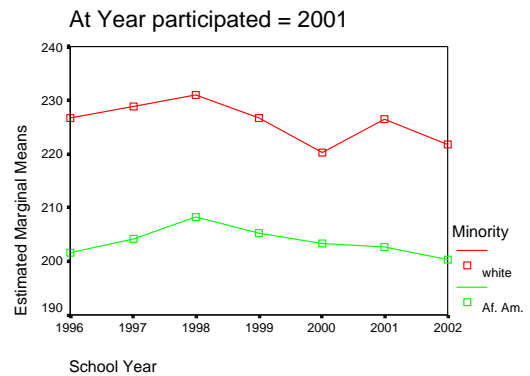
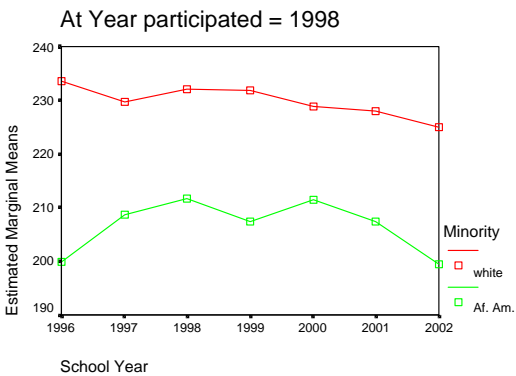
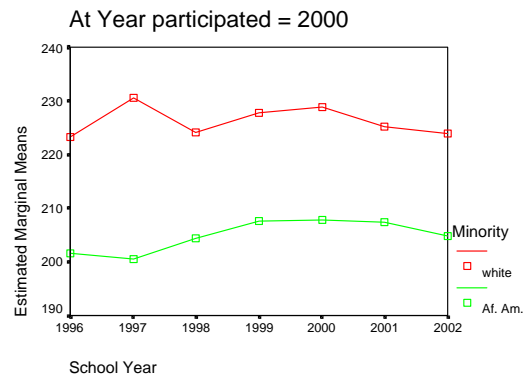
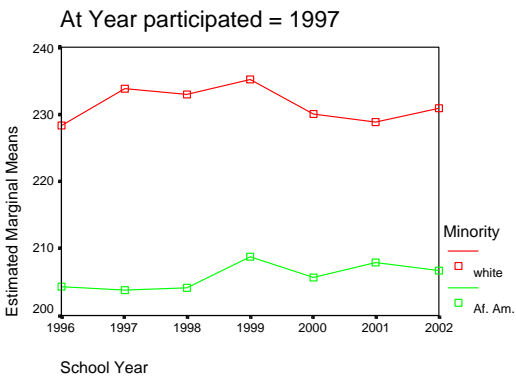
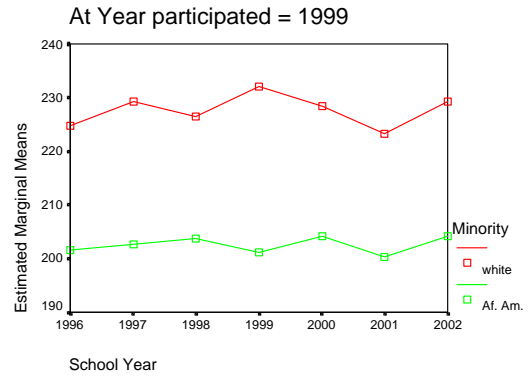
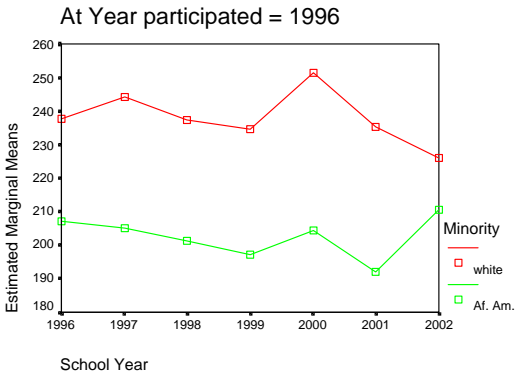
**Figure B-1.**  
**Estimated marginal means of science (gender effect)**



**Table B-4.**  
**Estimated means of ITBS science scores by race, cohort and time**

Year participated	School year	White		African American	
		Mean	Standard error	Mean	Standard error
1996.....	1996	237.7	3.9	207.1	5.3
	1997	244.1	3.6	205.2	6.7
	1998	237.5	3.8	201.4	6.5
	1999	234.7	3.6	197.2	6.4
	2000	251.5	3.7	204.4	5.6
	2001	235.4	3.5	192.1	7.3
	2002	225.9	4.2	210.7	5.9
1997.....	1996	228.3	1.5	204.3	2.3
	1997	233.9	1.6	203.8	2.4
	1998	233.0	1.5	204.1	2.3
	1999	235.2	1.5	208.8	2.3
	2000	230.0	1.5	205.7	2.3
	2001	228.9	1.5	207.9	2.3
	2002	231.0	1.5	206.7	2.3
1998.....	1996	233.5	2.0	199.8	3.4
	1997	229.7	1.8	208.6	3.1
	1998	232.0	2.0	211.6	3.1
	1999	231.9	1.9	207.4	2.9
	2000	228.8	1.9	211.6	2.9
	2001	227.9	1.9	207.4	2.9
	2002	225.0	1.8	199.5	2.7
1999.....	1996	224.8	1.9	201.6	2.9
	1997	229.2	1.9	202.8	2.6
	1998	226.4	2.0	203.6	3.1
	1999	232.1	2.0	201.1	3.0
	2000	228.4	2.1	204.2	3.1
	2001	223.3	2.0	200.3	3.3
	2002	229.3	2.0	204.1	3.1
2000.....	1996	223.3	1.8	201.5	2.3
	1997	230.6	1.9	200.6	2.4
	1998	224.1	1.8	204.4	2.4
	1999	227.8	1.9	207.5	2.5
	2000	228.9	1.8	207.9	2.6
	2001	225.1	1.7	207.4	2.5
	2002	223.9	1.8	204.9	2.3
2001.....	1996	226.7	2.0	201.5	2.1
	1997	228.8	2.2	204.1	2.1
	1998	231.0	2.1	208.3	2.0
	1999	226.8	2.1	205.3	2.0
	2000	220.3	2.1	203.3	1.9
	2001	226.5	2.0	202.7	1.9
	2002	221.8	2.1	200.2	1.8

**Figure B-2.**  
**Estimated marginal means of science (race effect)**



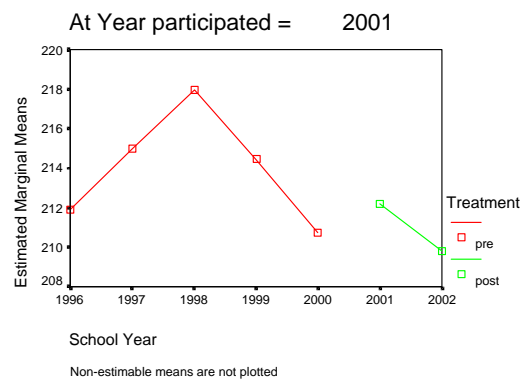
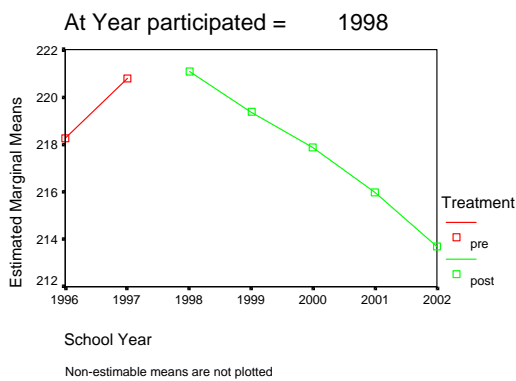
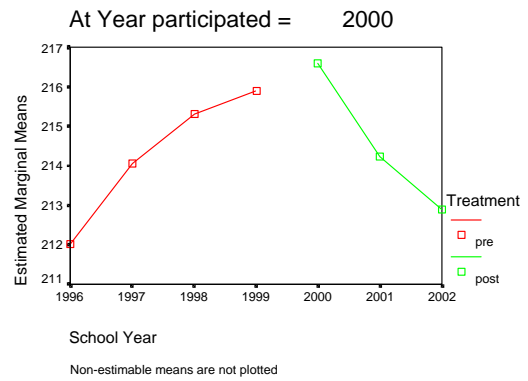
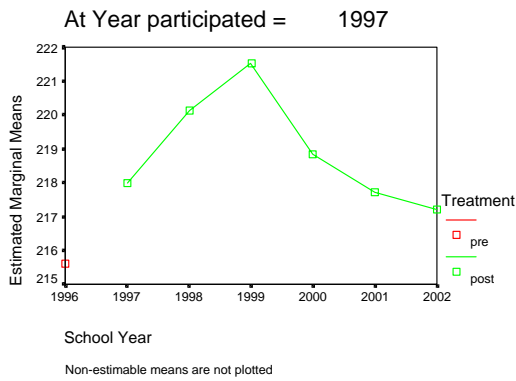
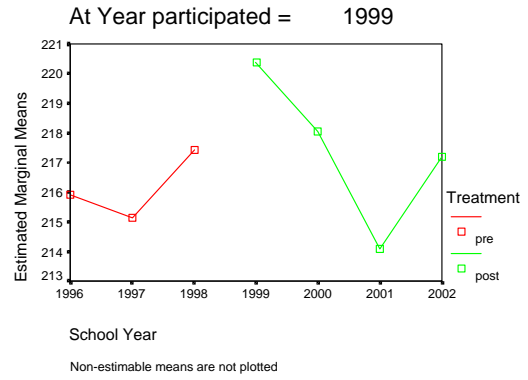
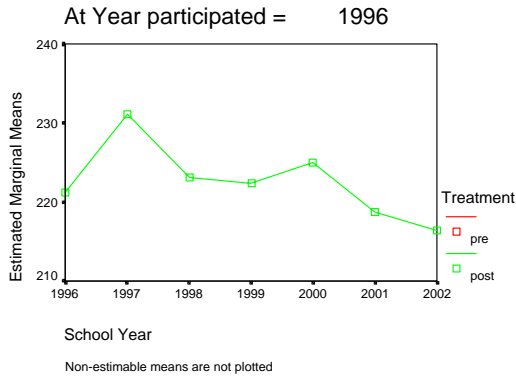
**Model 1-4. Treatment Effect**

To investigate whether there is any program effect on student test scores, a dichotomous variable “treatment” is created. Test scores collected before the school enrolled in the program are considered as pretreatment scores and those collected after the program are considered as post-treatment scores. Variables such as time and cohort of Model 1-1 are also included because of their statistical significance in the first analysis. A three-way interaction effect is added to capture the program effect across the years for different school cohorts. The results (Table B-2, Model 1-4) show that the three-way interaction term was significant ( $F=1.8$ ,  $P=.00$ ), indicating that the long-term program effects of the six cohorts were different. To explore the interaction further, we provide the estimated mean scores of each group and plots by cohort showing the “trend” of the science scores before and after program implementation in Table B-5 and Figure B-3 (1-6). The results suggest that the test scores of all five cohorts, with the exception of cohort 1996, which did not have pretest scores, increased modestly in the first year of the program implementation. For example, for cohort 1997, there was a 2.4-point rise in scores after one year of program. This is the difference between the first posttest score and the last pretest score. For Cohort 1987, posttest (218.0) minus pretest (215.6) equals 2.4. Similarly, the pre/post differences in science scores for cohorts 1998 to 2001 are 0.3, 3.0, 0.7, and 1.4 points respectively. However, the program effect seemed to diminish overtime, except for cohort 1997 where the post scores increased in three consecutive years from 1997 to 1999. The overall effect size was 0.002, indicating that the importance of these changes were practically trivial.

**Table B-5. Estimated means of ITBS science scores by time, treatment and cohort**

Year participated	School year	Pretest		Posttest	
		Mean	Standard error	Mean	Standard error
1996 .....	1996	.	.	221.2	2.8
	1997	.	.	231.1	2.7
	1998	.	.	223.2	2.8
	1999	.	.	222.3	2.5
	2000	.	.	225.0	2.4
	2001	.	.	218.7	2.5
1997 .....	2002	.	.	216.4	2.7
	1996	215.6	1.1	.	.
	1997	.	.	218.0	1.1
	1998	.	.	220.1	1.1
	1999	.	.	221.5	1.1
	2000	.	.	218.9	1.1
1998 .....	2001	.	.	217.7	1.1
	2002	.	.	217.2	1.1
	1996	218.3	1.4	.	.
	1997	220.8	1.3	.	.
	1998	.	.	221.1	1.3
	1999	.	.	219.4	1.3
1999 .....	2000	.	.	217.9	1.2
	2001	.	.	216.0	1.3
	2002	.	.	213.7	1.2
	1996	215.9	1.3	.	.
	1997	215.1	1.3	.	.
	1998	217.4	1.3	.	.
2000 .....	1999	.	.	220.4	1.4
	2000	.	.	218.0	1.4
	2001	.	.	214.1	1.4
	2002	.	.	217.2	1.4
	1996	212.0	1.2	.	.
	1997	214.1	1.2	.	.
2001 .....	1998	215.3	1.2	.	.
	1999	215.9	1.2	.	.
	2000	.	.	216.6	1.2
	2001	.	.	214.2	1.2
	2002	.	.	212.9	1.2
	1996	211.9	1.3	.	.
2002 .....	1997	215.0	1.3	.	.
	1998	217.9	1.2	.	.
	1999	214.4	1.2	.	.
	2000	210.8	1.2	.	.
	2001	.	.	212.2	1.2
	2002	.	.	209.8	1.2

**Figure B-3**  
**Estimated marginal means of science (treatment effect)**



### 3. School Participation (Multiple Regression)

In this section, two separate regression models will be used to analyze two datasets. The first analysis examines the impact of school participation in the LSC program on ITBS science scores. The basic analytical model can be explained as follows:

Model 2

- $Y = \alpha + \beta X + \delta Z$
- Y: ITBS scores (ITBSSC)
- X: school participation in the program (SPART)
- Z: control variables
  - students' gender (SEX), free/reduced lunch (FRL), ethnicity (ETH), living arrangement (LWP), gifted programs (GIFT), bilingual eligibility (BILG), ITBS math scores (ITBSSM), ITBS reading scores (ITBSSR)

- time

To investigate the overall program impact on student achievement, we choose ITBS science scores as the dependent variable. The independent variable SPART is a dummy variable (0,1) indicating whether a particular school was involved in the program at the time when the student took the test. For a school in cohort 3 that participated in the program in 1998, its SPART is coded as 0 from 1996-97, and 1 for the years 1998-2002. The control variables aim to isolate impacts from other confounding factors. These factors include measures of students' academic ability (gifted, scores in math and reading) as well as environment (SES and living arrangement). In addition, the time variable indicates the year when the data come from, effectively controlling for effect from maturation and other unexplained factors.

Table B-6 presents results from the first multiple regression analysis. The rows of the table contain different independent variables and the columns illustrate results from separate analysis using the

**Table B-6.**  
**Impact of school participation in LSC (SPS)**

vars \ DVs	ITBSSC			ITBSSC(BL)		ITBSSR		WASLR		ITBSSC	
	Unstd. Coeff.	Std. Coeff.	Sig.	Unstd. Coeff.	Sig.	Unstd. Coeff.	Sig.	Unstd. Coeff.	Sig.	Unstd. Coeff.	Sig.
Intercept.....	-8.73		0.79	-5.21	0.32	106.02	0	235.3	0	-8.30	0.81
SPART.....	1.84	0.03	0	1.18	0.16	-0.39	0.06	0.96	0	1.72	0
SEX.....	-3.18	-0.05	0	-2.52	0	2.32	0	3.51	0	-3.18	0
ETH.....	-1.55	-0.06	0			-0.48	0	-0.71	0	-1.55	0
BILG.....	-0.79	-0.01	0.05	-0.75	0.26	-11.59	0	-6.58	0	-0.81	0.04
FRL.....	-4.53	-0.06	0	-4.05	0	-4.49	0	-3.34	0	-4.52	0
GIFT.....	2.58	0.02	0	1.52	0.27	4.69	0	2.94	0	2.68	0
LWC.....	0.48	0.01	0.05	1.46	0.05	-0.43	0.02	0.77	0	0.46	0.06
GRD.....						-7.33	0.05	5.56	0.06	-0.39	0
ITBSSM.....	0.31	0.23	0	0.34	0	0.31	0			0.31	0
ITBSSR.....	0.76	0.57	0	0.70	0		0			0.76	0.95
ITBSSC.....						0.37	0				
WASLM.....								0.38	0		
CHT.....	-0.51	-0.03	0	9.6E-03	0.97	0.32	0.88	0.34		-0.49	0
IMP.....										0.95	0.04
R2.....	0.69			0.66		0.74		0.31		0.69	
F.....	4566			440		5632		1718		4186	
N.....	81028			14964		81028		64114		81028	

model. The results on the overall program effect (column 1) suggest that the program does have a positive impact on ITBS science scores ( $b=1.84$ ). This means that everything else being equal, a student from a school that participated in the LSC program would have 2 more points on average in ITBS science than he/she would have had the school not been in the program. Other control variables are significant and are in the expected direction. For example, while being on the lunch subsidy, an ethnic minority, and bilingual are negatively related to scores, living with both parents, being in the gifted program, and having high scores in other subjects (math, reading) are positively related to science scores. Being a female student has a negative association with science scores. The time variable appears to suggest a negative trend, which is consistent with findings from GLM.

While the unstandardized coefficients explain the substantive effects from the independent variables, standardized coefficients compare the relative impact among the identified variables. For each student, achievement in other subjects ( $\beta=0.23$  for math and 0.57 for reading) are the most important predictors of ITBS science scores. Eligibility for free/reduced lunch subsidy (-0.06), ethnicity (-0.06), gender (-0.05), and school participation (0.03) also explain some of the variances. The R square is 0.7, suggesting that the model explains 70 percent of the variation in student ITBS science scores.

We created a subset sample for African American students, who represent 20 percent of the students, in order to look at the program impact on minority students. If the school participation affects student learning overall, does it improve learning of African American students? Our analysis produces somewhat different results (column 2). Although the effects from other controls variables are similar, the program impact appears to be insignificant for African American students.<sup>19</sup>

The VIPS study reviewed in the previous chapter (section 3.12) suggests the spillover effect from science education on other subjects, especially

reading, for a school district with high percentage of minority students. The theory is that interactive-based instruction in science will improve students' reading skills. Our analysis shows mixed results. While the program does not seem to have a significant effect on students' ITBS reading scores (column 3), it seems to affect students WASL reading scores in a significantly positive way (column 4). The correlation in reading scores between ITBS and WASL appears strong and significant (0.71).

Various analyses in the first design suggest that the LSC program, measured by school participation, does have a significant positive effect on student achievement in science. This confirms the anecdotal evidence that the program has been successful in SPS. However, there are threats to the validity of the conclusion. First, SPS did not collect data on implementation, and different schools participated in the program in different ways. For example, to join the program, cohort 1-3 schools needed to show about 85 percent to 100 percent of teacher participation, but this requirement was relaxed in later years to allow enthusiastic teachers in reluctant schools the opportunity to participate in the program. Overall, the principal investigator observed,

For Cohort 1, the expectations were not clear as the program was just beginning and clear communications were not in place. In the second year, the expectations were more clear...by the third year, the expectations were very clear, teachers signed a Memorandum of Understanding that they were to complete 100 hours...this was the most excited and supportive group yet...the fourth and fifth year were the more reluctant schools. However, in each of these schools, there were pockets of enthused teachers and new teachers so there was never a black and white situation in almost any school.

---

<sup>19</sup>Results from other subpopulations such as Hispanics and Asian Americans are similar to that of the overall population.

In addition,

It is really hard to point out very active schools as most of the school's involvement changes from year to year depending on what else is going on, change in leadership, and change in teaching staff. Also, there are pockets in a number of schools where there is high involvement but where there are also teachers who are not so involved.

The principal investigator identified three top schools in terms of involvement. Regression results (column 5) based on this observation suggest that the quality of implementation does matter. Strong implementation appears to improve student achievement significantly.

Second, there is little evidence on the alignment of ITBS science test and the LSC program. The comparison of the program impact on reading between ITBS and WASL regarding the spillover effect suggests that the alignment and reliability are legitimate and important concerns. It is suspected that the impact of LSC program, which emphasizes on hands-on ability, might be more effective in improving the abilities better captured by criterion-referenced test such as WASL than norm-referenced test such as ITBS. If the reading scores are of any indication, one might expect to detect an even larger impact if student learning in science were measured by a well-aligned criterion-referenced test instrument.

### 3. Teacher Participation (Multiple Regression)

The second design analyzes data from 2000-2002. Compared with the first design, it has a closer link between LSC program and student outcomes, because teachers rather than schools are the direct recipients of the treatment. As shown in Model 3, the dependent variable is the same as the previous design. The independent variable is teacher participation, indicated by hours of training in the program. The control variables include student

characteristics and time, but also add teacher characteristics such as experience and education.

Model 3

$$Y = \alpha + \beta X + \delta Z$$

- Y: ITBS scores (ITBSSC)
- X: teacher participation hours (TPART)
- Z: control variables
  - students' characteristics: gender (SEX), free/reduced lunch (FRL), ethnicity (ETH), living arrangement (LWP), gifted programs (GIFT), bilingual eligibility (BILG);
  - teacher characteristics: experience (YRSEXP), education (EDUC);
  - time

Table B-7 contains results of teacher participation on students' ITBS science scores. Column 1 indicates an positive and significant effect ( $b=0.04$ ), which means that the longer teachers are involved in professional development activities, the higher their students score on ITBS science test. Substantively, the coefficient means that a completion of 100 hours professional development leads to an average 4-point increase in ITBS scores. Other results on the control variables are similar to the model on school participation (see Table B-7). In terms of teacher characteristics, neither teacher education nor experience is statistically significant.

The standardized coefficients show the relative impact of each identified variable. Again, achievement in other subjects is most important. However, teacher participation is just as important as other SES variables (lunch, gender, and ethnicity) in explaining the variances in ITBS scores. The adjusted R square is 0.68, meaning that the model explains 68 percent of variations in student scores.

Similar to the first design, we have conducted analysis on teacher participation on ITBS science scores from African American students (column 2)

and ITBS scores in reading (column 3). The results are the same as the first analysis.

Although the second design establishes a better nexus between program activities and student outcomes, it suffers, to a lesser degree, the similar drawbacks. For one, the quantity of professional

development cannot be translated into the quality of implementation by teachers, that is, teachers who took the same amount of professional development might apply what they learned in the classroom differently. In addition, concerns about the alignment and reliability of the test instrument remain.

**Table B-7.**  
**Impact of teacher participation in LSC (HSSPS)**

DVs	ITBSSC			ITBSSC (BL)		ITBSSR	
	Unstd. Coeff.	Std. Coeff.	Sig.	Unstd. Coeff.	Sig.	Unstd. Coeff.	Sig.
Intercept.....	1.50		0.72	31.74	0.01	69.43	0
TPART .....	0.04	0.05	0	-0.02	0.41	-0.008	0.14
SEX .....	-2.69	-0.04	0	-5.46	0.01	2.51	0
ETH .....	-4.58	-0.07	0			-4.38	0
BILG.....	-0.11	0	0.46	-4.81	0.06	-9.97	0
FRL.....	-4.13	-0.06	0	-7.91	0	-3.31	0
GIFT .....	0.45	-0.01	0.23	-2.20	0.54	2.61	0
LWC .....	1.58	0.02	0.02	0.74	0.38	-0.29	0.60
ITBSSM.....	0.27	0.21	0	0.37	0	0.32	0
ITBSSR.....	0.73	0.57	0	0.55	0		
ITBSSC.....						0.37	0
EDUC .....	0.0001	0	0.99	-1.40	0	0.15	0.09
YRSEXP.....	-0.03	-0.01	0.41		0.05	0.02	0.41
R2 .....	0.68			0.64		0.73	
F.....	602			57.2		773	
N.....	27393			2795		27393	