WCER Working Paper No. 2003-12
September 2003

Final Report on the Study of the Impact of the Statewide Systemic Initiatives: Volume II of Two Volumes

Principal Investigators

Norman L. Webb
Wisconsin Center for Education Research
University of Wisconsin–Madison
nlwebb@facstaff.wisc.edu

Iris R. Weiss
Horizon Research, Inc.
Chapel Hill, NC
iweiss@horizon-research.com

A Technical Report Prepared for
Bernice T. Anderson, Acting Division Director
Educational System Reform
Division of Research, Evaluation and Communication
Directorate for Education and Human Resources
National Science Foundation
Arlington, VA

Wisconsin Center for Education Research
School of Education • University of Wisconsin–Madison
http://www.wcer.wisc.edu/

WCER Working Paper No. 2003-12
September 2003

Study of the Impact of the Statewide Systemic Initiatives

Volume II: Final Report on the Use of State NAEP Data to Assess the Impact of the Statewide Systemic Initiatives

Norman L. Webb, Principal Investigator
Janet Kane, Jung-Ho Yang, Darwin Kaufman, Allen Cohen, Taehoon Kang, Chanho Park, and Linda Wilson

May 2003

A Technical Report Prepared for
Bernice T. Anderson, Acting Division Director
Educational System Reform
Division of Research, Evaluation and Communication
Directorate for Education and Human Resources
National Science Foundation

Wisconsin Center for Education Research
School of Education • University of Wisconsin–Madison
http://www.wcer.wisc.edu/

The research reported in this report was supported by the National Science Foundation under Grant No. REC–9874171 and by the United States Department of Education under Grant No. R902B020005 to the Wisconsin Center for Education Research, School of Education, University of Wisconsin–Madison. Any opinions, findings, or conclusions expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies, WCER, or cooperating institutions.

Table of Contents

Acknowledgements

Many people contributed to this project. We appreciate the excellent and persistent work of the researchers on the research teams who collected and analyzed data. This group of researchers was exceptional and enriched the project by being able to draw upon a wide range of methodologies. We received very thoughtful and timely advice from our advisory board early in the project, which helped to give focus to our work and better relate our study to the growing body of research on large-scale reform. The advisory board meetings were tremendous and intellectually stimulating due to the depth of experience and knowledge of our advisors. We were fortunate to have very capable and articulate staff members to edit our work, to format pages, tables, and graphs, and to construct the Web pages. Finally, we appreciate greatly the support we received from the National Science Foundation and project officer Bernice Anderson. She gave us excellent feedback and encouraged us to extend this study to its limits.

Advisory Board

Dennis Bartels, TERC, Cambridge, Massachusetts
Audrey B. Champagne, SUNY, Albany, New York
Jere Confrey, University of Texas
Andrew C. Porter, University of Wisconsin

NSF Program Officer

Bernice Anderson

Horizon Research Staff

Iris R. Weiss, Co-Principal Investigator
Daniel J. Heck, Senior Research Associate
Jonathan Supovitz, Consultant, University of Pennsylvania
Sally E. Boyd, Researcher
Michael N. Howard, Researcher
Susan Hudson, Secretary

WCER Research Staff

Norman L. Webb, Co-Principal Investigator

Daniel Bolt, Consultant
Allen Cohen, Consultant
Janet Kane, Research Scientist
Michael Kane, Consultant
Taehoon Kang, Program Assistant
Darwin Kaufman, Research Scientist
Jong-Baeg Kim, Program Assistant
Chanho Park, Program Assistant
Linda Wilson, Research Scientist
Jung-Ho Yang, Research Scientist

Lynn Lunde, Program Assistant
Mary Tedeschi, Administrative Assistant
Margaret H. Powell, Editor

All of the authors contributed to the total document, but major authors of specific chapters were as follows:

Chapter 1 Introduction, *Norman L. Webb*
Chapter 2 Project Goals, *Norman L. Webb*
Chapter 3 Review of Recent Studies, *Darwin Kaufman*
Chapter 4 SSI and Non-SSI Achievement Using the State NAEP Descriptive Analysis, *Jung-Ho Yang, Chanho Park and Norman L. Webb*
Chapter 5 Reform-Related Changes in Educational Practices in SSI and Non-SSI States, Janet Kane
Chapter 6 Using Supplementary Information in Evaluating Statewide Systemic Initiative Reform, *Janet Kane*
Chapter 7 Contrast of State NAEP with Three State Assessments, *Darwin Kaufman*
Chapter 8 SSI and Non-SSI Achievement Using State NAEP Data: Empirical Bayes and Bayesian Analyses, *Jung-Ho Yang*
Chapter 9 A Comparison of SSI and Non-SSI Performance on State NAEP Mathematics Assessment Items, *Chanho Park and Norman L. Webb*
Chapter 10 Conclusions, *Norman L. Webb*

Additional information and data can be found on the following web site:
http://www.wcer.wisc.edu/ssi

List of Tables

List of Figures

Executive Summary

In an effort to evaluate the impact of the SSIs on student achievement and the lessons that could be learned from the National Science Foundation's effort to reform mathematics and science education on a statewide basis, research studies identified the technical strategies, the political strategies, and the interactions with funders that were critical factors in the attempt to effect significant change in student learning over large populations. Documents were received on 21 of the 26 SSIs. More intensive data were collected via telephone interviews of key personnel in seven of these states and during site visits in six other states. Among a number of lessons learned were the following: it was vital to incorporate enough flexibility within the design so that information produced by research, evaluation, and monitoring could be effectively used—technical lesson; the creation of partnerships with policy organizations significantly advanced policy work—political lesson; and, SSI leaders and funders needed to develop a shared, in-depth understanding of the reform strategies as these fit the local context—interaction with funders.

In addition, an analysis of the NSF's systemic initiatives compared student mathematics test data for grades 4 and 8 in SSI states and non-SSI states with data from State NAEP assessments for three testing years, 1992, 1996, and 2000. Comparisons were made of 14 SSI states and 13 non-SSI states that participated in the State NAEP in each of these three testing years.

The close fit found between improved performance and SSI funding suggests that a relationship exists between such initiatives and student achievement. Of equal importance is the finding that change is most effective when multiple components are addressed in concert: i.e., when the SSIs served as catalysts for other reform efforts that states had initiated, they achieved optimum impact. When state policies are aligned with the goals of a systemic initiative and when state infrastructure supports teachers and schools as they change their practices, reform can result in substantial achievement gains in a relatively short time.

Summary

Study of the Impact of Statewide Systemic Initiatives

Early in the 1990s, the National Science Foundation embarked on an ambitious mathematics and science education reform effort that continued throughout the decade. Based on a commitment to systemic reform, the Statewide Systemic Initiatives (SSI) Program provided funding to qualifying states that enabled them to make simultaneous changes in multiple components to achieve improved student learning of challenging content. Over the decade, 25 states and the Commonwealth of Puerto Rico received millions of dollars from NSF for up to five years. Eight of the jurisdictions received funding for a total of ten years. The present study mined existing performance data from the National Assessment of Educational Progress (NAEP) and acquired new data from SSI leaders for evidence of the impact of the SSI Program on student learning and to determine what could be learned about the strategies, policies, and activities that were needed to advance large-scale reform. The study was driven by two main research questions: 1) What lessons have been learned about designing, implementing, evaluating, and supporting statewide systemic reform? 2) What differences were there on mathematics achievement as measured by NAEP between SSI states and non-SSI states over the period 1992 to 2000? The methods and findings related to these two questions are presented below.

Lessons Learned

In the qualitative analyses, external documents were reviewed on 21 of the 26 SSI jurisdictions. Internal documents produced by the SSIs were reviewed for 13 states, with telephone interviews conducted of key personnel in seven states, and site visits and more intensive interviews in the other six states. Data were analyzed using topical and thematic coding schemes that examined the technical and political strategies of the jurisdictions studied. The major conclusions of the study and the lessons learned were derived from cross-case analysis.

The lessons learned are derived from technical strategies and demands, political strategies and demands, and interactions with funders. These three areas are not considered independent, but identify three important functions that all of the SSIs faced. The SSI leaders learned a great deal about statewide systemic reform as a part of the enactment of reform in their states through the SSI program. Many of these lessons were similar across SSIs, and both positive and negative examples supported the lessons learned.

Technical strategies. Technical strategies are interventions needed to bring about those changes in teaching and learning that result in improved and more equitable student achievement. A sound technical strategy is one that:

- Operationalizes the reform vision through interventions;
- Monitors and refines the interventions, and provides evidence that they result in improved classroom practice and improved student outcomes;

- Increases capacity within the state to scale up the reform efforts; and
- Provides evidence that quality and impact are maintained during scale-up.

Lessons related to technical strategies and demands were:

1. Beginning with a manageable scope and scale in design was beneficial.
2. Establishing data systems to monitor progress, measure impact, and assure quality research was at least as important as focusing resources on scaling-up the interventions.
3. Designing interventions that incorporated the flexibility to effectively use what was learned through research, evaluation, and monitoring was vital.
4. Many kinds of capacity need to be developed to initiate and continue reform.
5. Developing a feasible plan for scaling-up within a reasonable time frame was important; scaling-up too quickly can become problematic.
6. Creating a healthy tension between capacity building for scale-up and the achievement of quality control was absolutely necessary.

Political strategies. Political strategies describe how the SSIs envisioned, and set to work, establishing a supportive context for reform. A sound political strategy is one that:

- Facilitates development of formal policies that provide guidance and incentives for the reform vision;
- Cultivates broad understanding of and support for the reform vision; and,
- Increases school and district leadership commitment to reform.

Lessons related to political strategies and demands were:

1. Housing SSIs within, or forming a partnership with, policy organizations positioned many SSIs to engage in policy work.
2. Involving education policy makers as leaders or partners of an SSI situated the initiative to be a natural contributor to policy decisions related to mathematics and science education.
3. Providing an "existence proof" of high quality, valued service, and contribution in one policy area often expanded the role of the SSI in a state's wider education policy arena.
4. Establishing the understanding and support of mathematics and science leaders in the state and education leaders in general increased the likelihood that the SSI would become a player by being informed and consulted.
5. Nurturing relationships with regional and local leaders and stakeholders, including superintendents, principals, curriculum supervisors, curriculum committees, and parents, was needed to reach schools and classrooms.
6. Establishing neutral political turf to bring constituencies together benefited the initiatives significantly. Maintaining a connection while staying at a reasonable distance from existing agencies enabled some initiatives to convene a broad array of stakeholders.

7. SSIs had to balance taking credit and sharing credit for successes with collaborators and other reform actors in their state.
8. The need and opportunity existed to develop new, expanded leadership for mathematics and science education reform in order to expand the reform statewide and to sustain the effort into the future.

Managing interactions with funders. The SSI Program launched a series of NSF programs that evolved out of a commitment to provide challenging, meaningful science and mathematics education to all students through changes in whole systems of education. It also pioneered new relationships between NSF and program awardees in the form of cooperative agreements. The critical aspects of how initiative leaders managed their interactions with NSF included:

- Developing a shared understanding of the strategy for reform;
- Negotiating appropriate changes to the design; and,
- Making a case that their initiative was having the desired impact.

There were three important lessons for managing interactions with funders and sustaining these relationships in the future:

1. The SSI leaders and the funders needed to develop a shared, in-depth understanding of the reform strategy as it fit the local context.
2. Appropriate changes in reform design needed to be negotiated through a shared understanding between the initiative leaders and funders, with careful attention to the trade-offs and balances associated with these changes.
3. A shared understanding of the reform strategies, expected impacts over time, and long-term outcomes of the initiative was needed to guide the collection, interpretation, and reporting of appropriate evidence.

Findings from the NAEP Analysis

State NAEP student mathematics achievement data from 1992, 1996, and 2000 and teacher-reported information on classroom practices from 1992 and 1996 were analyzed to compare student performances and practices in 14 SSI states and 13 non-SSI states. The states included in the sample analyzed were all states that had participated in the State NAEP assessments during the three years of the study. The states selected, although not randomly chosen, represented a cross-section of the states in each group and had characteristics (average public school enrollment, per capita expenditure, percent of White students enrolled, and 1992 average mathematics achievement) similar to those of all of the states in their respective groups.

A variety of analytic approaches was used with the State NAEP data to compare and contrast the SSI and non-SSI states. Overall achievement, as well as performance for population subgroups, was described via means and mean differences between groups. Hierarchical linear modeling was used to estimate rates of growth for each group across 1992, 1996, and 2000. Performance differences on individual NAEP items were

identified using differential item functioning. Scales of reform-related instructional practices were constructed from items on the NAEP teacher questionnaire, and changes over time were examined with regression methods. Qualitative methods were used to identify differences among the SSI states and to relate these features to achievement gains. Finally, results of state assessments in three SSI states were compared to the results of State NAEP for those states. The paragraphs below summarize the findings from each of these studies.

Descriptive analyses. The average composite mathematics performance of students at grades 4 and 8 in the 14 SSI states and the 13 non-SSI states remained nearly comparable over the eight years of the study, 1992 to 2000. During this time period, both groups improved on the average by 6 scale points at grade 8 and by 6.5 scale points at grade 4. In 1992, grade 8 students in the SSI states averaged slightly lower than those in the non-SSI states (1.2 scale points difference). All other differences between the two groups of states by year and grade were .7 scale points or less. Thus, there were no differences between the SSI states and the non-SSI states in average mathematics composite scores for each of the three testing times, except for grade 8 in 1992.

In 1992, there was considerable variance in mean performance among the states within the SSI group and within the non-SSI group: both groups included high-performing states and low-performing states. Over the eight years, the variance in mathematics performance among states decreased. Empirical Bayes and Bayesian Analyses confirmed that the average mathematics performance by SSI states at both grades 4 and 8 began below that of the non-SSI states in 1992, but increased at a faster annual growth rate, although not statistically significant, than the non-SSI states.

Population subgroups. There were no differences by gender between SSI states and non-SSI states in mathematics performance over the eight years from 1992 to 2000. There is evidence that Black students in SSI states made relatively higher gains than those in non-SSI states between 1992 and 1996. Hispanic students in SSI states made relatively higher gains than those in non-SSI states between 1996 and 2000. This evidence is apparent in both cross-sectional data and growth in performance from grade 4 to grade 8 by the same cohort of students. Black students in SSI states made a slightly higher gain from grade 4 in 1992 to grade 8 in 1996 than did White students on the Number/Operations strand and the Algebra/Functions strand. Over those four years, the performance of White students and Black students from SSI states was more similar than for these two populations in non-SSI states. However, over the next four years this trend was reversed, with the gap between White students and Black students in 2000 being smaller for non-SSI states than for SSI states. For Hispanic students, the finding was reversed. The gap between growth in performance from grade 4 to grade 8 of White students and Hispanic students was less for SSI states than for non-SSI states between 1996 and 2000 and was greater between 1992 and 1996.

Differential item functioning (DIF). Some differences in the underlying mathematical constructs of the performance of students from the SSI states compared to those from the non-SSI states were detected. At both grades 4 and 8, when the performances of students

with equal abilities were compared, students from SSI states performed higher on items from the Data Analysis content strand and items requiring problem solving. Both item types represent areas that have been emphasized in reform mathematics over the 1990s. Students from SSI states also performed better on an increasing number of multiple-choice items from 1992 to 2000. This finding, along with a reduction in the number of DIF items in more widely covered content strands of Number/Operations and Algebra/Function, indicates that students from SSI states improved in performance in relation to students from non-SSI states who were of equal ability both on a greater number of reform topics and on more traditional measures.

Scales of reform-related instructional practices. Classroom practices in SSI states incorporated a greater number of reform practices than classroom practices in non-SSI states. We analyzed six reform indicators—three on classroom practices and three on teachers' knowledge and professional development. As expected, SES (socioeconomic status) was the primary predictor of a states' mean mathematics composite in both 1992 and 1996. SSI states averaged significantly higher on an indicator of the relative emphasis on reasoning and communication at both grades 4 and 8 in 1996. In a regression model, this indicator was predictive of the mean State NAEP mathematics scores. Teachers in SSI states compared to those in non-SSI states reported giving students more opportunities for mathematical discourse in both 1992 and 1996. However, the difference was not significant. The use of criterion-referenced tests (CRT) in at least two grades from grades 3 through 8 was used as another variable for describing reform within states. Both SSI status and CRT were related to achievement gains across the three State NAEP administrations (1992, 1996, and 2000). The gains were the largest among states with criterion-referenced tests.

Use of qualitative research to analyze NAEP performance. We employed qualitative methodology to understand more fully what might explain the differences in performance by the 14 SSI states in the longitudinal trend sample. In addition to the mathematics performance data from the State NAEP, we used data from a number of sources. To gather more information on the independent variable for the time covered by the assessment data, we interviewed state mathematics supervisors and SSI leaders; we also reviewed documents to provide information on the percentage of teachers in the state reached by reform in 1990, 1992, and 1996, emphasis given to components of reform, and relative emphasis given by the SSI on the five State NAEP content strands. Information from these SSI state reports was supplemented with data from a policy analysis of state SSIs, evaluations of the SSI program by other analysts, and annual surveys of state student assessment programs. We also consulted the Horizon Research team on their findings from the SSI states in the trend sample. Using these data as a basis for our analysis, we divided the SSI states into three groups based on State NAEP mathematics performance for the three testing times in both grades 4 and 8—Steady Increase, Some Increase, and Little/No Increase.

Overall, our findings are consistent with the underlying theory of systemic reform. State assessments and accountability policies appear to be strong factors in improved student performance. Furthermore, we found that state policies aligned with the goals of a

systemic initiative, along with a sufficiently strong statewide infrastructure to support teachers and schools as they change their practices, can result in substantial achievement gains in a relatively short time. More specifically, we found:

- Statewide achievement gains across four years were more likely to be evident when reform efforts addressed state policy as much as or more than teachers and classroom practices.
- Statewide assessment policies and practices seemed to be important components of systemic reform. The existence of a state assessment program seems to be related to statewide achievement gains, particularly when criterion-referenced tests were used.
- There is some indication that when state policies were not supportive of SSI goals, reform efforts were compromised or even undermined.
- When assessments were aligned with the goals of the SSI, reform-related instructional practices increased; when they were not aligned, reform-related instructional practices did not change, or decreased.
- States with a strong infrastructure prior to the SSI generally had steady gains in achievement. States with large increases in reform indicators during the SSI were able to steadily increase achievement with Phase II funding. Achievement gains from 1996 to 2000 were unlikely to occur in SSI states that did not receive Phase II funding.
- In all SSI states, the alignment of state frameworks and assessment with the SSI goals appeared to be an important influence on statewide student achievement.

State assessments and State NAEP. The framework for State NAEP assessments in 1990 through 2000 was designed to use a number of sources that included state and district standards and the National Council of Teachers of Mathematics *Curriculum and Evaluation Standards for School Mathematics*. Although the State NAEP provides information on a range of mathematics performance, it was not designed for precise measurement of curriculum standards and frameworks from any one state or reform in mathematics in any one state. Students also do not have the same motivation to perform on the State NAEP as they do on state assessments where the results have some meaning to them. The assessments designed and administered by the state should be in a better position to do this. We conducted a focus study comparing results from assessments administered by three of the SSI states—Texas, Maine, and Massachusetts—using the State NAEP results to verify the findings attained from the State NAEP and as a basis for closer inspection of the relationship between an SSI intervention among schools within a state and mathematics performance.

In trying to use data from state assessments, we were confronted with a number of issues that included change in the state assessments over the time period, lack of year-to-year data, and insufficient documentation of data needed for longitudinal analyses. Our findings were mixed. For the time period between 1992 and 1996, the State NAEP results and the state assessment results for Texas and Massachusetts were comparable. However, the Maine state assessment for grade 4 showed improvements that were not apparent on the State NAEP. For the 1996 to 2000 time period, only state assessment data from Texas

could be used in our analysis. During this period, both State NAEP and the state assessments indicated some improved performance, but the Texas state assessment indicated substantially more improvement than did the State NAEP, similar to the previous results for Maine. Even though the state assessment scores in Massachusetts from 1992 to 1996 showed little gain, the cohort of schools with the most intense SSI involvement over this period did show improved scores. Thus, the State NAEP can be sensitive to some large group changes in performance as verified by state assessments, but is less sensitive to more subtle effects when reform efforts target specific subpopulations.

While state assessment data proved to have potential for the study of reform efforts within states, it was determined that continuity of content and program design is essential for such studies. Furthermore, test designs that reflect the knowledge, skill and cognitive development of disciplines as well as psychometric scales that allow for adequate measurement of growth are required if these state assessments are to detect achievement improvements over time.

Conclusions

We did not design this research to be, nor did we have the resources to conduct, the definitive study on the impact of the Statewide Systemic Initiatives. However, the study supports the finding that a tremendous amount of learning about how to engage in large-scale reform took place over the duration of the SSI program and that in states having an SSI, we found an increased rate of learning by students. The findings in this study have produced a number of lessons learned that are directly applicable to any attempt to make significant changes in student learning over a large population. We learned that it is not only critical to consider the technical issues concerning the functioning of a program, but it is essential to address the political decisions within the state and negotiations with the funder in order to garner the support necessary to sustain an effort long enough for a measurable impact on student learning to be achieved.

It was impossible in this study to isolate the specific impact of an SSI on student learning. When SSI states were studied as a group and compared to non-SSI states, there was evidence that student scores from 1992 to 1996 to 2000 in SSI states increased at a faster rate than did student scores in the non-SSI states. The variation among SSI states was as great as among non-SSI states. It was clear that SSI states with Phase II funding accelerated the rate of learning over the time period from 1996 to 2000, whereas the SSI states that did not receive continued funding and the non-SSI states as a group maintained or declined in the rate of learning over this time period. The close fit between improved performance and SSI funding suggests a possible relationship between the statewide systemic initiatives and student performance, but it was impossible to discount other alternative hypothesis including a selection bias. It was also clear that teachers in SSI states were using a greater number of reform practices than those in non-SSI states and that students from SSI states were performing more favorably on those mathematics areas that were given greater emphasis in reform mathematics curricula—Data Analysis and Problem Solving. The findings from this study are very compatible with the theory of

systemic reform and the need to change multiple components in concert rather than independently. The findings are consistent with NSF's vision that the SSIs serve as catalysts for other reform efforts in states. Those states with a more developed infrastructure prior to the SSI were able to take greater advantage of the SSI funding. Overall, the SSI program was related to an increased rate of student performance in some states. The variation in improved student performance among SSI states appeared to be related to prior conditions in their education systems, accountability, and duration of funding.

A number of methodologies were used to complete this study. The SSI leaders interviewed were a deep source of information on implementing large-scale reform. The State NAEP proved to be a viable source of data that could be used to study differences among states and compare SSI states with non-SSI states. Even though only about 60% of the states in each of these groups participated in the State NAEP for 1992, 1996, and 2000, the states that did participate were representative of the larger groups. An important condition for this study was our capacity to secure data and information on the nature and the quality of the SSI implementation. For this information, we drew heavily upon the work of researchers who received funding from NSF to describe and analyze the implementation of the SSI program—SRI, the National Institute for Science Education, RAND, the Council of Chief State School Officers, the Consortium for Policy Research in Education, COSMOS, and Abt Associates. The study and its findings were greatly enhanced by combining both qualitative and quantitative methodology.

CHAPTER 1

INTRODUCTION

At the outset of this study of the National Science Foundation's (NSF) Statewide Systemic Initiatives (SSIs), our primary intent was to use existing data to measure the impact of the SSIs on student achievement and to determine what could be learned about designing and implementing these major state initiatives on the basis of document analyses of SSI proposals, interim reports, and participants' reflections. As our work evolved, it became apparent that it was equally important to address various approaches to the study of large-scale educational change. Documenting what we sought to do and could not do because of the lack of relevant data, the size of the problem, or the lack of adequate analytic procedures became as important as distilling results from the National Assessment of Educational Progress (NAEP) and interviews of key people in selected sites. In order to produce meaningful findings that could be associated with state efforts supported by NSF funding, it also was important for us to think differently about the analysis of the masses of data that were available. In the process of doing this, we employed a number of analytic techniques and different means of reporting as we developed our case for the impacts of the SSIs and the conclusions that could be reached about these very ambitious efforts for reforming K-12 mathematics and science education on a statewide basis.

In the early 1990s, NSF made a massive effort to improve mathematics and science education in the states. NSF sustained its funding of the SSI program for over a decade, expanding from a statewide systemic initiatives model to a program designed to include urban systemic initiatives (1993), rural systemic initiatives 1994), and local systemic initiatives (1995)—e.g., the Comprehensive Partnerships for Minority Student Achievement Program. Through the SSI program, a total of 26 jurisdictions, 25 states and Puerto Rico, were each awarded up to $10 million over five years. These awards were made in three cohorts:

1991 Cohort (N = 10):
Connecticut, Delaware, Florida, Louisiana, Montana, Nebraska, North Carolina, Ohio, Rhode Island, and South Dakota

1992 Cohort (N = 11):
California, Georgia, Kentucky, Maine, Massachusetts, Michigan, New Mexico, Texas, Vermont, Virginia, and the Commonwealth of Puerto Rico

1993 Cohort (N = 5):
Arkansas, Colorado, New Jersey, New York, and South Carolina.

During the initial five-year period, NSF withdrew its funding from four states—Florida, North Carolina, Rhode Island, and Virginia—because the agency judged that they were not fulfilling the full intent of the program. After the first five years of funding, SSI states

were permitted to apply for another five years of funding under a second phrase of the program. NSF awarded Phase II funding to eight states—Connecticut, Louisiana, Massachusetts, New Jersey, Puerto Rico, South Carolina, Texas, and Vermont.

States varied in the strategies they adopted to attain systemic reform. Nearly all states claimed that mathematics and science were a major focus. Eleven Phase I states focused on grades K through 16. Another six focused on grades K through 12. The other states concentrated their initiatives on the primary or middle grades. Only the Montana SSI addressed primarily high school. Eighty percent of the SSIs developed a strategy for supporting teacher professional development and approximately 90% had a strategy for creating an infrastructure for capacity building among teachers, schools, or institutions, the two most common approaches to change (Zucker, Shields, Adelman, Corcoran, & Goertz, 1998). Other strategies identified by the SRI International evaluation included developing, disseminating, or adopting instructional materials (13 SSIs), supporting model schools (7 SSIs), aligning state policy (16 SSIs), funding local systemic initiatives (9 SSIs), reforming higher education and the preparation of teachers (13 SSIs), and mobilizing public and professional opinion (14 SSIs) (Zucker et al., 1998).

State educators who engaged in systemic reform encountered a major challenge in their efforts to design programs that addressed statewide reform. Systemic reform required strategic thinking about the technical aspects of reforming mathematics and science teaching and learning; it required planning to go to scale within a state, and a commitment to sustain reforms over time. In their initial efforts, SSIs had too much to attempt in delivering needed services to districts, schools, teachers, and students. Although a great deal was learned about which initiatives proved especially promising or productive, leaders of SSIs found that in the first five years critical choices had to be made. The SSIs had to balance attention to direct services with attention to building infrastructure and capacity and to reforming state policy systems that would support changes in teaching and learning. Moreover, whether delivering services, building infrastructure and capacity, or working in the policy arena, the SSIs required keen political strategizing in order to make systemic reform work (Heck, Weiss, Boyd, & Howard, 2002).

In the ten years since the initial funding of the SSIs, states have been engaged in a range of reform activities. The number of states that developed curriculum standards dramatically increased from a handful in 1992 to 49 in 2002 (Mid-Continent Research for Evaluation and Learning, 2003). The number of states with assessment systems that incorporated some student accountability measures increased to 30 by Spring, 2000 (Council of Chief State School Officers, 2001). A total of 44 states instituted some form of school accountability by the end of the 1999-2000 school year. In addition to statewide changes, school districts and schools engaged in a range of reform activities such as adopting new curricula, increasing professional development for teachers, building learning communities among staff, and incorporating more hands-on learning activities for students. In states with SSIs, these other reform activities may or may not have been related to NSF funding. The data included in this part of our analysis were insufficient to determine fully which results could be attributed to specific programs and initiatives.

As the National Science Foundation embarks on another large-scale reform effort through the funding of mathematics and science partnerships (MSPs) in 2003, the findings from the present study have much to offer regarding the planning and construction of the evaluations of the new initiatives. Although it will be inappropriate to use State NAEP data in evaluating the MSPs because of their focus on school districts rather than on states, the present study demonstrates the importance of employing comparable measures of achievement in a multisite study. Because the same assessments were used to measure student achievement in a large number of the states, it was fairly straightforward to make comparisons among the SSI states as well as between the SSI and non-SSI states, using the state as the unit of analysis. When we compared NAEP scores with scores produced from states' own assessments over the same time period (see Chapter 3), it became apparent that the State NAEP results were not consistent with state assessments. This implies, for several reasons, that state mathematics assessments clearly vary when compared to the NAEP assessment—in part, because they measure somewhat different content and offer varying incentives for students to do their best work. Even when comparing results from different instruments, using effect size based on standard deviation units can be problematic and may not account for all of the variation in performance between comparison groups.

Measures of Classroom Practices and Teacher Knowledge

Up until 1996, the State NAEP database included 1) reports by teachers and students on classroom practices and 2) information about schools. This made it possible to create several indicators of reform practices that could be used to document any statewide changes over time, as well as to identify differences in practices among the states. The State NAEP in 2000, however, did not include teacher questionnaires as detailed as those administered in 1990, 1992, and 1996. This lack of comparability among the questionnaires prohibited us from continuing the analysis using classroom practices for the 2000 data. Because data on classroom practices accompanied the student assessment data for the 1990, 1992, and 1996 State NAEP, we have been able to probe the data more deeply in our effort to explain variations in student performance. This clearly points to the need to collect data on classroom practices concurrently with student achievement data, an opportunity that rarely exists when using state or district assessment data. In evaluating the MSPs, it will be important not only to have data on classroom practices along with data on student achievement, but also to document or monitor the participation of teachers and others in MSP activities. Our analysis of the State NAEP did not include data on SSI activities, which prohibited us from doing more precise investigation on differences in implementation and emphasis by the different states beyond considering only general statewide changes. We did collect, post hoc, some information on content emphasis by the SSIs as reported by key individuals, but this level of information was insufficient as a basis for fully analyzing the relationship between student performance and SSI activities. We also used data from other sources to describe the nature of each SSI (Zucker et al., 1998; Clune, 1998), but, although informative, these data could not be directly related to the State NAEP timeframe.

Vertically Scaled Data

All State NAEP student achievement data are reported on the basis of a 500-point vertical scale. Use of a common scale is a desirable feature that enabled us to consider change in performance, by cohort, of students from grade 4 to grade 8. In particular, analysis of growth in performance by a cohort of students presented findings that distinguished the performance of different ethnic groups, findings not produced in other analyses. Although racial groups varied greatly, by more than 30 points in the mean performance at grade 4 and grade 8, the growth in mean scores from grade 4 to grade 8 by the different ethnic groups varied by less than eight points. With a vertical scale, it is evident that much of the difference in performance among White, Black, and Hispanic students existed at grade 4 and only slightly increased between grade 4 and grade 8. This implies that differences in performance among racial groups are apparent at school entry in the early grades and that the achievement gap among these groups only slightly increases from grade to grade.

Reporting Achievement by Content Topics

Another desirable feature of the NAEP data is that information is reported by content strands for a content area. In mathematics, student scores are reported for Number Sense, Properties, and Operations; Measurement; Geometry and Spatial Sense; Data Analysis, Statistics, and Probability; and, Algebra and Functions (National Assessment Governing Board, undated). This feature was helpful in discerning a relationship between the assessment results and the systemic initiatives by relating the pattern of achievement by content strand to the degree of emphasis in content by the systemic initiative. Variation in the change of student performance by content strand over time is helpful in explaining group differences. For example, in SSI states, the achievement of Black students and White students increased from grade 4 to grade 8 on Number Sense, Properties, and Operations at nearly the same rate. However, grade 4 to grade 8 increases for White students far exceeded those for Black students in Measurement, Geometry and Spatial Sense, and Data Analysis, Statistics, and Probability. This suggests that Black students received sufficient instruction in Number Sense, Properties, and Operations, but received less than adequate instruction on the other topics. Most state assessments require all students to take the same assessment instrument and are limited in the number of content area results that can be reported. An important benefit in using State NAEP data is that the design allows data from NAEP assessments to be reported by content strands. The NAEP assessments use a matrix-sampling procedure, a balanced incomplete block spiraling design (Allen, Jenkins, Kulick, & Zelenak, 1997, p. 7), where over 140 items at grade 4 and over 160 items at grade 8 were distributed among 13 blocks. Any one student took only three blocks. However, the total number of items by the five content strands ranged in 1996 from 17 to 59 at grade 4 and from 25 to 47 at grade 8. These items included multiple-choice items, constructed-response items (scored dichotomously and polytomously), and cluster items. Based on the sampling design, by student and item, inferences could be made about the performance of students in the state on the five mathematics content strands. The added information that can be gained by considering the differences in student performance by

mathematics topics is important both for observing changes in patterns of performance over time within a state and for comparing the differences in performance between states.

Item Statistics and Parameters

Access to individual item statistics and parameters also can be of value in deciphering the differences among states and programs. Although some information can be gained by considering variations in student responses to one item by those in different groups, we found that it was informative to consider the set of items that differentiated one group from another. Using differential item functioning (DIF), typically used to study item bias, we attempted with some success to detect differences in the construct being measured in the SSI and non-SSI groups. Just the variation in the number of DIF items over time suggested convergence between the groups in the underlying construct being measured. For example, the SSI and non-SSI student performance became more similar in 1996 than in 1992 and 1990, as indicated by the lower number of DIF items in 1996. We identified those DIF items that favored the SSI group and those that favored the non-SSI group. By analyzing the content measured by DIF items that favored one group compared to the other group, we were able to describe more explicitly what mathematics SSI students were better able to do compared to non-SSI students. There was an indication that at least some of the items differentiating the SSI students from non-SSI students addressed topics emphasized more by reform mathematics (National Council of Teachers of Mathematics, 2000). Our analysis indicated the feasibility of using DIF analysis for investigating the differences among SSI-related state reforms and other evaluations of reform efforts.

Lessons Learned

NAEP achievement and questionnaire data only served to characterize the SSI and non-SSI states very generally over time. In order to probe more deeply into what constituted systemic reform and to capture more cogently what drove the states' decision-making process in their efforts to achieve reform in mathematics and science, we engaged in intensive interviews of the leaders in most of the SSIs and extensive document review for all of the SSIs. The challenge of designing and carrying out systemic reform and of sustaining the effect over time demands the development and enactment of a plan that includes 1) effective activities for making changes on the required scale and of the needed scope; 2) a feasible means for putting those changes into place in a timely and coordinated manner; and, 3) attention to the interests and influence of a wide range of stakeholders within the state and within the initiative. The processes by which the leaders of systemic reform address these demands are described as the *strategic thinking* of the initiative. The choices SSI leaders made about which elements of the education system the initiative would target and how those elements would be addressed constitute the initiative's *technical strategy*. Common elements of the system the technical strategy might address include teacher capacity, infrastructure for delivering assistance to schools and teachers, and policies such as curriculum standards, instructional materials adoption, and student assessment. Each initiative had to make key decisions about how the SSI would position itself to reform long-standing mathematics and science education systems,

which partners to include in the initiative and how each partner's interests could be made to fit within the SSI, and how to address interests that did not fit well with those of the SSI. These decisions constitute the *political strategy* of the initiative. By analyzing the strategic thinking employed by the SSIs along with performance measures, we were better able to produce a comprehensive view of the SSIs. Interestingly, the student achievement of some SSIs that focused more on strategic thinking did not increase greatly compared to SSIs that were less strategic in their planning. By incorporating data from the two perspectives—student achievement and planning—we were better able to put the findings from each of the analyses into a more comprehensive context, which would not be possible if only one of the approaches were applied in isolation.

Attribution

The data used in this study are of insufficient quantity to be used as a basis for discerning causal relationships between the NSF Statewide Systemic Initiatives Program and student achievement. However, the data do provide a sufficient base for developing plausible arguments for the relationship between a state's participation in the SSI program and student achievement. We were able to distinguish clearly between those states that participated in the SSI program and those that did not and compare the two groups on the change in student achievement over time. We used all of the states that participated in the State NAEP for the three administrations—1992, 1996, and 2000—for the longitudinal study. This included 14 SSI states (64% of the 22 SSI jurisdictions that were funded for at least the first five years) and 13 non-SSI states (54% of the 25 states not participating in any way in the SSI program). Both the SSI states and the non-SSI states included in the longitudinal study represented all of the states in their respective groups on a number of variables, including total educational expenditures per capita, public school enrollment, percentage of White students enrolled in public schools, and average grade 8 mathematics achievement as measured on the State NAEP in 1992 (see Table 1A.1 in the Appendix to this chapter). In our analysis of group comparison, the means of the subgroups fell well within the 95% confidence intervals for the means of both the SSI and non-SSI total groups and none were statistically different from the means for the total group. However, for both the SSI and non-SSI groups, students in those states in the longitudinal groups performed at a lower level than those in states not in the longitudinal analysis. The states in our longitudinal study were also distributed nearly equally among the four regions of the country, as defined by NAEP (see Table 1A.2 in the Appendix). Thus, although the analyses were only performed on subgroups of states, the 14 SSI states and the 13 non-SSI states in the subgroups are characteristic of the total population—increasing the confidence that the results for the subgroups are representative of the total groups.

SSI states do not, however, represent the fifty states in the most general sense. It is clear that NSF awarded SSI funds to those states with the highest public school enrollments. The four states with the highest student enrollment all received SSI funding (California, Florida, New York, and Texas). Of the ten states with the highest enrollment, in fact, eight received SSI funding. However, the differences in enrollment between SSI states and non-SSI states were not statistically significant. Also, SSI states when

compared to non-SSI states had, on the average, a smaller proportion of White students enrolled in public schools. It is not surprising that NSF would seek to award funds through the Systemic Initiatives Program to those states with the greatest need. These data indicate that the SSI states and the non-SSI states differed as discrete groups at the beginning of the program in 1992. In looking for SSI impact on student achievement, an important indicator is the change in student scores over time. Any increase in student achievement by SSI states compared to non-SSI states is worthy of note because the SSI states in general had lower values on major indicators in 1992, such as State NAEP achieved scores.

State NAEP data for both grade 4 and grade 8 existed over a span of eight years and three test administrations. This provided an opportunity to report on the change in student achievement over two four-year periods: The first four years, 1992 to 1996, occurred at the peak of the SSI program when all of the SSI states were receiving funds from NSF and were engaged in the major activities associated with this funding, such as providing professional development for teachers, strategic planning, developing infrastructure, and instituting new mathematics curricula. It took SSIs some time to develop an organization and to get their programs underway, but by the time of NAEP's testing early in 1996, states would have had at least three years of funding to initiate changes. If student achievement were associated with SSI activities, then we could expect to find the greatest increase in student achievement between 1992 and 1996. In the second time period, from 1996 to 2000, only eight of the SSIs (five that consistently participated in NAEP and are included in the longitudinal sample) received continued NSF funding in Phase II of the SSI program. The other 14 SSI states either had to find other funding for continuing their SSI-related initiatives, or discontinue them altogether. NSF clearly placed a condition on the original SSI funding with its expectation that states would sustain the work of the SSI after the funding period ended. Thus, it was not unreasonable to expect that students would continue to show achievement gains after NSF funding ceased.

As a result of differentiation in funding over the 1996 to 2000 time period, a natural experiment was possible—we were able to compare SSI states that had NSF Phase II funding ($N = 5$), SSI states that completed five years of NSF funding and then were left to their own resources ($N = 9$), and the states that had not received any NSF SSI funding ($N = 13$). A study of the changes in student performance in these three contrasting groups allowed an opportunity to develop some plausible arguments for the relevance of NSF funding. Were NSF SSI funding not a factor, we would expect to see little or no difference in the change in student performance among the three groups of states over the two four-year periods, 1992 to 1996 and 1996 to 2000. If NSF's SSI funding was a critical factor, then we would expect that both SSI groups would show greater changes in student performance than the non-SSI group over the first time period and only the Phase II states over the second time period. If the first five years of funding were sufficient to produce sustained systemic reform, it is not unreasonable then to expect that changes in student achievement for both SSI groups would have continued over the second time period, from 1996 to 2000.

Clearly, there were too many variables could not be controlled for in this study for us to be able to conclude with certainty that NSF's SSI funding was the major contributing factor to improved student performance. Over the eight-year time period, states developed, changed, and implemented assessment systems, accountability systems, new curricula, and teacher-certification requirements. Many of these broad systemic educational changes went beyond mathematics and science, the two content areas targeted by NSF funding. The systemic initiative in a state could have contributed to these changes, but it is unreasonable to expect that the relatively modest funding NSF provided relative to states' total education spending would be the sole cause, or even a significant cause, of whatever changes in student achievement these education reforms may have produced. The program was not designed to effect changes on its own—but to serve as a catalyst for reform. It also is not clear whether the SSI activities or initiatives impacted a sufficient number of teachers and students within a state to represent a saturation point that would be sufficient to produce a change in student learning as measured on the basis of a sample of students. Through June, 1995, based on data on all of the mathematics and science teachers in 25 jurisdictions with SSI funding, Zucker and his colleagues found that 15% of the elementary teachers, 37% of the middle grade teachers, and 20% of the high school teachers had been served in some way by SSI professional development activities (Zucker et al., 1998, p. 23). However, if an SSI devoted its efforts directly to developing and implementing state standards or frameworks for mathematics and science—a less costly strategy for reaching a large number of teachers and students—then it is possible that a larger proportion of state students would be affected.

NAEP sampling procedures could have inhibited its sensitivity to detecting SSI impact. The actual State NAEP sample of public school students for each state ranged from 2,000 to 2,700. The SSI implementation strategy of some states, such as that of the SSI in New York, was to begin small by focusing on a limited number of schools with the expectation of expanding later (Zucker et al., 1998). This could result in the case that none of the students directly influenced by the SSI were included in the State NAEP assessment. This would be less likely the case if an SSI addressed a wider spectrum of schools, teachers, and students. Along with the sampling issue, the data collection activities for the State NAEP were the responsibility of each participating jurisdiction. Although some quality control of the State NAEP was provided by a national contractor monitoring 50% of the administrations in jurisdictions participating for the first time and 25% of the administration in the other jurisdictions, it is unknown whether the State NAEP results can withstand the same scrutiny as the other NAEP administrations because of the dependence on local people to implement the sampling plan and test administration.

As of 1997, about 40% of the SSIs had allocated funds to develop state mathematics and science frameworks (Zucker, et al., 1998). Six of the 14 SSI states (43%) in the longitudinal group had provided such funding. The information that exists on the saturation of SSI funding suggests that the SSIs had reached at least some mathematics and science teachers either directly through professional development activities or indirectly through advancing modified curriculum frameworks. Both of these

activities could conceivably have led to a proportion of the students in these states receiving more effective instruction and instruction on new content that could result in improved group performance on assessments. If SSIs served as a catalyst or leverage for reform activities, as intended, then the impact of the SSI could have a wider effect on student performance through indirectly motivating incentives for change rather than indirectly influencing classroom practices and student learning through specific SSI activities, such as professional development workshops. Without more detailed information than existed or than we could produce on what each SSI did, we are unable to parcel out any effects that may have occurred due to sampling issues.

What we have done to add to the credibility of our plausible arguments is to identify intermediary variables that can be used to link possible SSI activities to change in student performance: We have collected information from SSI sites on the emphasis placed on the five mathematics topics tested and reported by NAEP, analyzed classroom process variables obtained from the NAEP questionnaires, gathered data on the SSI, and on other reforms from other sources, and disaggregated NAEP data to look for patterns that may reveal meaningful differences that can be associated with different types of reform activities. Through this collection of data, we believe a case can be developed to support the claims that 1) states with SSI funding have engaged in activities advanced by current reform efforts and 2) that student achievement has improved as a result.

Summary

There is no question about the difficulties faced in studying the impact of NSF's SSI program. We have proceeded with this clearly in mind. We have tried to make clear what choices we made in conducting the analyses and how much credence should be given to the reported findings. An important part of the study is that it builds upon both quantitative and qualitative data throughout a lengthy period of time. The NAEP data span a ten-year period. SSI leaders and documents were consulted nearly ten years after NSF made its first state award. We have uncovered factors and produced findings that coincide with some SSI effect on student learning and that reveal details of the strategic thinking leaders engaged in while implementing statewide reform. Along with these findings, our approach to addressing our basic research question illustrates important characteristics of conducting an inquiry of large-scale reform that has strong methodological implications for investigating other educational programs, such as the Mathematics and Science Partnerships.

Appendix

Table 1A.1
SSI States and Non-SSI States Included and Not Included in the Longitudinal Study
On Selected Indicators

Table 1A.2
Number of SSI States and Non-SSI States Included and Not Included in the Longitudinal Study by Region of the Country

Table 1A.1

SSI States and Non-SSI States Included and Not Included in the Longitudinal Study on Selected Indicators[1]

	N[2]	Total Expenditures per Capita 1990-91 (in dollars)		Public School Enrollment Fall 1993		Percent White Enrolled in Public Schools Fall 1993		Average Math Achievement NAEP Grade 8 1992[3]	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
SSI States	22	3569	655	1,043,563	1,305,908	68	23	265	7.76
Longitudinal Sample	14	3559	754	1,302,079	1,533,876	67	18	264	8.41
Other	8	3588	444	591,160	615,775	70	30	268	4.55
Non-SSI States	25	3643	1409	623,454	458,246	77	17	268	10.71
Longitudinal Sample	13	3425	780	563,733	315,870	73	21	265	11.30
Other	12	3879	1884	688,151	583,552	81	11	275	5.62

[1] Data were obtained from Snyder, T. D.. & Hoffman, C. M. (1995). *Digest of education statistics 1995*. National Center for Education Statistics (NCES 95-029). Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement.

[2] The 22 SSI States included Puerto Rico. The four states from which NSF withdrew funding are not included—Florida, North Caroline, Rhode Island, and Virginia.

[3] Of the 22 SSI states, 18 participated in the 1992 State NAEP (14 in the longitudinal group). Of the 25 non-SSI states, 19 participated in the 1992 State NAEP (13 in the longitudinal group).

Table 1A.2

Number of SSI States and Non-SSI States Included and Not Included in the Longitudinal Study by Region of the Country

	N	Northeast	Southeast	Central	West
SSI States	21	7 (33%)	5 (24%)	4 (19%)	5 (24%)
Longitudinal Sample	14	4 (29%)	5 (36%)	2 (14%)	3 (21%)
Other	7	3 (43%)	0 (0%)	2 (28%)	2 (28%)
Non-SSI States	25	2 (8%)	5 (20%)	8 (32%)	10 (40%)
Longitudinal Sample	13	1 (8%)	4 (31%)	4 (31%)	4 (31%)
Other	12	1 (8%)	1 (8%)	4 (33%)	6 (50%)

CHAPTER 2

PROJECT GOALS

This study sought to answer two research questions. The first is central to understanding the importance of the systemic initiatives as a strategy for attaining large-scale education reform in states and how statewide systemic initiatives have improved mathematics and science achievement and participation of students. The answer to the second question enables building on the large amount of knowledge people have gained about systemic reform in mathematics and science. Together the two research questions address both what statewide systemic initiatives have accomplished and what has been learned from this reform effort.

The original research questions were:

1. What impact has NSF's Statewide Systemic Initiatives (SSI) Program had on student learning, on student participation, and on other important variables such as classroom practices and differential performance by ethnic group?

    A. What differences between SSI states and non-SSI states were evident in mathematics achievement and student participation variables (e.g., course completion) as measured by NAEP over the period 1992–2000? What explanations exist for observable differences or for the absence of observable differences?

    B. Were there improvements in statewide achievement and in student participation variables for mathematics and science on multiple measures, including NAEP and state assessments in a selected cluster of SSI and non-SSI states? What explanations are there for improvements or for no observable improvements in relation to SSI, state reform initiatives, and other activities within the states?

    C. How does improvement in mathematics and science outcomes on multiple measures (e.g., state assessments and district assessments) relate to the degree of implementation of systemic reform, level of SSI participation, and other variables?

2. What lessons have been learned about designing, implementing, evaluating, and supporting statewide systemic reform?

    A. What knowledge can be gleaned from three states about how systemic reform by a state can be approached and be successful?

    B. How can mistakes be avoided?

C. Which kinds of data from one state can be generalized to other states and which are state-specific?

Certain conditions were imposed on this study by of the availability of data. The impact study only addressed mathematics specifically and not science because State NAEP data were only available in mathematics. However, the study of lessons learned attended to both mathematics and science. State NAEP data were collected only on grade 4 and grade 8 students. We therefore limited our focus to analysis of student achievement at these grades. NAEP data were not available on what courses students pursued in high school. This prevented us from studying variations in student participation that could be related to SSI activities. The group of states included in the impact study was limited to those states that had participated in the State NAEP in 1992, 1996, and 2000. This group included 14 SSI states and 13 non-SSI states. The three-year project was extended for a year in order to incorporate State NAEP data from 2000 in the analysis. Thus, most of the State NAEP analyses were based on three data points that represented an eight-year time period—1992, 1996, and 2000. The State NAEP produced data for both grade 4 and grade 8 mathematics for each of these years. Only grade 8 mathematics data were collected in the trial State NAEP in 1990. We reported data from 1990 in our first technical report (Webb, Kane, Kaufman, & Yang, 2001). However, because fewer states participated in the 1990 trial State NAEP, we decided not to include the 1990 data in our final analysis in order to increase the number of states with a complete set of data.

Study Design

In mining the State NAEP data to address the first research question, we have used multiple methods in our effort to determine what the major impacts of the SSI program were on student performance in participating states. In our analyses, we contrast a subgroup of the SSI states with a subgroup of the non-SSI states, using state as the unit of analysis. Each state is weighted equally, rather than by population. States in each subgroup were determined by their participation in the State NAEP, first for 1990, 1992, and 1996 (17 SSI states and 11 non-SSI states) and then for 1992, 1996, and 2000 (14 SSI states and 13 non-SSI states).[1] State NAEP data for 1992 in both grades 4 and 8 were used as baseline data prior to the implementation of the SSI activities in states. We analyzed change in achievement scores from 1992 to 1996 and from 1996 to 2000. Confidence intervals and standard errors of measurement are reported when appropriate to determine the statistical significance in results. Results for our analysis, which contrasted data from SSI states and non-SSI states with State NAEP data for 1990, 1992, and 1996 are reported in a previously published technical report (Webb, Kane, Kaufman, & Yang, 2001) that is available on the Web.[2] In this final project report, findings are reported for our analysis of State NAEP data for 1992, 1996, and 2000.

[1] The 17 SSI states and 11 non-SSI states include those states that participated in the State NAEP for 1990, 1992, and 1996. The 17 SSI states include all states that participated in the State NAEP that received any funding from NSF, including three states whose funding was terminated before the end of the five-year grant period. The 14 SSI states and the 13 non-SSI states include those states that participated in the State NAEP for 1992, 1996, and 2000; the 14 SSI states only include states that received NSF funding for at least the first five years.

[2] http://facstaff.wcer.wisc.edu/normw/technical_reports.htm

The sample of states included in the State NAEP analysis reported here includes 14 SSI states and 13 non-SSI states:

SSI States	**Non-SSI States**
Arkansas	Alabama
California	Arizona
Connecticut	Hawaii
Georgia	Indiana
Kentucky	Maryland
Louisiana	Minnesota
Maine	Mississippi
Massachusetts	Missouri
Michigan	North Dakota
Nebraska	Tennessee
New Mexico	Utah
New York	West Virginia
South Carolina	Wyoming
Texas	

As reported in Chapter 1, the SSI states and the non-SSI states as a group are comparable to the total groups for each category on a number of variables.

A series of analyses were conducted to describe the grouping of states included in the study, to report the differences between SSI and non-SSI states over time, and to identify variables related to student performance. Descriptive statistics were used to report on how the state groups varied on demographic variables, trend scores over time, and cohort growth in achievement. These analyses were performed on the total group, by gender, and by race. One technique we used to better detect possible relationships between student achievement as measured on the State NAEP and content emphasis by the SSIs was to analyze the change in achievement on the total score and the five content strands for the cohort of students who were in grade 4 in 1992 and in grade 8 in 1996. We did a similar cohort analysis on the growth in achievement by grade 8 students in 2000 who were in grade 4 in 1996.

Using the responses to teacher questionnaires administered with the State NAEP in 1992 and 1996, we identified six indicators of mathematics reform. When the measures were the same across time, a repeated measures analysis of variance was used, with SSI as a between-subjects factor and time as a within-subjects factor. We used a two-step regression model when the measures were similar, but not exactly the same. On the average, the SSI states had lower mathematics achievement at grade 4 and grade 8 on the State NAEP in 1992 and prior to the SSI program. We employed Bayes and fully Bayesian methods (Raudenbush, Fotiu, & Cheong, 1999) to compare the overall trends of SSI states and non-SSI states for 1990, 1992, and 1996 and to detect differences between the two groups.

Because of significant variation in achievement and growth in achievement over time, we developed state profiles for each of the SSI states using State NAEP data. These profiles are located at http://www.wcer.wisc.edu/SSI/Profiles/state%20profile.htm. Other researchers (Zucker, Shields, Adelman, Corcoran, & Goertz, 1998; Clune, 1998) have classified each SSI on different categories based on the activities of the SSI (e.g., professional development, curriculum standards, infrastructure building). None of these classifications addressed the emphasis an SSI gave to the five mathematics content strands measured on the State NAEP. Therefore, we thought it imperative that we gather this data ourselves. One researcher identified and contacted individuals in the SSI states who would be most knowledgeable of the mathematical content emphasized by the SSI over the period of time from 1990 to 1996. Information was elicited from at least two sources in each state on the background of mathematics reform, target population, saturation (percent of students reached by reform), the nature of the reform, and the degree of emphasis given to each mathematics strand and abilities. When those interviewed could not provide information on the emphasis given to content strands, the proportion of items on the state assessment allocated to each content strand was used. This information was combined, with state demographic information and the key activities of the SSI in the state, to create the state profile.

On the basis of the State NAEP data, it was evident that some states did significantly better than others in raising student achievement scores from one test administration to the next. For example, from 1992 to 1996, Texas and Michigan were two SSI states that exhibited higher grade 4 mathematics scores than other SSI states. We assigned the 14 SSI states included in the analysis to three categories based on mean mathematics achievement gains from 1992 to 1996 and from 1996 to 2000. Six SSI states showed steady growth over both four-year periods; four SSI states had some growth over one of the periods but little growth over the other time period; and four SSI states showed little growth over either of the periods. We then compared the three categories of states on a number of other variables, including state-level policies, tests tied to the standards, state assessments aligned with the SSI goals, accountability policies, and statewide change in classroom practices. Because the number of states included in the analysis—all 14 SSI states with State NAEP data—and the variation in the form of data on other variables, we turned to qualitative analysis procedures (Miles & Huberman, 1984) in our effort to understand the approaches to implementing the SSI, and other factors that were associated with steady improvement in mathematics achievement. We employed three analysis activities—data reduction, data display, and conclusion drawing /verification. In the latter, we considered consistent findings for states within an achievement change category.

When State NAEP 2000 data became available from NCES in 2002, we were able to analyze the achievement of the 14 SSI states and 13 non-SSI states in relationship to state participation in the SSI program. Seven of the 14 SSI states had only had one round of funding, which spanned the 1992 to 1996 period. Five of the SSI states received a second round of funding that extended over both four-year time periods bracketed by State NAEP administrations, 1992 to 1996 and 1996 to 2000. The 13 non-SSI states, of course, had no SSI funding for the eight years. With this configuration of data, we were

able to analyze the differences in the change in mathematics achievement related to SSI funding over the two time periods, along with other state education accountability characteristics. Even though the number of states in each classification was small, we were able, because of certain conditions, to make some reasonable inferences from the data by using regression analysis to estimate the effect size and error of measurement. We were able to precisely define the variables that obtained and identify significant differences among the groups.

We used differential item functioning (DIF) analysis as another technique to detect possible differences between students' mathematics achievement in SSI states compared to non-SSI states. By using DIF analysis, we were able to identify those assessment items on which one group or the other performed significantly differently when compared with the performance on all of the items. Then, by analyzing the content of the DIF items that favored students in the SSI states and the content of the DIF items that favored students in the non-SSI states, we were able to distinguish differences in the underlying constructs of the performance of SSI students compared to non-SSI students. Even though two groups of students may have nearly the same total score on an assessment, the groups may vary systematically in the items they each answered correctly. This variation could represent certain significant differences in what mathematics students in each group know. During this study, we were particularly interested when SSI group items measured content emphasized in reform documents (e.g., data analysis) could be considered more challenging content than routine problem solving. Given NSF's strong SSI emphasis on student achievement vis a vis challenging content in mathematics and science and on content alignment with state standards, if SSI students were found to do better on comparable items administered by State NAEP, then at least the results would prove compatible with the goals of the program.

Some of the SSI states identified schools, teachers, and students engaged specifically in SSI activities or that could have been influenced by SSI activities. A few of these states then related SSI participation to scores on the assessments administered statewide, frequently at grades 4 and 8, similar to the State NAEP. States with such data provided us an opportunity to study how the State NAEP results related to student performance on state assessments and whether it would be possible, using state assessment data, to draw some inferences between the State NAEP and SSI participation. This was done to respond to research question 1C above regarding how improvement on mathematics and science outcomes relates to reform measures. We conducted a focus study using state assessment data from three states (Maine, Massachusetts, and Texas) over nearly the same period of time as that measured by the State NAEP. This study has particular relevance to the *No Child Left Behind* legislation that will use NAEP results as one means to validate performance as measured by state assessments (U.S. Department of Education, 2001).

To determine what lessons could be learned from the SSIs on designing, implementing, evaluating, and supporting statewide systemic reform (research question 2), the Horizon research team first developed an analytic model depicting the steps in

strategic planning and implementation.[3] This model was then used to structure questions and protocols for reviewing SSI proposals, interim and final reports, and evaluation reports for 21 SSIs. More intensive data were gathered from six SSIs by interviewing key leaders responsible for implementing the activities in mathematics and science. Researchers conducted site visits at four additional SSIs. The data obtained were used to prepare four in-depth case studies, six analytic reports, and a cross-case analysis. Data and conclusions from these analyses were then used to revise the original analytic model. The analytic model itself, presented as a research study, is one of the important products of this work.

Findings from Previous Work

The work described in this final report builds upon research that was detailed in an earlier technical report (Webb et al., 2001). Those findings were based on the analysis of State NAEP data for 1990 (grade 8 only), 1992, and 1996 from 17 SSI states (including three states that did not receive funding for the full five years) and 11 non-SSI states. In that analysis, we found that the SSI states had higher percentages of Black and Hispanic students than non-SSI states. Otherwise, the SSI states differed very little from the non-SSI states on demographic variables. Student performance on the State NAEP improved for both SSI and non-SSI states from 1990 to 1996 at grade 8 and from 1992 to 1996 at grade 4. Prior to the beginning of the SSI program, students in the SSI states included in the analysis had performed significantly lower than students in the non-SSI states at both grade levels. Over the course of the SSI program, up to 1996, SSI states improved at a slightly faster rate than did the non-SSI states for both grades 4 and 8. However, non-SSI states were more successful in reducing the achievement gap between male and female students. In 1992, male students out-performed female students in mathematics in both SSI states and non-SSI states. By 1996, this gap was eliminated in the non-SSI states, but a difference of over two points continued in the SSI states at grade 8. Although differences in achievement between White students and Black students remained in both SSI and non-SSI states (about 30- to 34-point differences), there was some narrowing of the gap for specific content strands. For the SSI states at grade 8, the gap increased for Measurement, but narrowed slightly on both Geometry and Algebra and Functions. For the non-SSI states at grade 8, the gap increased on the composite scores and on all five content strands. The growth by Black students in SSI states from grade 4 to grade 8 was greater on the Algebra and Functions scale than that of Black students in non-SSI states. In SSI states, Black students gained more than White students on the Algebra and Functions scale over the four-year period, from grade 4 to grade 8.

Using State NAEP teacher questionnaires, we developed six indicators of mathematics reform:

Relative Emphasis on Reasoning and Communication
Students' Opportunities for Mathematical Discourse
Teachers' Knowledge of the NCTM *Standards*

[3] Iris Weiss, co-principal investigator, led a research team from Horizon Research, Inc., to conduct the part of the impact study that analyzed what lessons could be learned.

Last Year's Professional Development
Reform-Related Topics Studies
Calculator Use

All of these indicators related in some way to practices associated with reforms in mathematics in the 1990s. In general, both SSI and non-SSI states increased on the six indicators. At grade 8, from the responses of those teachers who taught the students tested by the State NAEP, SSI states as a group scored significantly higher than did non-SSI states on five of the six indicators. At grade 4, SSI states scored significantly higher than non-SSI states on four of the indicators. These results signify that by 1996 teachers of those students tested in SSI states were more apt to be using practices advanced in the reforms than teachers in the non-SSI states. A review of data from 1992 to 1996 shows that the SSI states increased more than non-SSI states on the Mathematical Discourse and Reasoning and Communication indicators. However, there was considerable within-group variation for both the SSI states and the non-SSI states.

Summary

Multiple research methods were employed to address the two major questions. The State NAEP tests for 1990, 1992, 1996, and 2000 were the main source of data for studying the impact of the SSI program on student achievement and classroom practices. Our analyses, using State NAEP data for 1990, 1992, and 1996, indicated that SSI states, although beginning with an average score below that of the non-SSI states, increased in performance at a faster rate than non-SSI states and more readily incorporated classroom practices that were compatible with reform initiatives into their curricula. Similar analyses were performed with State NAEP data for 2000 to see whether these trends continued. The actual group of states included in the analyses varied by time period because not all states participated in the State NAEP each time it was given. Our analyses for the three testing times—1992, 1996, and 2000—were based on data from 14 SSI states, all of those that received full NSF funding for at least five years, and from 13 non-SSI states. In addition to analyzing trends over time, we employed DIF analysis to study variation in the underlying mathematical constructs achieved by the two groups. We also conducted a focus study relating the State NAEP results to the results from the state assessments administered by each of three states. Our analyses of SSI strategic planning processes utilized qualitative methods to glean lessons learned. Because our study spanned two NSF funding periods, at least some data were analyzed from a total of 21 SSI states. Site visits were made to six of these SSI states and telephone interviews were conducted with SSI leaders in four other states. This work has produced a total of six case studies, four analytic studies, and the refinement of an analytic model that details how a strategic planning process used by complex education systems can advance reform initiatives.

CHAPTER 3

REVIEW OF RECENT STUDIES[1]

In recent years, several studies (Grissmer, Flanagan, Kawata, & Williamson, 2000; Raudenbush, Bryk, Cheong, & Congdon, 2000; Raudenbush, Fotiu, & Cheong, 1999; and, Wenglinski, 2000, 2002) have examined the impact of a variety of factors on mathematics achievement as measured by the National Assessment of Educational Progress (NAEP). The correlates studied by each of these research groups can be categorized within four classifications: student, family, and home characteristics; educational resources and teacher characteristics; schooling characteristics; and, classroom practices. While the first of these is beyond the control of educational institutions, the latter three are variables that are under the influence of state, district and school policies, as well as of teacher practices.

Grissmer et al. (2000) and Raudenbush et al. (1999) found the typical associations between parental educational levels, family income, and race/ethnicity on the one hand and NAEP mathematics achievement on the other. From the perspective of reform efforts developed to improve achievement, effective policies and practices are of greatest interest. Raudenbush, Fotiu, and Cheong (1998) found that social background and ethnicity are associated, in part, with the quality of schooling and that:

> School means that are not adjusted for student composition will typically convey an overly negative picture of school process in those schools with the most disadvantaged students. (p. 255)

Furthermore, this study identified an inequality of resources based on social and ethnic factors that is " . . . much more pronounced in some states than others" (p. 265). Because of the potential bias associated with the study of unadjusted means, together with the differences across states in the allocation of resources and social and ethnic factors, Raudenbush et al. (1999) suggested that

> . . . state comparisons might productively focus on state differences in policy-relevant correlates of proficiency rather than on state differences in mean proficiency. (p. 413)

Additional policy-related factors were identified by Grissmer et al. (2000). Controlling for student, family, and home characteristics, these researchers found that per-pupil expenditures, pupil-teacher ratios at early grade levels, higher levels of teacher resources, percentage of children participating in public prekindergarten, and the availability of high school algebra for grade 8 students were all positively related to

[1] This literature review extends the review completed for the previous technical report (Webb, Kane, Kaufman, & Yang, 2001, pp. 11-21).

student achievement as indicated by an aggregate of NAEP mathematics and reading scores. Wenglinski (2000, 2002), also controlling for student, home, and family characteristics, found that the policy-relevant variables- teachers' majors/minors in mathematics, professional development that included a focus on working with disadvantaged students, applying problem solving techniques to unique problems, and teacher higher-order thinking skills were positively related to mathematics achievement on NAEP. Contrary to traditional findings (Coleman et al., 1966), Wenglinski's most recent study found that teacher variables had a greater effect on student achievement than did SES (Wenglinski, 2002).

Another school-related characteristic that influences achievement is student behavior. Barton (2001) reported on the increase in the incidence of discipline issues in the 1990s and notes that the statistics rose more dramatically for Black and Hispanic students.

These recent studies suggest that understanding the necessary focus and effectiveness of reform efforts requires attention to policy and practice outcomes as well to as student achievement. Furthermore, when comparing states both on achievement and on policy and practice variables, it is important to control for social and ethnic factors.

Mathematics Content

The National Assessment of Educational Progress (NAEP) has been conducted, in some form, since 1969.[2] Over time, three distinct NAEP projects have evolved: the Main NAEP, the Long-Term Trend NAEP, and the State NAEP. The Main NAEP periodically assesses students' achievement in reading, mathematics, science, writing, U.S. history, civics, geography, the arts, and other subjects at grades 4, 8, and 10. The State NAEP has measured writing, reading, mathematics, and science at grades 4 and 8. Student samples for this program are drawn to permit inferences about the achievement levels for each participating state. The content of both the Main and State NAEP programs follows curriculum frameworks developed by the National Assessment Governing Board (College Board, 1996), which have been adapted to changes in the nation's curricula. Since 1989, the mathematics tests have followed the recommendations of the National Council of Teachers of Mathematics' *Curriculum and Evaluation Standards for School Mathematics* (National Council of Teachers of Mathematics, 1989), often referred to as the NCTM *Standards*. Test-item types for the Main and State NAEP assessments that are consistent with the current state-of-the-art in achievement testing also have evolved.

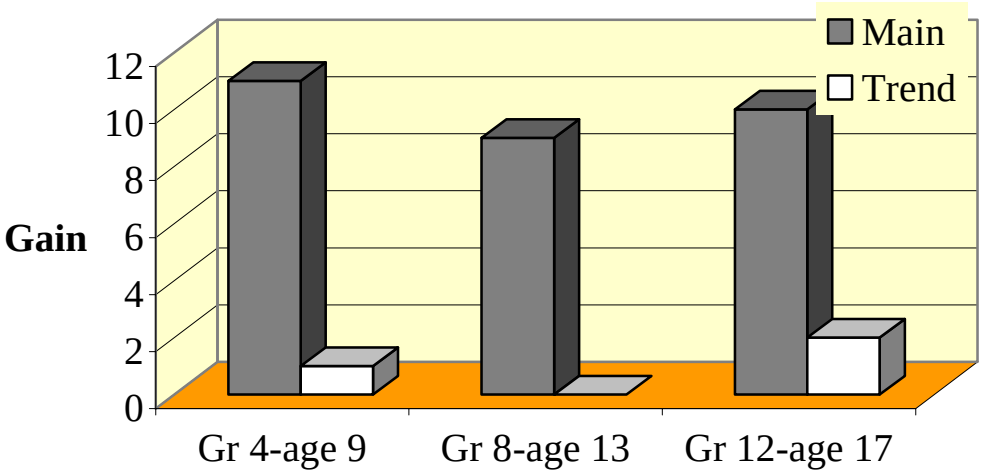
In contrast, both the student sampling framework and the content of the Long-Term Trend NAEP have remained essentially unchanged. The Trend program, which began 30 years ago and was intended to monitor general trends in achievement, was NAEP's original program. Rather than focusing on grade levels, the Trend assessment has targeted students at ages 9, 13, and 17 in mathematics, reading, and science. Unlike

[2] http://nces.ed.gov.nationsreportcard/site/whatis03/

the Main and State NAEPs, the content of which evolves to match changes in curriculum and instructional practice, the content blueprints of the Trend tests have not changed.

Since 1990, the frameworks for Main and State NAEP assessments in mathematics have covered five content areas and three mathematical abilities. The content areas are: Number Sense, Property, and Operations; Measurement; Geometry and Spatial Sense; Data Analysis, Statistics, and Probability; and, Algebra and Functions. The mathematical abilities measured are conceptual understanding, procedural knowledge, and problem solving. Three types of items were employed: multiple-choice, open-ended, and extended open-ended, first used in 1992 (College Board, 1996).

Figure 3.1. Comparison of Main and Trend NAEP scale score gains between 1990 and 1996.



Loveless and Diperna (2000) compared trends on the Main and Trend NAEP programs and observed that improvements between 1990 and 1996 shown on the Main NAEP mathematics tests were not replicated on the Trend version. Figure 3.1 compares the gains for these two NAEP programs between 1990 and 1996. The researchers explained the discrepancies in trends on the basis of the differences in the content of the two NAEP programs. Their " . . . analysis suggests that the Main test is more oriented toward NCTM-like topics (geometry and problem solving) and the Trend more toward pre-NCTM topics (arithmetic)" (p. 18). Table 3.1 shows the definition of " arithmetic" that Loveless (2002) used in his analysis. In their analysis of the percent correct on the items, the researchers were able to show that student performance in the arithmetic clusters had remained steady for age levels 9 and 13 and declined somewhat at age 17 (Loveless & Diperna, 2000). Furthermore, they observed that students at all three age levels showed gains on the geometry items. The authors used these data, together with their observation that the Main NAEP contained a considerable larger proportion of geometry items than did the Trend version as a basis for concluding that the difference between the scale-score trends on the two NAEP programs was attributable to the difference in emphasis on arithmetic between the two programs.

Table 3.1
Arithmetic Clusters for Trend NAEP

Test Level	Number of Items	*Item Content*
Age 9	21	Addition, subtraction, multiplication, and division of whole numbers.
Age 13	23	Addition and subtraction of whole numbers, fractions and decimals
Age 17	17	Addition of whole numbers, fractions, decimals, converting decimals to numbers

Based on his understanding of the facts, Loveless (2001) argued that the item composition of the 2004 Main NAEP tests should reflect an increased emphasis on arithmetic. Representatives of the National Assessment Governing Board have questioned this interpretation of the data, as well as their conclusion about needed changes in the NAEP framework.[3] The draft of the content framework for the 2004 Main NAEP that is currently being circulated for public review is designed to: 1) reflect recent curricular emphases and objectives; 2) include what various policy makers, scholars, practitioners, and interested citizens believe should be in the assessment; 3) maintain the short-term trend lines begun with the 1990 mathematics assessment to permit reporting of changes in student achievement over time; and, 4) include clearer and more specific objectives for each grade level (Loveless, 2001, pp. 3-4). The framework recommends that, for grade 4, the proportion of items covering Number Sense, Properties, and Operations remain the same as for 1996 and 2000. Because at grades 8 and 12, mathematics increasingly requires that number computation and operations be done in the context of other content such as data analysis and probability, geometry/measurement, or algebra, the draft content framework recommends that there be a reduction in the proportion of items in the number category (Council of Chief State School Officers, National Assessment Governing Board, U.S. Department of Education, 2001). Since the framework specifies content for the tests at the grade level at which it is expected to be learned and because mathematics abilities are built on a foundation of number computation and operational skills, it is understandable that the NAEP should include less direct assessment of arithmetic in the middle and high school years. While most would agree with Loveless and Diperna's contention (2000) that efforts to improve reform skills should not come at the expense of the basics of computation with whole numbers, decimals, and fractions, it does not follow that this basic content needs to be assessed beyond the grade levels at which it is appropriately taught and learned.

Trends

Barton and Coley (1998) compared NAEP trends that resulted from both cohort and cross-sectional analysis. Using the Trend NAEP, these researchers compared achievement growth between ages 9 and 13 for two cohorts for the periods between 1978 and 1982 and between 1992 and 1996. Their analysis showed no difference in the amount of growth. However, a cross sectional study comparing both 9- and 13-year-olds'

[3] Mathews, J. "Computational Skills Slipping," *The Washington Post*, September 3, 2002

mathematics achievement in 1978 and 1996 suggested improvement for both groups, demonstrating that the way questions about growth are framed affects the nature of the finding. Focusing on State NAEP results for two states—one, the top achieving state in 1992 and 1996 and the other near the bottom of the list based on achievement—these observers noted that the achievement gain between the two years was identical. Thus, depending on the question asked of the longitudinal data, the effectiveness of mathematics education in the states could be considered either very different or equivalent.

Focusing on trends in Black/White achievement, Phillips (2000), based on a meta-analysis of many studies, reported an average effect size of roughly .8 for the gap between the two races over 12 grades. Interestingly, this same author noted that, based on two national surveys, at grade 6 the gap between races is about .1 *SD*, but increased by this amount in each succeeding year, thus suggesting a widening of the gap in middle and, probably, in high school. However, the analysis of State NAEP mathematics results by Barton and Coley (1998) and Shaughnessy et al. (1997) did not show such a widening of the gap between races at either grade 4 or grade 8.

Comparing State Assessments and NAEP Trends

Several researchers have noted the limitations inherent in comparing the State NAEP with statewide assessment results. Linn (2000) and Amrein and Berliner (2002) observed that the familiarity effect, which results when annual use of tests have the same format and cover the same domain, often results in an upward bias of estimates of student learning. Feuer, Holland, Green, Berthenhall, and Hemphill (1999) cited differences in content coverage, item format, test administration procedures, and intended use and their associated consequences as factors limiting comparisons. Even where state and NAEP test frameworks match well, subtle differences between state curriculum and instruction and those areas targeted by the State NAEP may compromise the comparability of the results from the two assessments (Kenney & Silver, 1998).

While the State NAEP is designed to reflect "many of the states' emphasis and objectives in addition to what various scholars, practitioners, and interested citizens believed should be included in the curriculum" (Allen, Jenkins, Kulick, & Zelenak, 1997), the large scale assessments administered by states are intended to cover the specific objectives and frameworks mandated by each specific jurisdiction. Some (Klein, Hamilton, McCaffrey, & Stecher, 2000b; Amrein & Berliner, 2002) have argued that NAEP should be considered a benchmark for judging the validity of state tests. These writers seem to believe that there is a generalized domain of "mathematics" that is best represented by the NAEP domain and for which state tests should provide valid inferences. However, citing Texas as an example, Mehrens (2000) argues that if state tests are well aligned with their state frameworks, they meet the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, National Council of Measurement in Education, 1999) for content validity. Because the criteria for validity are related to the purposes for which a test is intended, state mandates and frameworks should serve as the criteria for

judging validity. It is important that those who interpret the results of assessments keep in mind the purposes for which the tests were designed and the specific domains for which valid inferences can be made.

Using State NAEP mathematics results to evaluate the impact of SSIs, Kane (2002) hypothesized that the assessment and accountability context of a state is an important factor affecting achievement gain. Haney (2000) contended that the Texas graduation requirement associated with the Texas Assessment of Academic Skills (TAAS) has had the effect of causing attrition among low-achieving students, thus contributing to the impression that student performance is improving when the more plausible explanation is that the testing population is changing. While studies by Carnoy, Loeb, and Smith (2001) confirm increases in student retention since the beginning of reform in Texas, their analysis did not support the attribution of this phenomenon to the TAAS graduation requirement. Amrein and Berliner (2002) have argued that, even though results may indicate improved achievement, high stakes accountability programs based on state assessment results have an unintended, negative effect on students' ability to transfer and apply mathematics from one context to another (p. 18). As evidence for this deduction, the authors cite achievement trends on the ACT, SAT, AP, and State NAEP for 18 states with high school graduation tests (HSGTs). Their argument for using these tests is that they " . . . overlap the same domain as state tests" (Amrein & Berliner, 2002, p. 2). While there is, no doubt, some overlap between the content domains of these tests, the differences in purposes, designs, and target populations can be considerable. Because of these differences, the expectation that their results will be similar is dubious.

Table 3.2 shows the gains and losses on the means of State NAEP scale scores in 18 states with HSGTs. Correlations between NAEP mathematics score change and exclusion rate change is also depicted in Table 4.2 because of the authors' (Amrein & Berliner) contention that, in some states, changes in exclusion rates were responsible for gains.[4] Based on the data, the authors concluded that HSGTs did not consistently improve

Table 3.2
Gains and Losses in State NAEP Scale Score Means for Grade 4 and Grade 8 in 18 States with High School Graduation Tests over Four Time Periods[1]

		Time Period											
		1990-1992			1992-1996			1996-2000			1990(92)[3] -2000		
Grade		Gains	Losses	r^2	Gains	Losses	r	Gains	Losses	r	Gains	*Losses*	r
4	**Total**				6	7	.00	6	5	.45	8	3	.39
	SSI				3	5		4	1		4	2	
	Non-SSI				3	2		2	4		4	2	
8	**Total**	2	8	.00	5	5	.00	9	2	.35	5	4	.53
	SSI	2	5		2	5		6	1		2	3	
	Non-SSI	0	3		3	0		3	1		3	1	

Note 1 Data summarized from Amrein and Berliner (2002).
Note 2 Correlation between NAEP mathematics score change and exclusion rate change.
Note 3 Grade 4 was not tested in 1990 so the time period is 1992-2000.

[4] State NAEP weighting procedures are designed to compensate for such changes in exclusion rates (Allen et al., 1997)

performance on the grade 8 NAEP mathematics test. However, acknowledging that the data are scant and that grade 8 is more likely than grade 4 to be influenced by an HSGT, during the most recent period, a relationship between having a HSGT and showing a gain in State NAEP mathematics achievement is evident. There would also appear to be an advantage for the SSI states with HSGTs during that period.

Another potential negative consequence of high-stakes accountability based on statewide assessment is narrowing test domains and lowering standards. Linn, Baker, Herman, and Koretz (2002) note that this is most likely in cases where accountability expectations are unreasonable. These writers cite the requirement of the *No Child Left Behind Act of 2001* (*NCLB*) (Public Law 107-110) that calls for all students in schools to score above the proficient level within 12 years. Because of the difference in the rigor of standards across states, this federal legislation, in its present forms, will likely have the effect of motivating states with more rigorous standards to reduce the breadth and depth of their standards to a point where the requirements of *NCLB* become feasible. In the case of mathematics, such a consequence might result in reducing the similarity between state tests and NAEP. The differences in content between the Texas statewide assessment and NAEP, noted by Klein et al. (2000b), possibly resulted from that states' desire to establish accountability requirements that were perceived to be within reach of most schools. Because of the requirements of *NCLB*, we might expect additional states to adopt Texas-like standards, thus reducing the comparability of NAEP and various statewide tests.

CHAPTER 4

SSI AND NON-SSI ACHIEVEMENT USING THE STATE NAEP: DESCRIPTIVE ANALYSIS

Introduction

Use of the State NAEP data allows us to track the change in academic performance in each state that voluntarily participated in the assessment. At present, State NAEP results are available for grade 8 for four years—1990, 1992, 1996, and 2000—and for grade 4 students for three years—1992, 1996, and 2000. This analysis includes the years of 1992, 1996, and 2000, when both grade 4 and grade 8 data were available. The achievement scales used in the State NAEP range from 0 to 500. The results are provided for a composite score and scores on each of five mathematics content strands (i.e., Number Sense, Properties, and Operations; Measurement; Geometry and Spatial Sense; Data Analysis, Statistics, and Probability; and Algebra and Functions). Using IRT procedures, the scale scores from each of the State NAEP assessments are linked to each other to make them comparable across assessment years. Thus, these scores and procedures enable us to monitor the trends of student performance in each state over the years of 1992, 1996, and 2000. In this chapter, we focus on identifying differences in mathematics scale scores between SSI and non-SSI states for grades 4 and 8. The results of the descriptive trend analysis are based on 27 states with data available over the three assessment years. Of the 27 states, 14 are SSI states and 13 are non-SSI states. Additional comparisons are provided in Appendices A and B for this chapter, which present results for all participating states. Previous work with a different sample is summarized in Appendix C.

This chapter consists of two main sections: one on the results of the trends of grades 4 and 8 students, and the other on the cohort growth results from grade 4 (1996) to grade 8 (2000). Within each of these sections, results are presented for the total group, for males and females, and for Whites, Blacks, and Hispanics. In addition to the comparison of composites and of the five content strands, the gaps found between different gender and ethnic groups are also reported.

Trends in Average Scale Scores over 1992, 1996, and 2000

Composite Scores

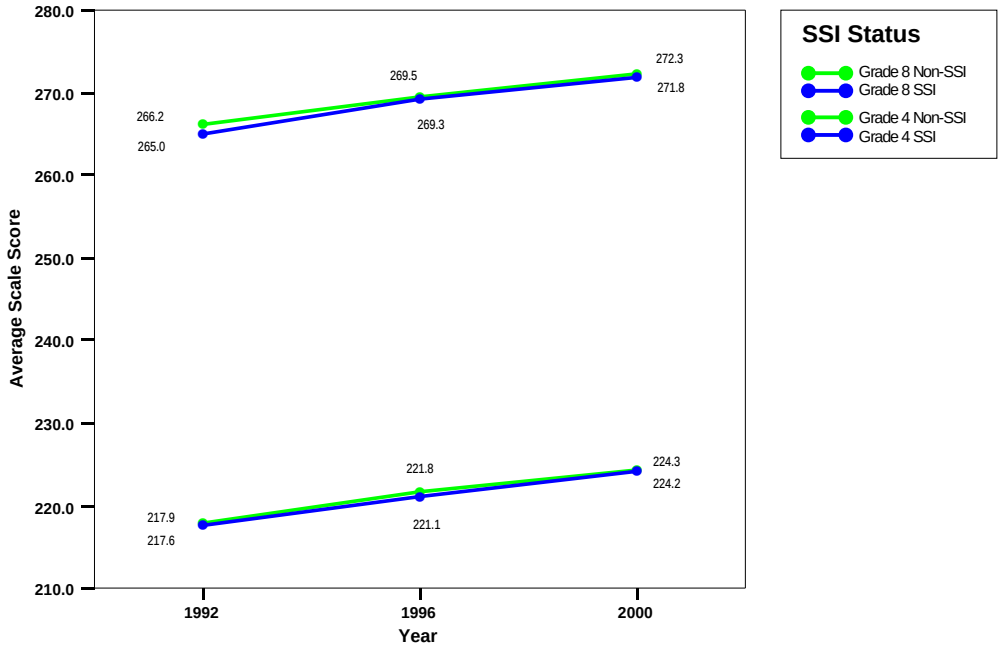
Total Group

In mathematics in both the SSI states and non-SSI states, students show continued increases in average scale scores from 1992 to 2000 for both grade 8 and grade 4. Overall, the average performance across the 14 SSI states was slightly lower than the average performance for the 13 non-SSI states across all three years. The gap between SSI states and non-SSI states was, however, hardly noticeable after 1996.

The average scale score for grade 8 mathematics from 1992 to 2000 showed a 6.8-point increase for the 14 SSI states and a 6.1-point increase for the 13 non-SSI states. The increase by students in the SSI states was slightly higher than by students in the non-SSI states by 0.7 points. In 1992, at an early stage of the SSI program, the 14 SSI states averaged 1.2 points less than the 13 non-SSI states. Since 1996, however, the difference was within 0.5 points, although the average score of the SSI states was still lower than that of non-SSI states.

Similarly in grade 4, both the SSI states and non-SSI states made a slight performance gain in average scale scores from 1992 through 2000. The SSI states scored only slightly lower than the non-SSI states. Over the eight years, the increase in the average scores by students from SSI states was 6.6 points, while the increase in the average scores by students from non-SSI states was 6.4.

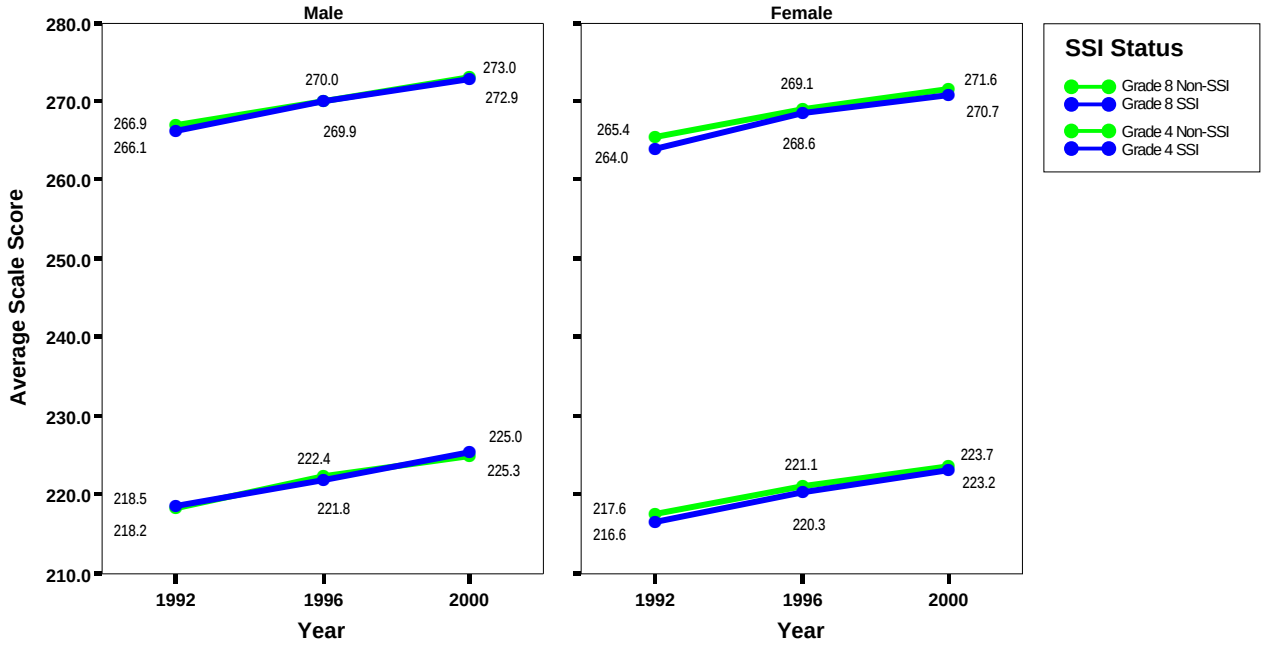
Figure 4.1. Trends in average scale scores, by SSI status: Trend Group 92-00 (14 SSI and 13 non-SSI states).



Gender

The trend in average scale scores for male and female students shows similar, but slightly different, patterns across grades and SSI status (Figure 4.2). Male and female students in both SSI and non-SSI states showed increases in average scale scores in all years, with male students scoring higher than female students. As shown in previous overall scale-score trends (Figure 4.1), SSI states in general had lower average scale scores than non-SSI states; however, male grade 4 students of the SSI states scored higher than those of the non-SSI states in 1992 and 2000. Overall, both male and female students in SSI states gained slightly more than those in non-SSI states, which narrowed the difference in scores between SSI states and non-SSI states.

Figure 4.2. Trends in average scale scores, by gender and SSI status: Trend Group 92-00 (14 SSI and 13 non-SSI states).



Ethnicity

Overall, the average scale scores of racial subgroups (Whites, Blacks, and Hispanics) show increases in mathematics performance across the assessment years. But substantial variety in mathematics performance among racial subgroups was evident at grades 4 and 8. Although the SSI states had lower scale scores than the non-SSI states across years in both grades (Figure 4.1), the SSI states, in general, showed higher scores than the non-SSI states across the three racial subgroups, which may have been caused by the insufficient sample size of the subgroups. Nonetheless, in both grades 4 and 8, White students outperformed Black and Hispanic students, while Hispanic students scored higher than Black students (Figure 4.3).

From 1992 to 2000, both grade 4 and grade 8 White students in the SSI and the non-SSI states gained in their NAEP mathematics composite scores, reflecting the gains for the state as a whole.

Average scale scores in both grades for Black students improved more for the SSI states than they did for the non-SSI states—12 SSI, 8 non-SSI, the number of states with minority populations that were large enough to report data for all three years.[1]

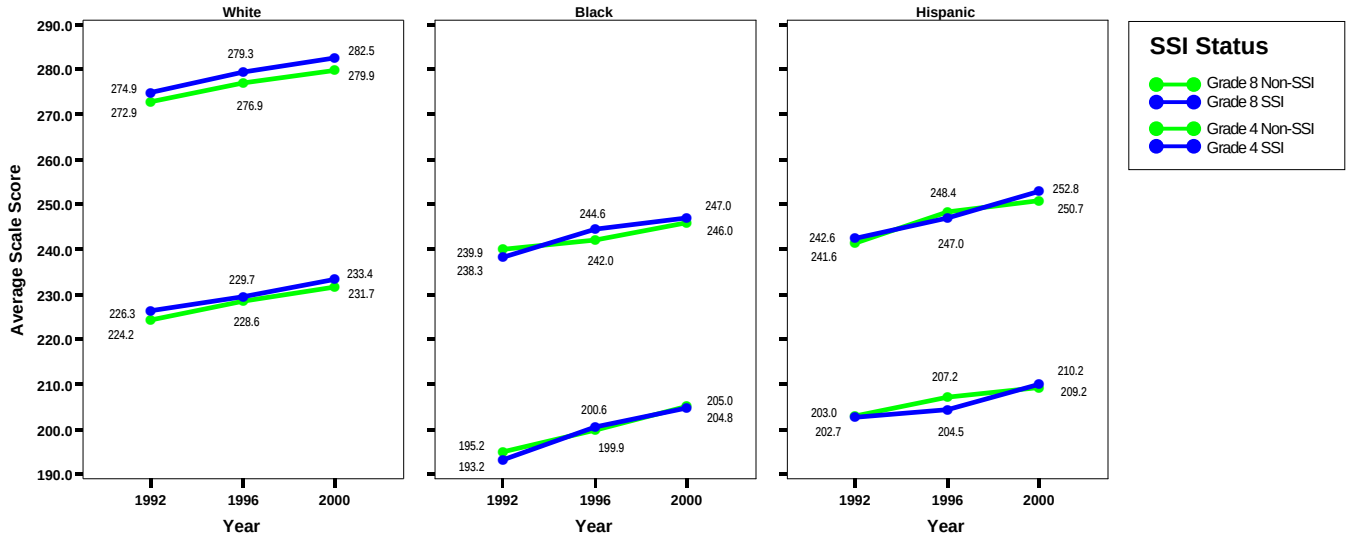
In 1992, grade 8 Black students in the SSI states had a mean mathematics score that was 1.6 points below Black students in the non-SSI states. Black students in the SSI states outperformed those in the non-SSI states by 2.6 points in 1996 and by 1 point in 2000. From 1992 to 2000, the mean score for grade 8 Black students increased both for SSI states and non-SSI states, 8.7 and 7.1 respectively. From 1992 to 1996, the mean score for grade 8 Black students from SSI states increased considerably, compared to increases by students from non-SSI states, a 6.3-point increase compared to a 2.1-point increase.

In grade 4, the mean mathematics score of Black students in the SSI states increased by 7.4 points between 1992 and 1996, compared to the 4.7-point increase over this period by Black students in the non-SSI states; thus, Black students from SSI states outperformed those from non-SSI states. In 2000, however, the mean score for grade 4 Black students in the SSI states was slightly lower than the mean score for Black students in non-SSI states, by 0.2 points.

Hispanic students also showed an overall increase in scores across the three years in both grades. Grade 8 Hispanic students in the 11 SSI states with sufficient numbers to be included in the analysis outperformed those in the 12 non-SSI states by one point in 1992; however, the non-SSI states scored higher than the SSI states by 1.4 points. The score of the SSI states regained superiority over that of the non-SSI states in 2000 by 2.1 points. In grade 4, the score of the SSI states was slightly below the score of the non-SSI states by 0.3 points, this gap widening in 1996 to 2.7 points. However, grade 4 Hispanic students in the 11 SSI states outperformed those in the 12 non-SSI states by 1 point in 2000.

[1] A minimum sample size of 62 students per state was required to report the results for any subgroups (Mullis et al., 1991.)

Figure 4.3. Trends in average scale scores, by race and SSI status: Trend Group 92-00 (14 SSI and 13 non-SSI states*).



* Due to the insufficient sample size of the subgroups, results are based on 12 SSI states and 8 non-SSI states for Blacks, and 11 SSI states and 12 non-SSI states for Hispanics.

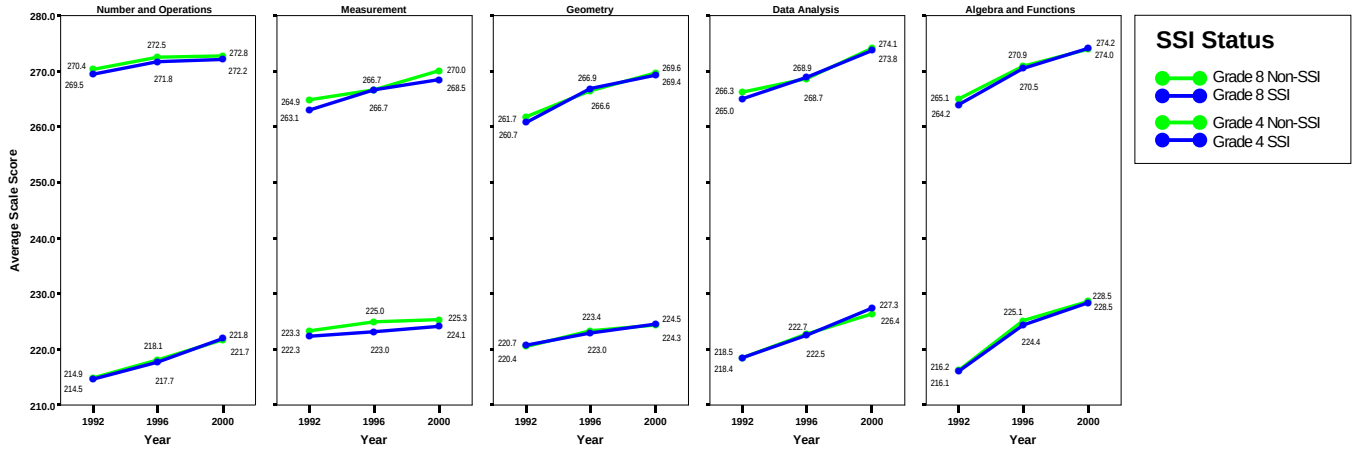
Subtopic Scores

Total Group

Very few differences existed in the pattern of achievement among the five mathematics topics tested by NAEP in the three testing years and between SSI and non-SSI states (Figure 4.4).

On each of the five mathematics topics, SSI states had average scale scores quite similar to or slightly below those of the non-SSI states for all testing times for both grade 4 and grade 8. As expected from previous examples, both SSI states and non-SSI states had increased gain in mathematics performance on each of the five mathematical topics. In general, in 1992 grade 8 students scored higher on Number/Operations followed by Data Analysis, Algebra/Functions, Measurement, and Geometry.[2] Grade 4 students in general scored lower on Number/Operations than on the other four topics. The greatest gains at both grade 4 and grade 8 levels were in Algebra/Functions. The smallest gain was in measurement for grade 4 and in Number/Operations for grade 8. In addition, the biggest gap between the SSI states and the non-SSI states was shown in Measurement.

Figure 4.4. Trends in average scale scores on content strands, by SSI status: Trend Group 92-00 (14 SSI and 13 non-SSI states).



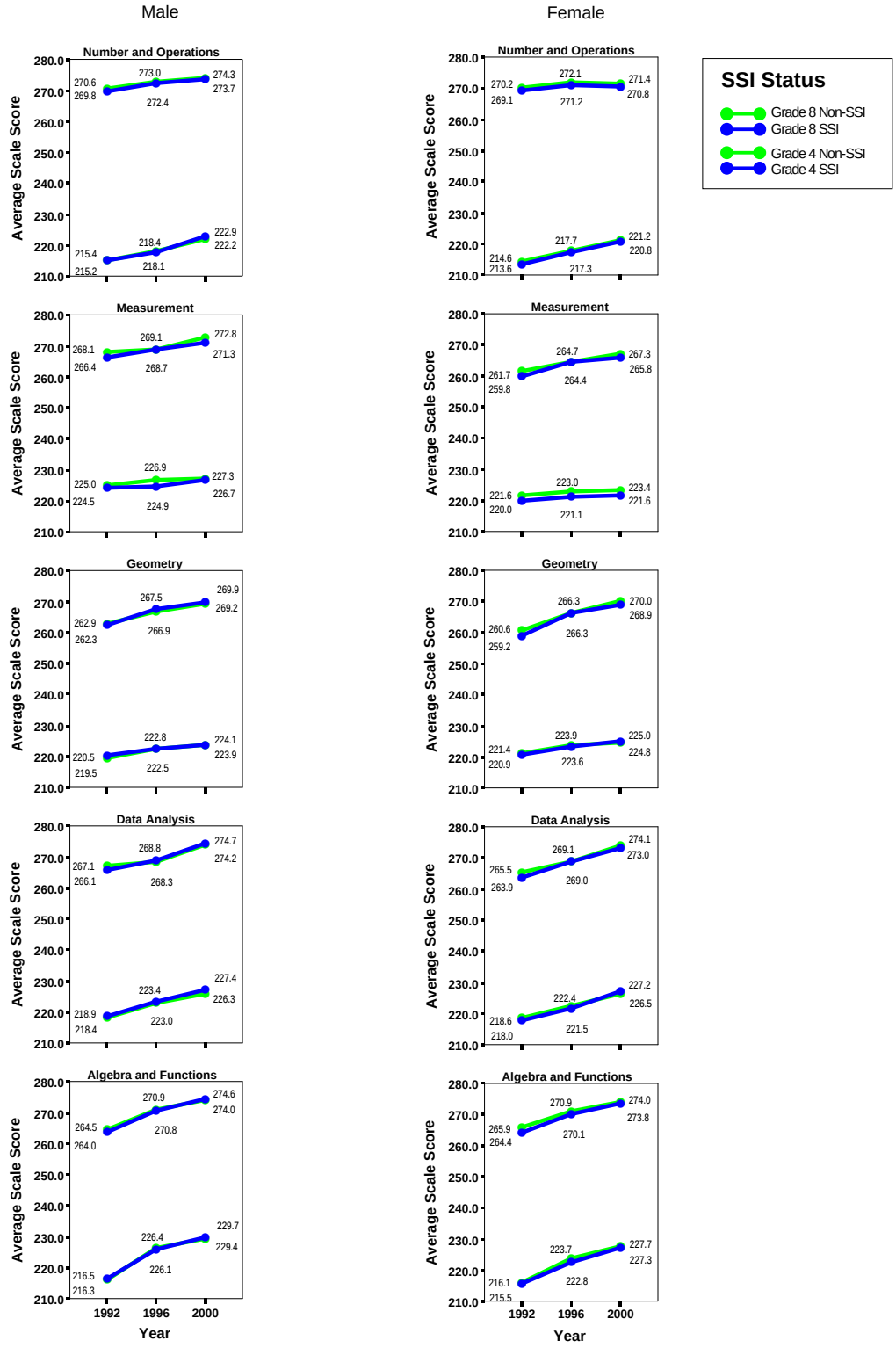
[2] However, by 2000, grade 8 students had the highest scale scores in Data Analysis and Algebra and Functions and the lowest scores in Measurement and in Geometry.

Gender

Gender trends by content area for SSI states and non-SSI states were similar to those for all students (Figure 4.5). Male and female students made substantial progress in the five content strands over the three assessment years for each grade, while the differences between SSI and non-SSI states were hardly noticeable. Gender differences were not great on any of the content-strand scale scores, although males performed somewhat better than females.

The greatest gains on average scale scores in content strands were in Algebra and Functions at both grade levels. From 1992 to 2000, grade 8 male students gained 10 points for SSI states, while male students from non-SSI states gained 10.1 points. The gains for male grade 4 students were greater for both SSI states and non-SSI states (around 13 points). For female students in grade 8, SSI states and non-SSI states improved similarly, up to around 9 points in the Algebra/Functions content strands; Grade 4 female students gained around 12 points from 1992 to 2000 in Algebra/Functions.

Figure 4.5. Trends in average scale scores on content strands, by gender and SSI status: Trend Group 92-00 (14 SSI and 13 non-SSI states).



Ethnicity

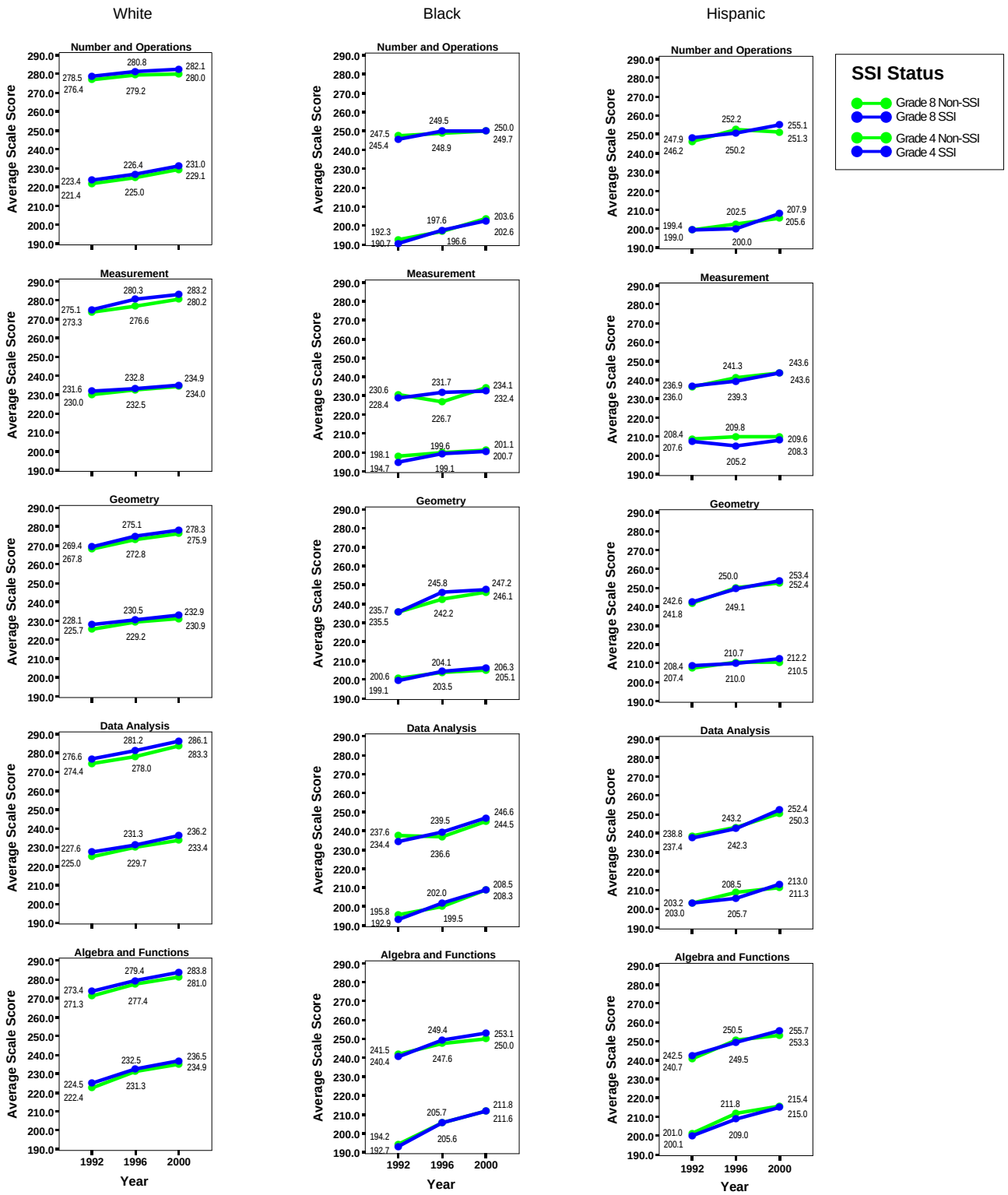
In the five mathematics content strands, the overall changes in performance among three racial subgroups were small, but there were cumulative increases, in general, from the assessment in 1992 to 2000 in grade 4 and grade 8. The score differences observed for racial subgroups in composite scale scores were also observed in five content-strand scale scores. White students scored higher than Hispanic students, who outperformed Black students across grades and assessment years.

There were varied patterns of average scale-score gains in five content strands for White, Black, and Hispanic students by SSI group status. Nonetheless, the SSI states gained slightly more in average scores across the three ethnicities and across the five content strands. Grade 8 White students from the SSI states, for example, gained more than those in the non-SSI states by 1.2 points.

For Black students in the SSI states, the gains were steady over the assessment years, while the scores of Black students in the non-SSI states dropped at some points. The increase in scores of SSI-state Black students was more apparent from 1992 to 2000. For example, in 2000, Black students from SSI states outperformed those from non-SSI states in both grades across the five content strands except for two cases. The reverse was true in 1992.

Overall, Hispanic students from both SSI and non-SSI states had the greatest gains in scale scores among the three ethnic groups at both grade 4 and 8. SSI states showed greater gains than non-SSI states in grade 4 and 8 across four content strands, while in measurement non-SSI states showed more gains. The gain scores of Hispanic students in Algebra/Functions—for grade 8, 13.2 points in SSI states and 12.6 in non-SSI states; for grade 4, 14.9 points in SSI states and 14.4 in non-SSI states—were the highest among all the increases made by any other ethnic groups in any of the content strands.

Figure 4.6. Trends in average scale scores on content strands, by race and SSI status: Trend Group 92-00 (14 SSI and 13 non-SSI states*).



* Due to the insufficient sample size of the subgroups, results are based on 12 SSI states and 8 non-SSI states for Blacks, and 11 SSI states and 12 non-SSI states for Hispanics.

Gaps Between Different Groups

Gender

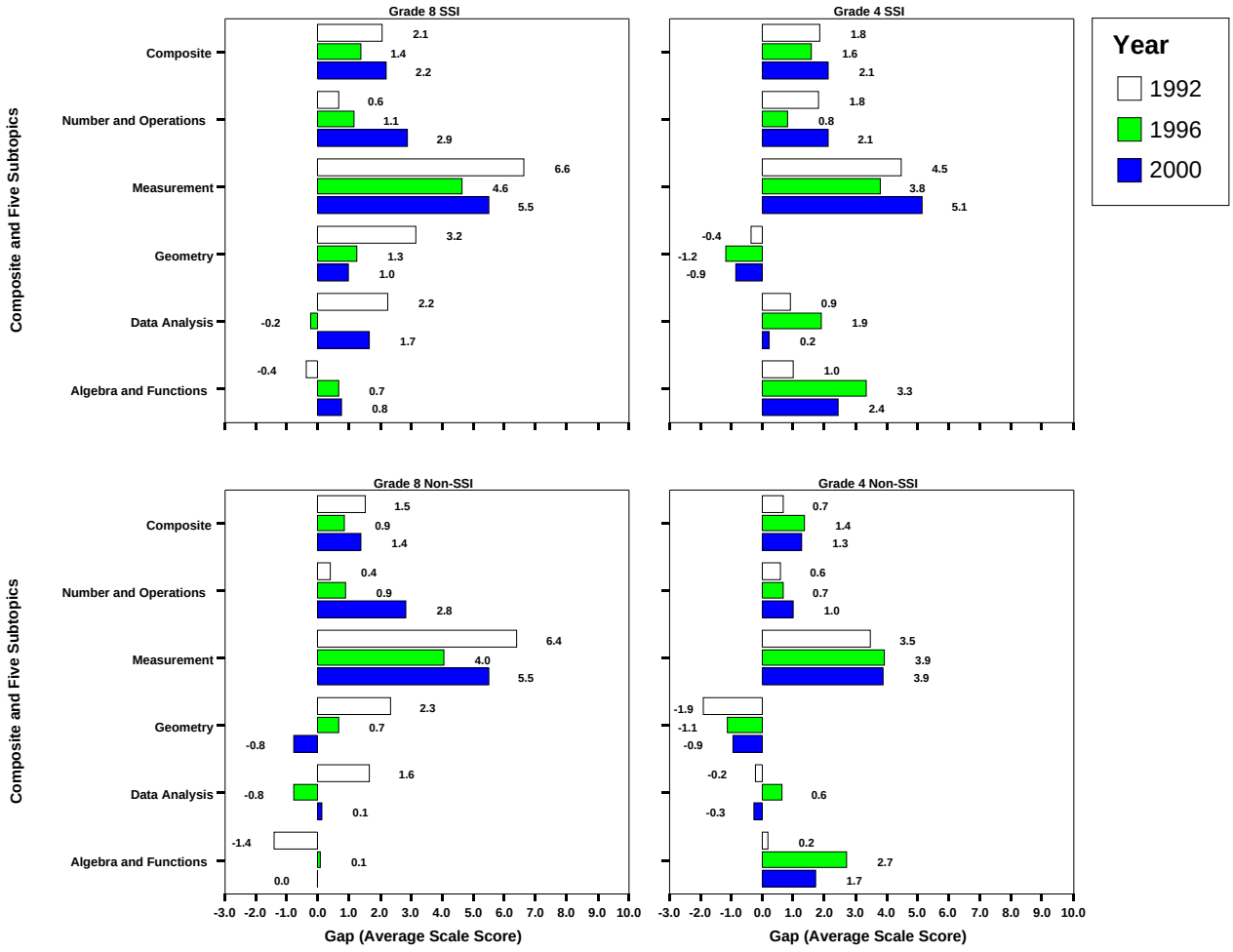
Although the gap between males and females on the composite score was moderate, ranging between .9 points and 2.2 at grade 8 and between .7 points and 2.1 at grade 4, the gaps on specific content strands were more noticeable. Males outperformed females in all groups (SSI and non-SSI in grade 8 and grade 4) on the Measurement strand.

The pattern of change within each content strand across the three years is really identical at grade 8 between the SSI and non-SSI states. This would imply no difference in variation in the gender gap by SSI states at grade 8.

Grade 8 male students in SSI states scored around two points higher than females in their mean mathematics composite score from 1992 through 2000. In non-SSI states, males had around a 1.5-point advantage across the years. Grade 8 students in SSI states thus showed greater consistent gaps between male and female students than did those in non-SSI states. Grade 4 students showed similar gender difference in average scores across the SSI and non-SSI states. The visible changes occurred in the Measurement strand. In both grades in SSI states and in non-SSI states, male and female gaps in Measurement were the highest of all. What is the most interesting is the difference in the Geometry strand score in grade 8. While SSI male students showed consistent gains over female students across the three years, in non-SSI states, female students reversed the pattern and outperformed male students by 0.8 points in 2000.

But, there is little evidence from the NAEP data to indicate that a state's participation in an SSI had any relationship to lowering the achievement gap between male and female students. The mean mathematics composite score for female students and male students differed at most by two points at grades 4 and 8, at any of the testing times, and for both SSI and non-SSI states. The mean score for grades 4 and 8 for both male and female students increased over time. However, the two-point difference in male scores in non-SSI states in grade 8 in 1992 had decreased by one point by 1996. The two-point gap at grade 8 in SSI states remained the same over the three testing times. In 1992 at grade 4, the mean mathematics composite score between male and female students differed by two points for both SSI and non-SSI states. This gap was lowered to one point in SSI states in 1996, but remained the same in non-SSI states.

Figure 4.7. Gender differences (males versus females) in average scale scores, by SSI status: Trend Group 92-00 (14 SSI and 13 non-SSI states).



At grade 4, the pattern of change by SSI and by non-SSI states varies some in that males in SSI states improved in performance more than females over time. This is particularly true in Number and Operations and Algebra/Functions. The small variation between the SSI group and the non-SSI group at grade 4 is not great enough to be significant. Overall, SSI status was not related to variation in scores between male and female students, as evident in average scale scores.

Whereas, at grade 8, the gap between White and Black students from SSI states decreased in 1996 and then increased some in 2000, at grade 4, the SSI states generally maintained the reduction in the gap attained in 1996 in 2000. The gap by students in SSI states declined on each of the six scales. In contrast, the gap by students in non-SSI states was higher in 2000 than in 1992 on two of the six scales, Measurement and Geometry.

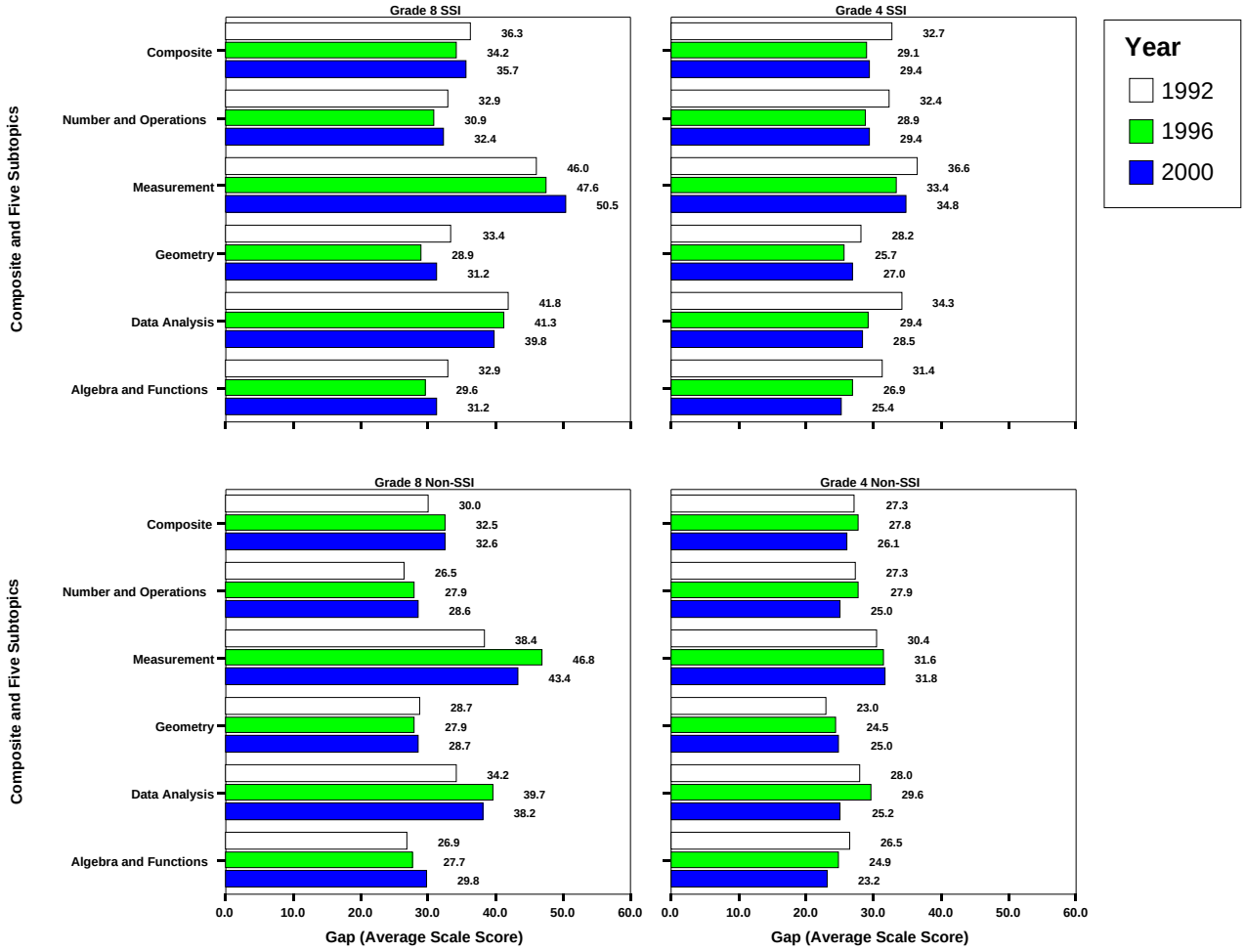
Although the gap between White and Black students grew smaller in the SSI states, the gaps still remained higher than those of the non-SSI states.

Ethnicity

There were significant differences in the composite score and in the five content strand scale scores for White and Black students, but the gaps between the two groups remained rather stable over time (Figure 4.8). Although White students scored higher than Black students in composite scores and in the five content strands scores in both grades, there were no consistent patterns across six scale scores among SSI and non-SSI states. Again, however, the Measurement strand discriminated the most between White and Black students for both grades across the three years, regardless of the SSI status.

Grade 4 score gaps of White and Black students showed a somewhat contrasting pattern for SSI and non-SSI states. The gaps between White and Black students were greatest for SSI states in 1992, reducing the gaps across years after that. On the other hand, the gaps between Whites and Blacks were rather stable for non-SSI students.

Figure 4.8. Differences in average scale scores between White and Black students, by SSI Status: Trend Group 92-00 (14 SSI and 13 non-SSI states*).



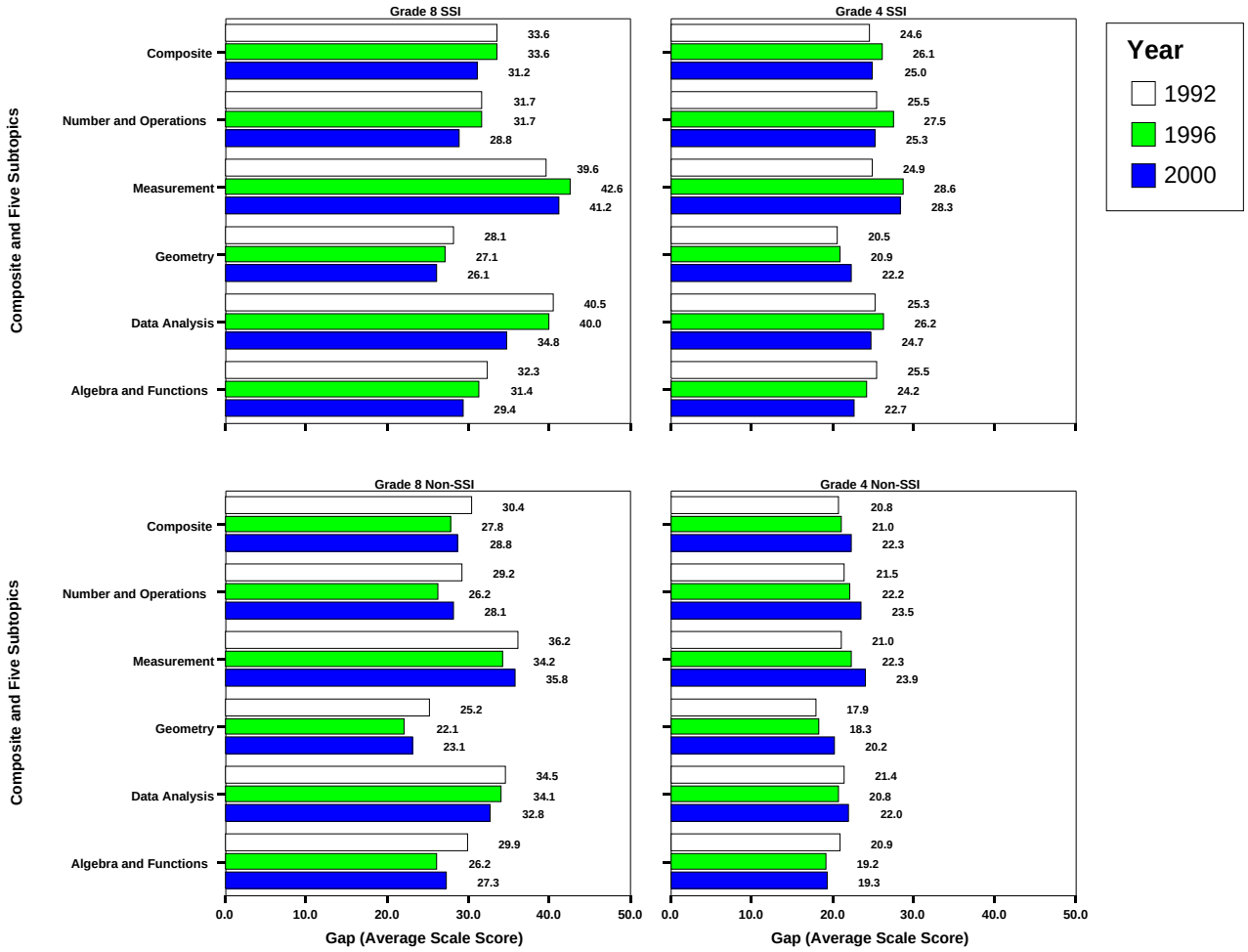
* Due to the insufficient sample size of these subgroups, results are based on 12 SSI states and 8 non-SSI states.

At grade 8, with the exception of Measurement, the gaps between White and Black students from SSI states decreased between 1992 and 1996, with a slight rebound in 2000. However, the gap for grade 8 SSI students in 2000 was less than in 1992, except for Measurement. This is in contrast to non-SSI students, whose gap increased on the composite scale and on three of the five content strands.

Regarding score differences of Whites and Hispanic students, there were different trends for students in grades 4 and 8 (Figure 4.9). In grade 8, the gaps in SSI states and non-SSI states narrowed between 1992 and 2000. However, there were greater increases in score gaps for grade 4 students from 1992 to 2000. Grade 8 on the Measurement content strand showed a different pattern of score-gap change. The gap in SSI states increased in 1996 and dropped in 2000, but the score gap in non-SSI states dropped in 1996, followed by an increase to almost the initial level in 2000.

The gaps between White and Hispanic students in non-SSI states were smaller than those in SSI states in the composite and the five average scale scores.

Figure 4.9. Differences in average scale scores of White and Hispanic students, by SSI status: Trend Group 92-00 (14 SSI and 13 non-SSI states*).



* Due to the insufficient sample size of these subgroups, results are based on 11 SSI states and 12 non-SSI states.

At grade 8, the gaps between White and Hispanic students grew smaller over time for those from SSI states. The most noticeable reduction was between 1996 and 2000. The gap between White and Hispanic students from non-SSI states declined some between 1992 and 1996, but increased on all six scales between 1996 and 2000. This was also true for grade 4 in the non-SSI states.

At grade 4 in the SSI states, the gap between Whites and Hispanic students in general increased from 1992 to 1996 and then declined from 1996 to 2000, to a level at or less than the gap in 1992 for four of the six scales.

Cohort Growth in Average Scale Scores from Grade 4 (1992) to Grade 8 (1996)

This performance comparison of cohort growth at two grade levels (grade 4 and grade 8) allows us to track achievement growth of the same group of students after four years.

Composite Scores

Total Group

Both SSI and non-SSI states had substantial cohort growth between grade 4 in 1992 and grade 8 in 1996 (Figure 4.10a). Students in non-SSI states scored essentially the same as those in SSI states in the two assessment years; for example, the grade 4 scale score in SSI states was 217.6, compared to 217.9 points in non-SSI states. After four years, grade 8 students in SSI states scored 269.3 points and their counterparts in non-SSI states scored 269.5. The cohort growth for SSI states and non-SSI states also was nearly the same, 51.7 points and 51.6 points, respectively. The relative pattern of cohort growth between SSI and non-SSI students as they progressed from grade 4 in 1996 to grade 8 in 2000 was nearly the same as for the previous cohort. However, of special note is that even though scores at grade 4 and grade 8 were higher than those grades in the previous four years, the growth over the four years was 1 point less.

Figure 4.10a. Cohort growth in average scale scores from 1992 to 1996, by SSI status: Trend Group 92-00 (14 SSI and 13 non-SSI states).

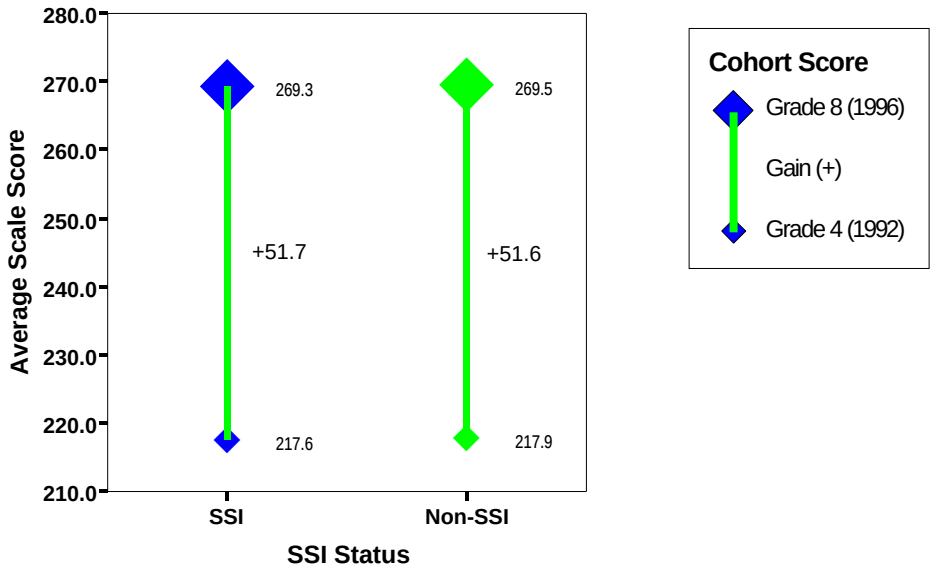
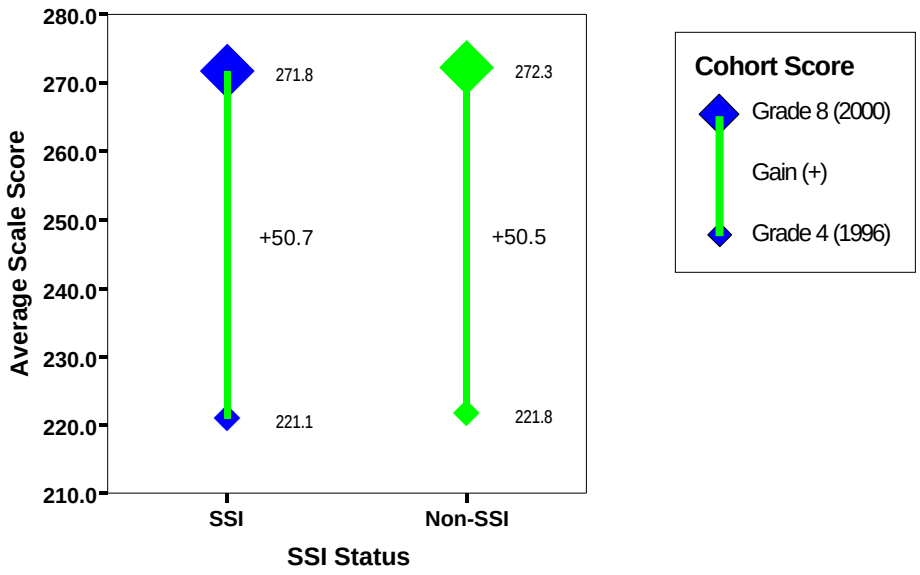


Figure 4.10b. Cohort growth in average scale scores from 1996 to 2000, by SSI status: Trend Group 92-00 (14 SSI and 13 non-SSI states).



Gender

Male and female cohorts in both SSI and non-SSI states showed performance improvement in average scale scores from 1992 to 1996 (Figure 5.11a). The average scores and gain scores made by both males and females in SSI and non-SSI states were all quite similar. The results were similar for the next four years, from 1996 to 2000, with one exception. Male students in SSI states maintained the same level of growth from 1996 to 2000 as in the comparable group in the previous four years. The other three groups had a slightly smaller growth of 1 point or more.

Figure 4.11a. Cohort growth in average scale scores from 1992 to 1996, by gender and SSI status: Trend Group 92-00 (14 SSI and 13 non-SSI states).

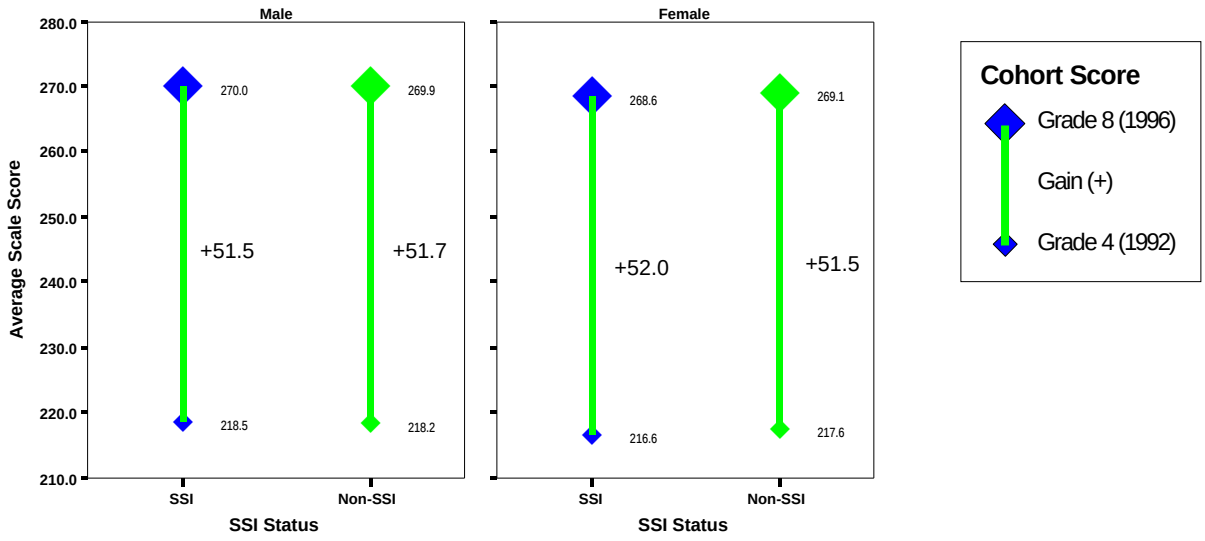
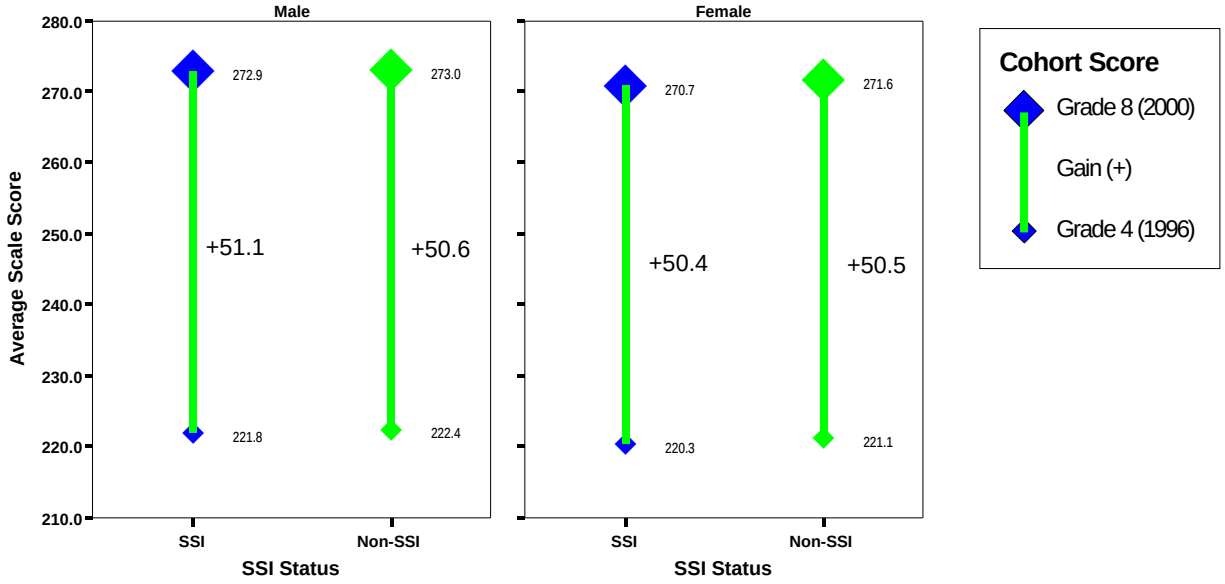


Figure 4.11b. Cohort growth in average scale scores from 1996 to 2000, by gender and SSI status: Trend Group 92-00 (14 SSI and 13 non-SSI states).

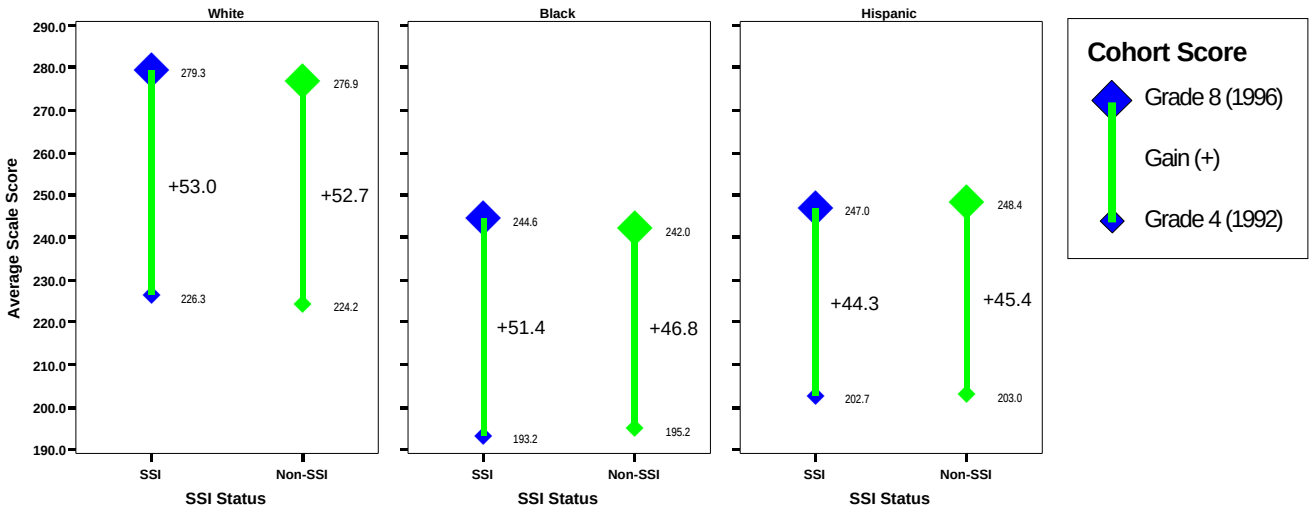


Ethnicity

The cohort growth of the three racial groups shows increases in average scale scores between the two assessment years, 1992 and 1996. The pattern of cohort growth varied among White, Black, and Hispanic students. White students made greater grains than Black and Hispanic students, while Black students made greater gains than Hispanic students (Figure 4.12a).

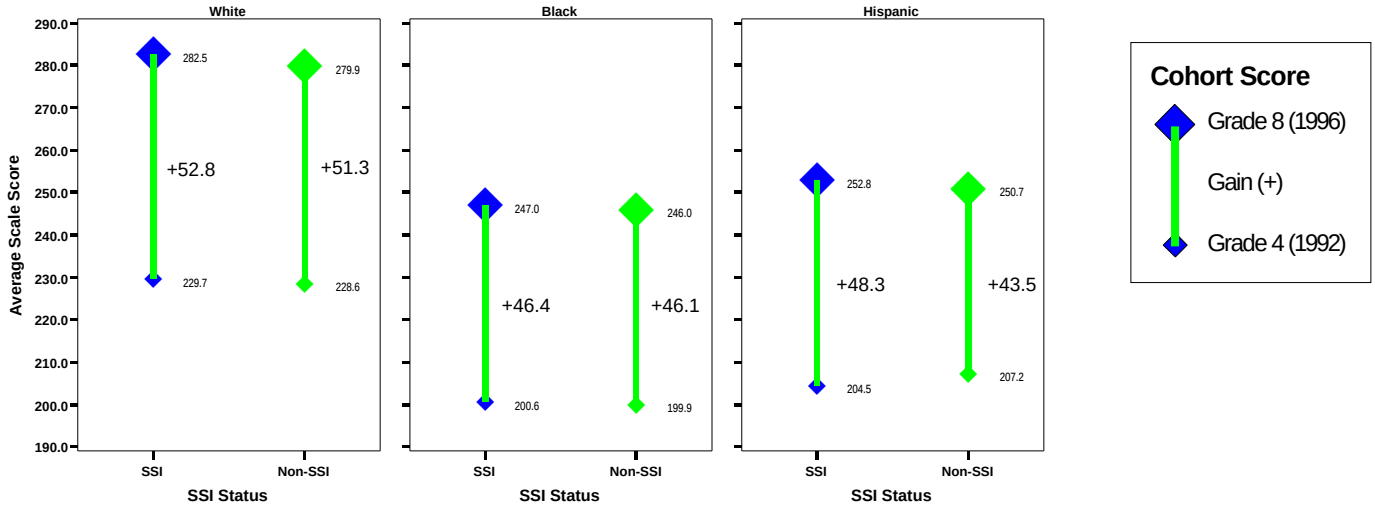
The results for Black students are encouraging for SSI states. Cohort growth of Black students in SSI states was 4.6 points higher than the growth in non-SSI states over the four-year timeframe (1992 to 1996). However, this large difference in growth from grade 4 to grade 8 was not sustained by the next cohort of Black students. From 1996 to 2000, Black students from SSI states experienced nearly the same growth as did Black students from non-SSI states (46.4 points and 46.1 points, respectively) (Figure 4.12b). However, Hispanic students from SSI states had a growth in NAEP scores from grade 4 (1996) to grade 8 (2000) that was higher than Hispanic students from non-SSI states (48.3 and 43.5, respectively). The growth in scores over the four years by Hispanic and Black students in both groups, still was 4 to 6 points less that the growth experienced by White students. It is interesting that White students from SSI states maintained the same level of growth between grade 4 and grade 8 as did the previous cohort, but the growth by White students from non-SSI states from grade 4 (1996) to grade 8 (2000) declined by 1.4 points.

Figure 4.12a. Cohort growth in average scale scores from 1992 to 1996, by race and SSI status: Trend Group 92-00 (14 SSI and 13 non-SSI states*).



*Due to the insufficient sample size of these subgroups, results are based on 12 SSI states and 8 non-SSI states.

Figure 4.12b. Cohort growth in average scale scores from 1996 to 2000, by race and SSI status: Trend Group 92-00 (14 SSI and 13 non-SSI states*).



* Due to the insufficient sample size of the subgroups, results are based on 12 SSI states and 8 non-SSI states for Blacks, and 11 SSI states and 12 non-SSI states for Hispanics.

Subtopic Scores

Total Group

Both SSI and non-SSI states had similar cohort growth in the five mathematics content strands over four years from 1992 to 1996. The results show some variations of cohort growth across five content strands (Figure 4.13a). Cohort students in both SSI and non-SSI states gained more in Number/Operations (up to 58 points), Algebra/Functions (up to 55 points), and Data Analysis (up to 51 points) than they did in Geometry (46 points) and Measurement (up to 44 points).

The cohort gains from 1996 to 2000 by SSI and non-SSI states on the five content strands only varied slightly from the prior cohort over the previous four years (Figure 4.13b). There were essentially no differences between those from SSI states and those from non-SSI states. From 1996 to 2000 both SSI and non-SSI groups gained slightly more than the previous cohorts in Measurement and Data Analysis. The cohort in both groups gained the same in Geometry. The 1996-2000 cohort did not gain as much as the prior cohort in Number/Operations and Algebra/Functions. This analysis indicates that the growth by cohorts from grade 4 to grade 8 did not differentiate between SSI and non-SSI states by different content strands. There was variation by content strand, but students in both SSI and non-SSI states had similar patterns in this variation.

Figure 4.13a. Cohort growth in average scale scores on content strands from 1992 to 1996, by SSI status: Trend Group 92-00 (14 SSI and 13 non-SSI states).

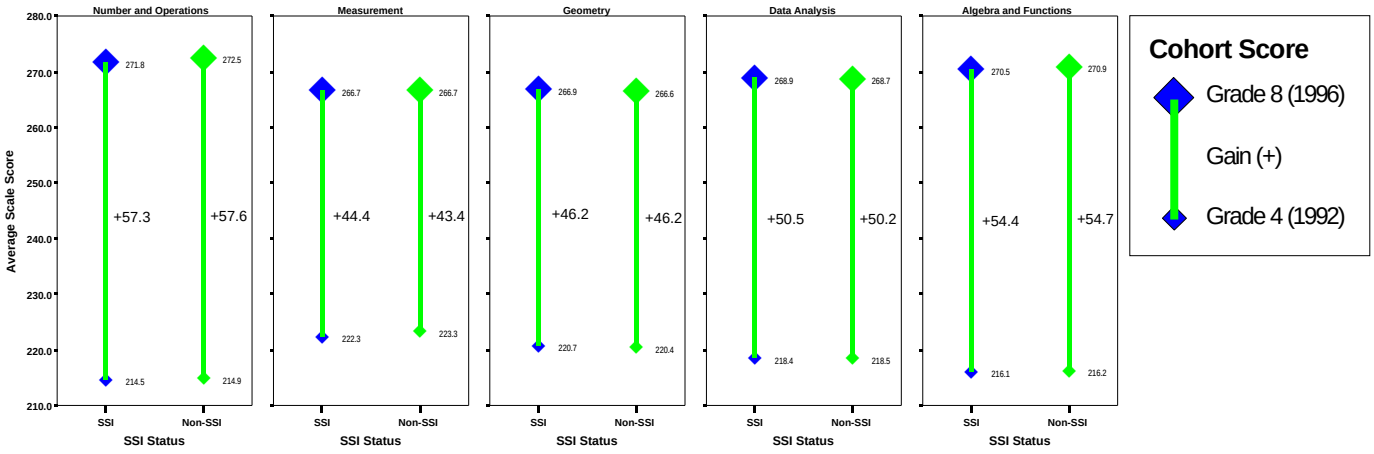
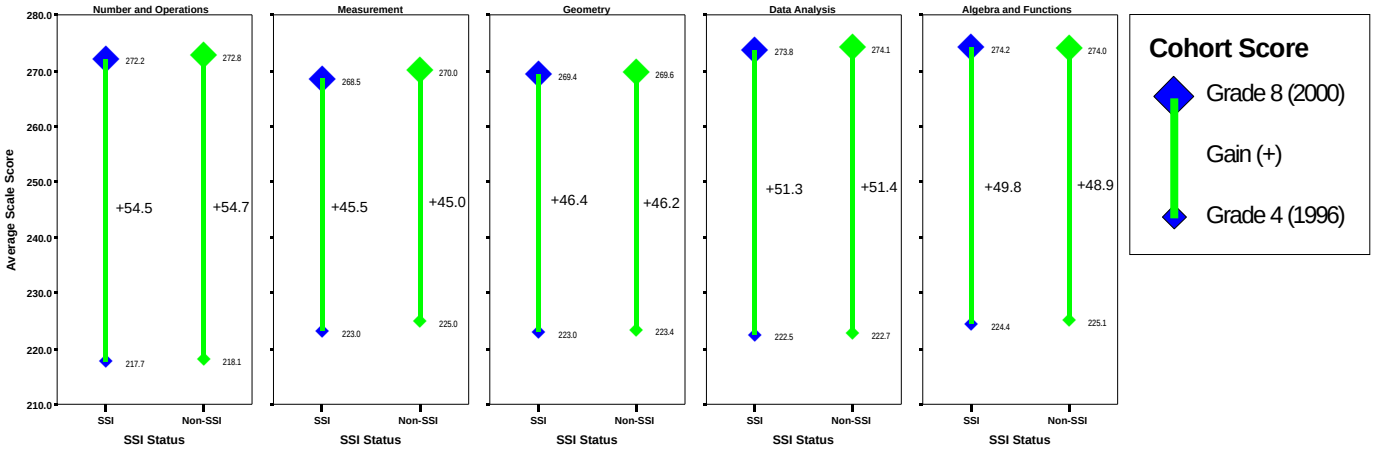


Figure 4.13b. Cohort growth in average scale scores on content strands, by SSI status from 1996 to 2000: Trend Group 92-00 (14 SSI and 13 non-SSI states).



Gender

We can see that the scores of male and female cohort students were quite similar in the five content strands (Figure 4.14a). They showed cohort growth of between 43 points and 58 points from 1992 to 1996. Most cohort gains were observed in Number/Operations and Algebra/Functions, and least growth was in Measurement for both male and female in both cohorts. From 1996 to 2000, the female cohort gained slightly less (around 2 scale points) than did the male cohort from grade 4 to grade 8 in four of the five content strands (Figure 4.14b). The female and male cohorts only gained the same in Data Analysis. For both four-year periods, 1992 to 1996 and 1996 to 2000, there were no differences between the SSI states and the non-SSI states in comparative performance of male and female students by content strand.

Figure 4.14a. Cohort growth in average scale scores on content strands from 1992 to 1996, by gender and SSI status: Trend Group 92-00 (14 SSI and 13 non-SSI states).

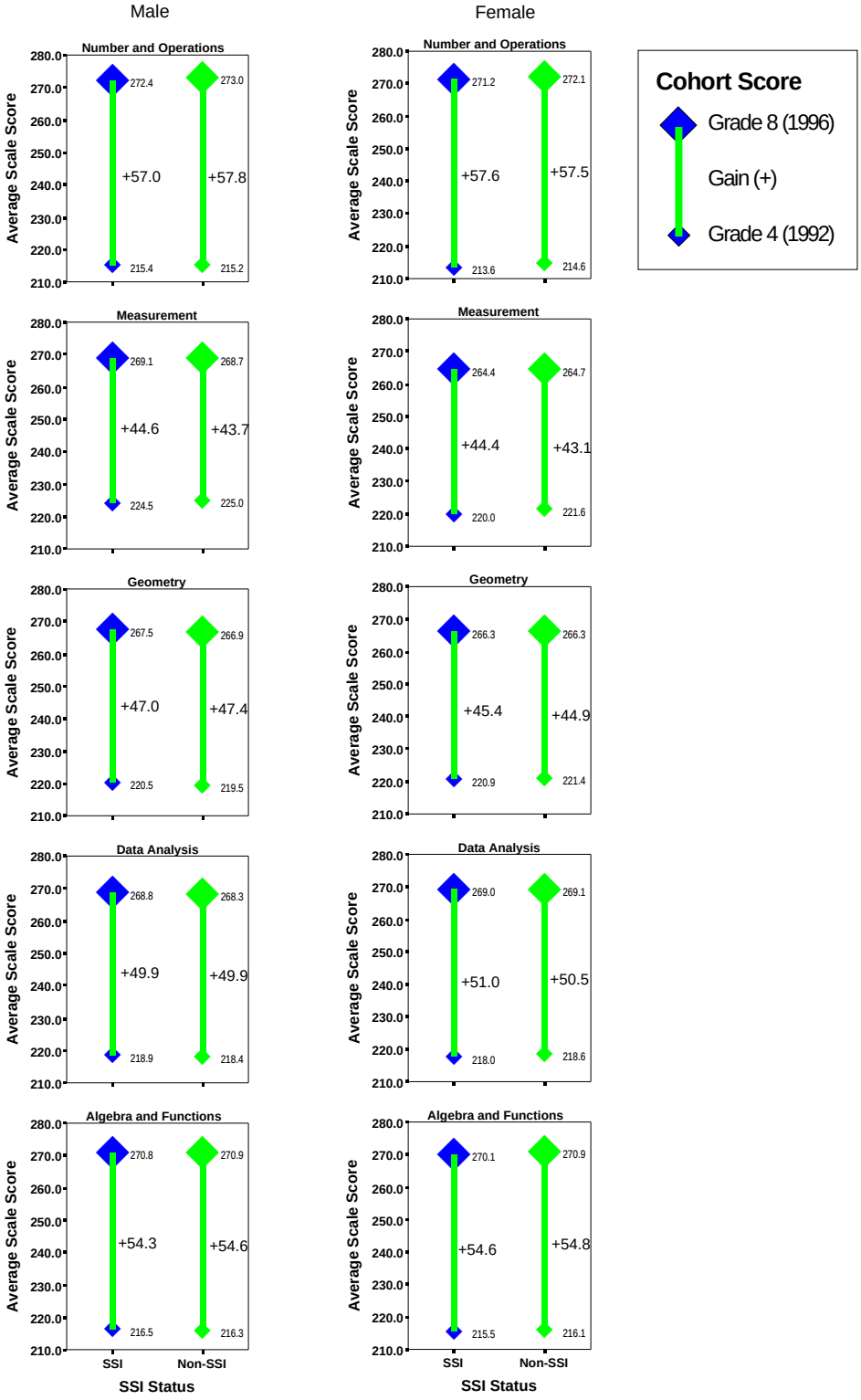
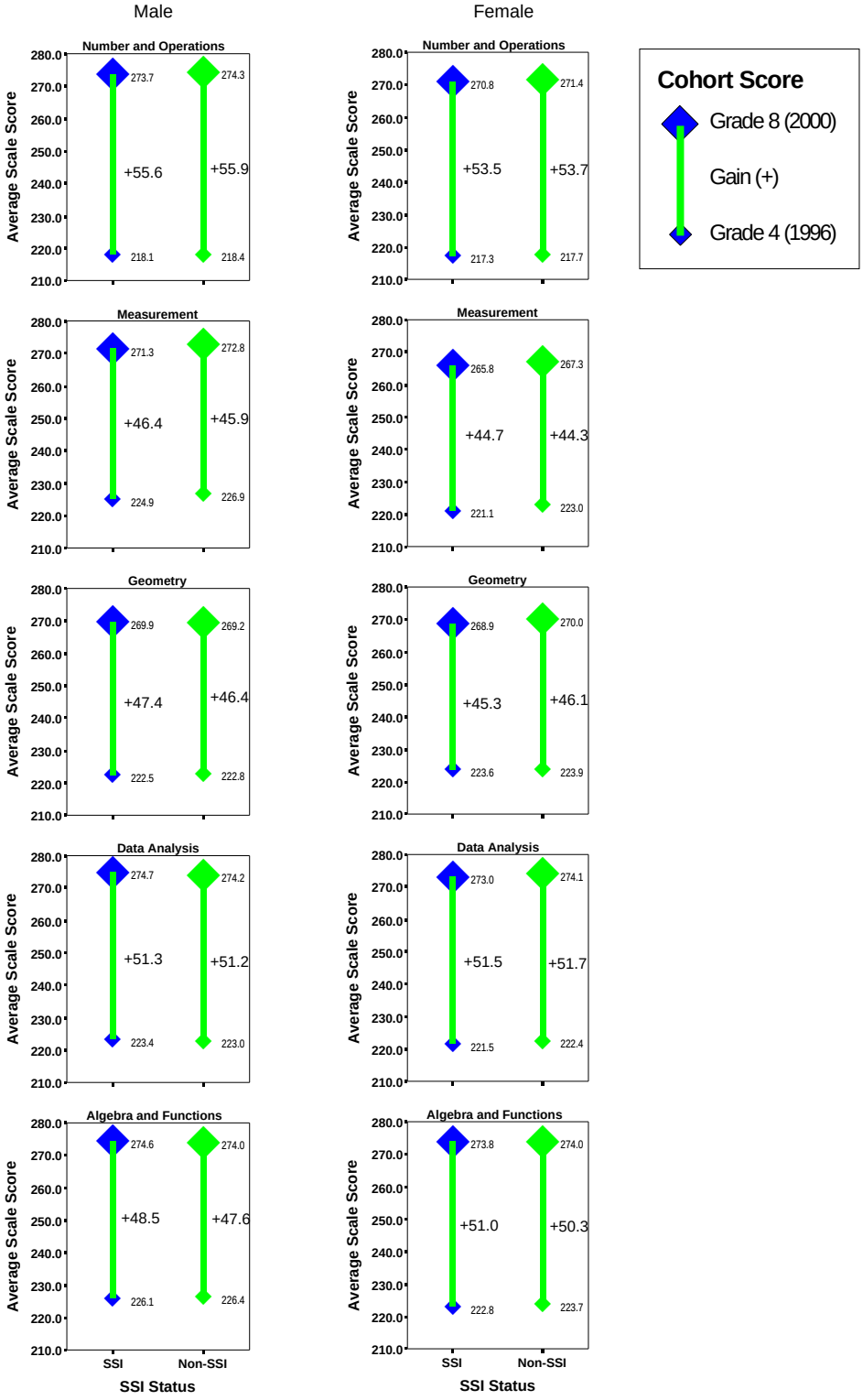


Figure 4.14b. Cohort growth in average scale scores on content strands from 1996 to 2000, by gender and SSI status: Trend Group 92-00 (14 SSI and 13 non-SSI states).

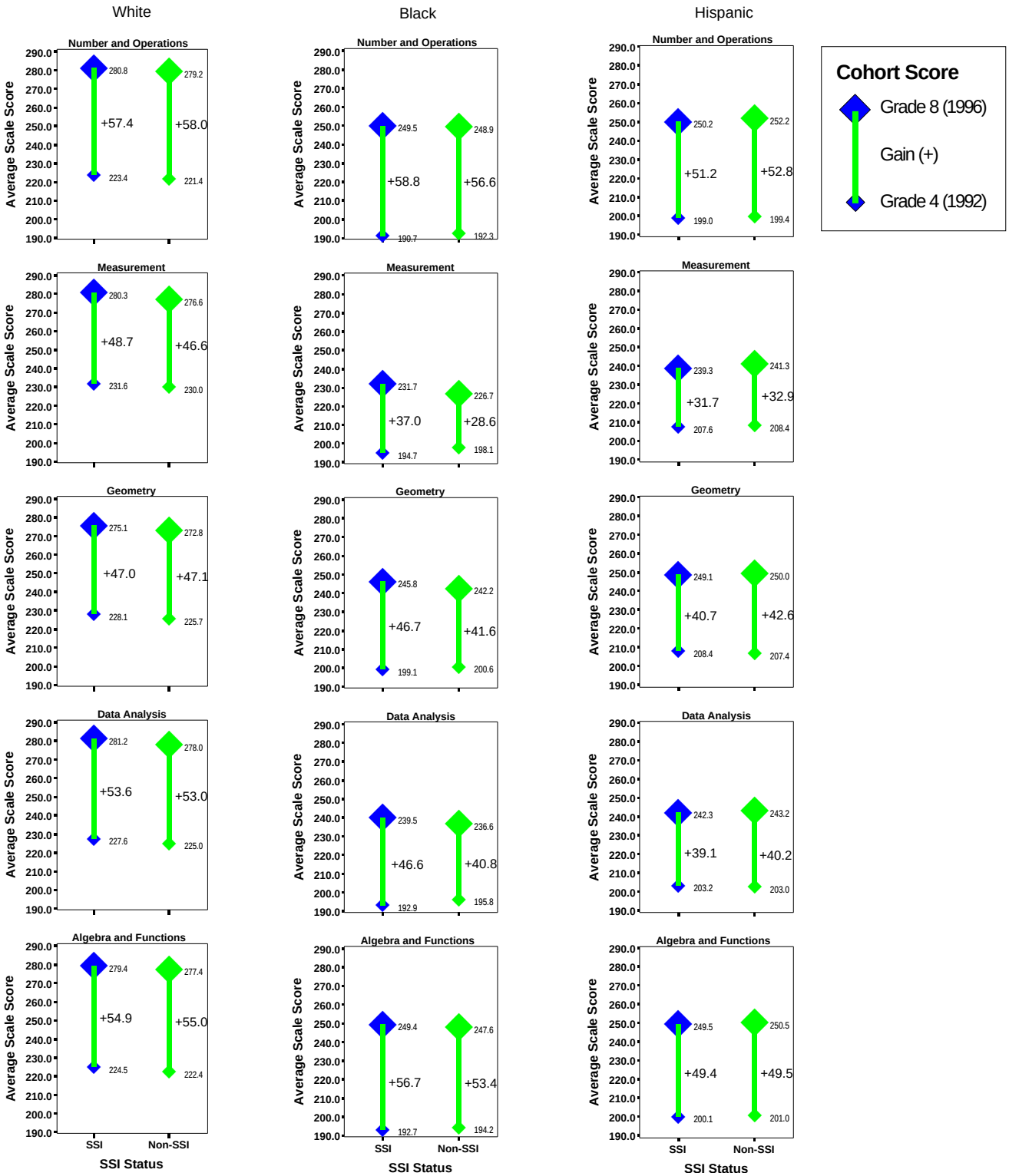


Ethnicity

The results for White, Black, and Hispanic students in cohort growth from 1992 to 1996 indicated differences among the three groups in five content strands (Figure 4.15a). It should be noted that only states that participated in the NAEP and that had a sufficient sample size of the subgroups were included in this analysis (12 SSI states and eight non-SSI states for Blacks and 11 SSI states and 12 non-SSI states for Hispanics). In general, White students outperformed Black and Hispanic students, and Black students gained more than Hispanic students. There was little variation in the cohort growth for white students from SSI states compared to white students from non-SSI in four of the five strands. White students from SSI states did gain more (2.1 scale points) than white students from non-SSI states in Measurement. As observed in previous figures, Black students from SSI states did have a higher growth (from 2.2 to 8.4 scale points) from 1992 to 1996 from grade 4 to grade 8 than did Black student from non-SSI states in all five content strands. The greatest difference was in Measurement (8.4 scale points in favor of SSI Black students). Although grade 4 Black students in SSI states started below their counterparts in non-SSI states in 1992, four years later their gaps were reversed in all of the five content strands. Hispanic students from non-SSI states gained slightly more (.1 to 1.9 scale points) than Hispanic students from SSI states in all five content areas.

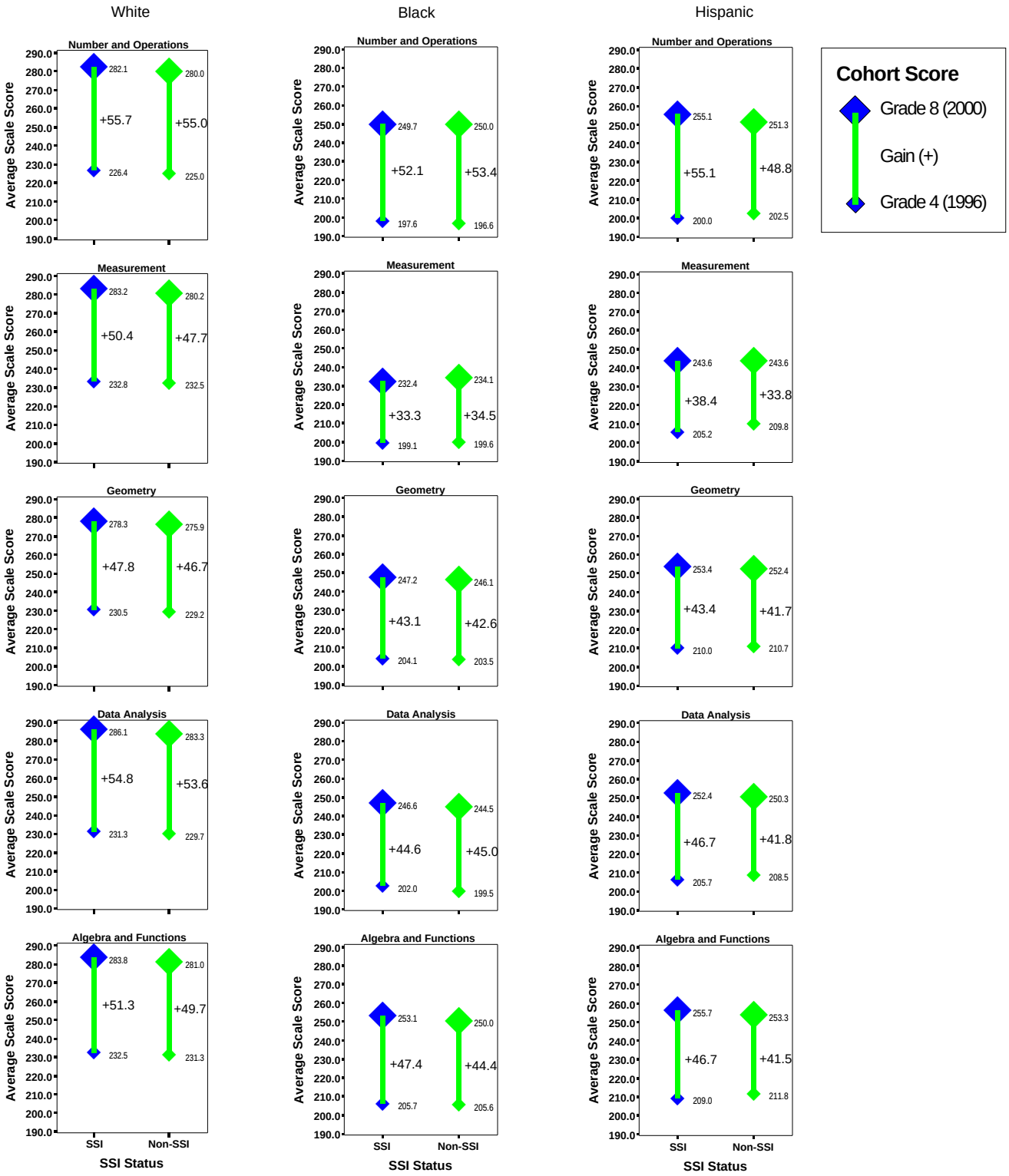
For the early years of the SSI program from 1992 to 1996, these data provide some evidence that Black students in SSI states made higher gains in all strands than did Black students in non-SSI states. We were particularly interested if this finding would be sustained over the next four years, from 1996 to 2000, for the next cohort of students. As shown in Figure 4.15b, Black students from SSI states from 1996 to 2000 did gain more (3 scale points) in Algebra/Functions than Black students from non-SSI states as they went from grade 4 to grade 8. However, in the other four content strands, the cohort gain was nearly the same between Black students in SSI states and Black students in non-SSI states or only slightly favoring those from the non-SSI states (1.3 scale points or less). Hispanic students in SSI states compared to those from non-SSI states made large gains of four or more scale points on four of the five content strands, all but Geometry, from grade 4 in 1996 to grade 8 in 2000. The greater gains Hispanic students in SSI states made from 1996 to 2000 are comparable to those experienced by Black students in SSI states the previous four years. The cohort of White students in SSI states made slightly greater gains (from .7 to 2.7 points) from 1996 to 2000 than did White students in non-SSI states. These data support a pattern of the SSI states having the greatest influence on Black students in the early years of the program, but having more relative effect on Hispanic students and White students in the latter years of the program.

Figure 4.15a. Cohort growth in average scale scores on content strands from 1992 to 1996, by race and SSI status: Trend Group 92-00 (14 SSI and 13 non-SSI states*).



* Due to the insufficient sample size of the subgroups, results are based on 12 SSI states and 8 non-SSI states for Blacks, and 11 SSI states and 12 non-SSI states for Hispanics.

Figure 4.15b. Cohort growth in average scale scores on content strands from 1996 to 2000, by race and SSI status: Trend Group 92-00 (14 SSI and 13 non-SSI states*).



* Due to the insufficient sample size of the subgroups, results are based on 12 SSI states and 8 non-SSI states for Blacks, and 11 SSI states and 12 non-SSI states for Hispanics.

Gaps in Cohort Growth Between Different Groups

The set of graphs in this section depict the difference in gain over a four period between two subgroups by cohort group on the mathematics composite score and on the five mathematics strands. The gain is the change in scale scores in grade 4 to the scale scores in grade 8. First, male students are compared with female students. All positive values indicate that male students made a larger gain than female students, or the gap favored male students. The further the values are from zero, the more was the gap. The analysis by gender is followed by a comparison of the difference first between White students and Blacks students and then between White students and Hispanic students.

Gender

From 1992 (grade 4) to 1996 (grade 8), there was very little differences between the SSI states and the non-SSI states in the achievement gap between female students and male students (Figure 4.16a). The gap between female students and male students on the six measures generally was less than one scale point. Only in geometry, did the gap between female students and male students exceed 1.5 scale points, favoring male students.

In the next cohort, 1996 (grade 4) to 1996 (grade 8), the male students scored higher than female students on three of the mathematics content strands—Number/Operations, Measurement, and Geometry (Figure 4.16b). Female students scored higher than male students on the Algebra/Functions strand. Both groups scored about the same on the mathematics composite score and on Data Analysis. There was very little variation between the SSI states and the non-SSI states in the gap between male and female students over this period of time. This suggests that the reform efforts did not differentiate by gender.

Figure 4.16a. Differences in cohort growth from 1992 (grade 4) to 1996 (grade 8) between male and female students by SSI status: Trend Group 92-00 (14 SSI and 13 non-SSI states).

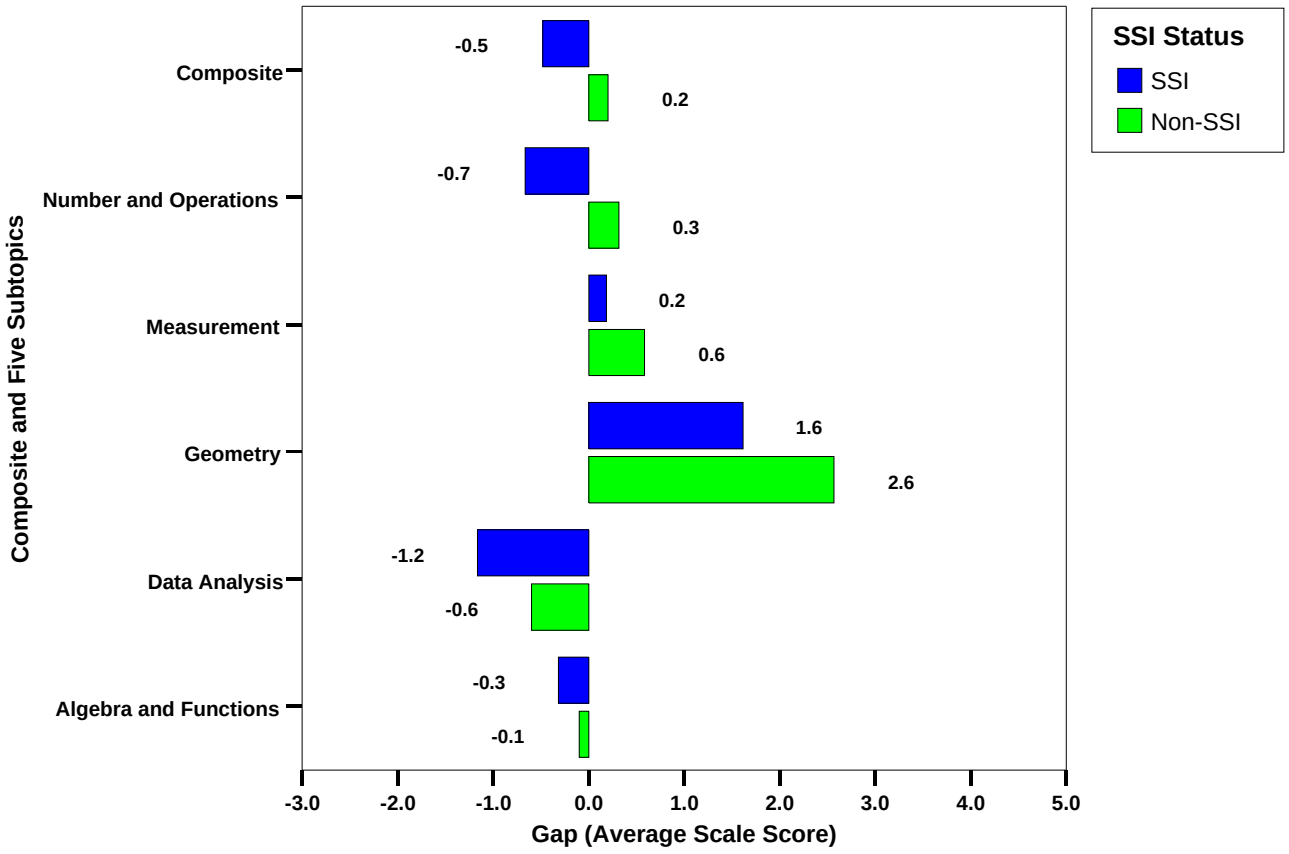
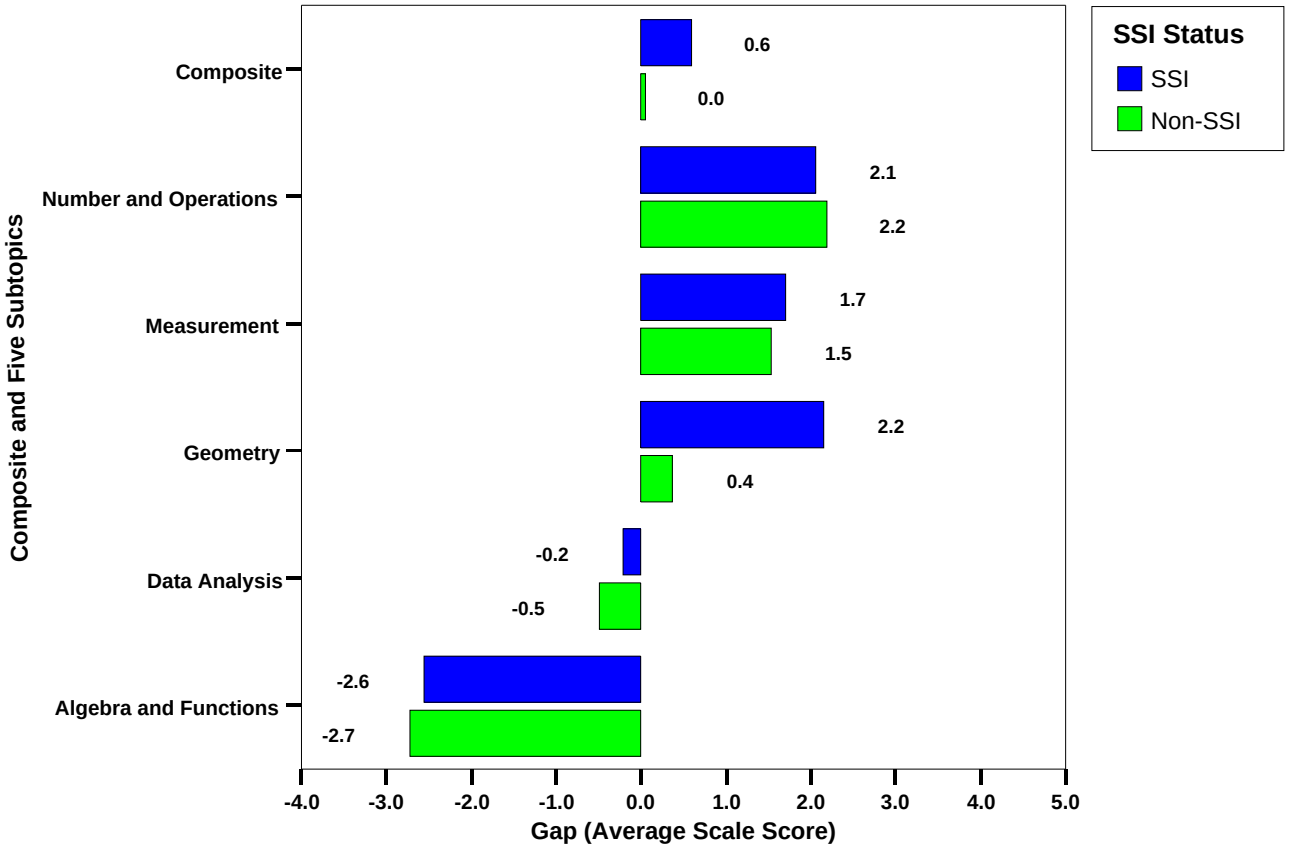


Figure 4.16b. Differences in average scale scores between male and female students from 1996 to 2000, by SSI status: Trend Group 92-00 (14 SSI and 13 non-SSI states).



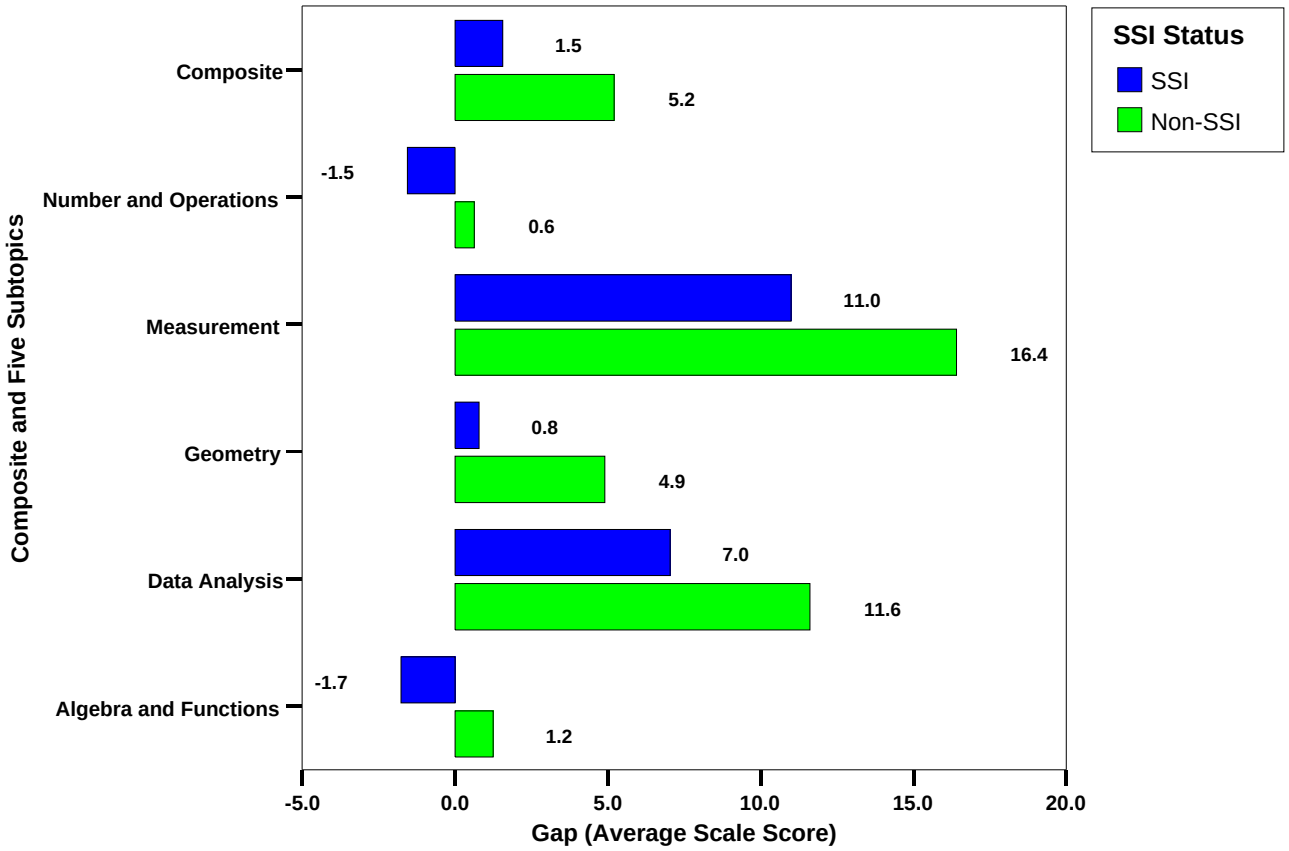
Ethnicity

The results for cohort growth differences between White and Black students were quite interesting. SSI states were successful in reducing the gap of cohort growth between White and Black students (Figure 4.17a). The cohort growth was computed as the difference in scores in grade 8 in 1996 and in grade 4 in 1992. In the composite and in the five content-strand scale scores, cohort growth gaps in SSI states were smaller than those in non-SSI states, or were even reversed. The biggest score difference in cohort growth between SSI and non-SSI states was noted in Measurement. Non-SSI states had a 16.4-point difference, while SSI states had an 11-point difference.

The most interesting pictures in cohort growth differences were displayed in Number and Operations and Algebra and Functions. In SSI states in 1996, Black students in the cohort of students who were in grade 4 in 1992 gained more in Number and Operations and in Algebra and Functions over four years than White students. As a result, the gap between White and Black students was reversed, with Black students gaining more than White students—1.5 points in

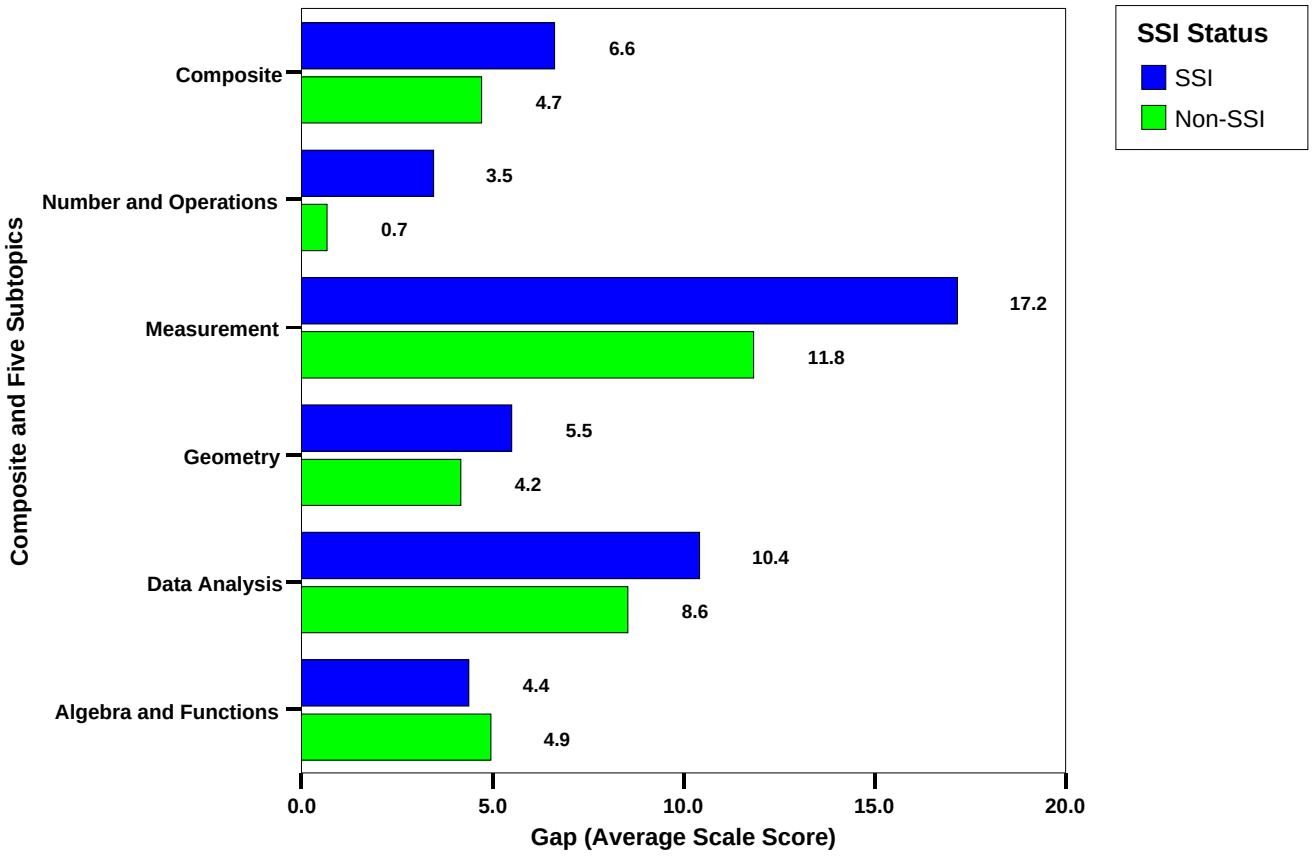
Number and Operations and 1.7 points in Algebra and Functions. In non-SSI states, on the other hand, White students gained slightly more than Black students, 0.6 points in Number and Operations and 1.2 points in Algebra and Functions. The fact that Black students gained more than White students is noteworthy, considering that all of the other comparisons show White students performing better than Black students.

Figure 4.17a. Differences in average scale scores between White and Black students from 1992 to 1996, by SSI status: Trend Group 92-00 (14 SSI and 13 non-SSI states*).



* Due to the insufficient sample size of these subgroups, results are based on 12 SSI states and 8 non-SSI states.

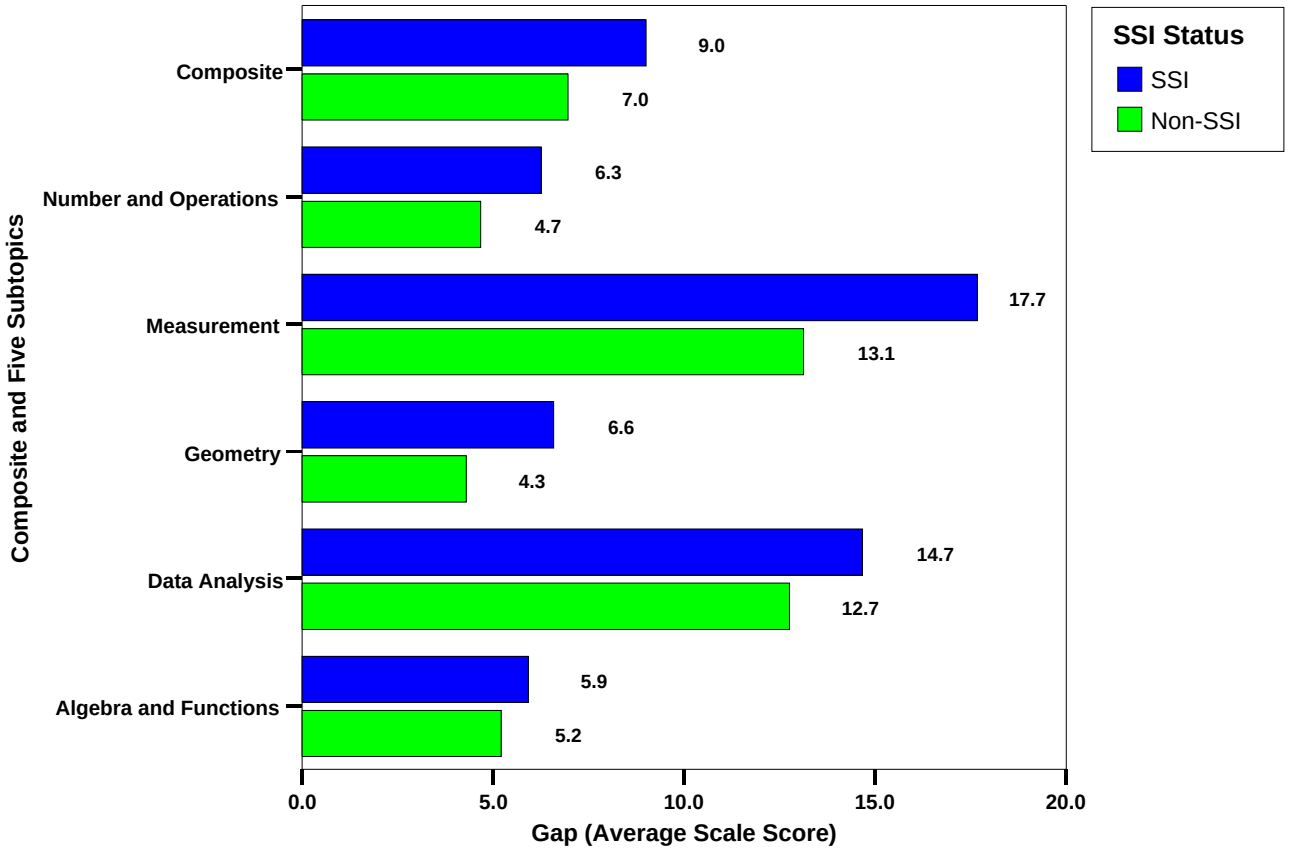
Figure 4.17b. Differences in average scale scores between White and Black students from 1996 to 2000, by SSI status: Trend Group 92-00 (14 SSI and 13 non-SSI states*).



* Due to the insufficient sample size of these subgroups, results are based on 12 SSI states and 8 non-SSI states.

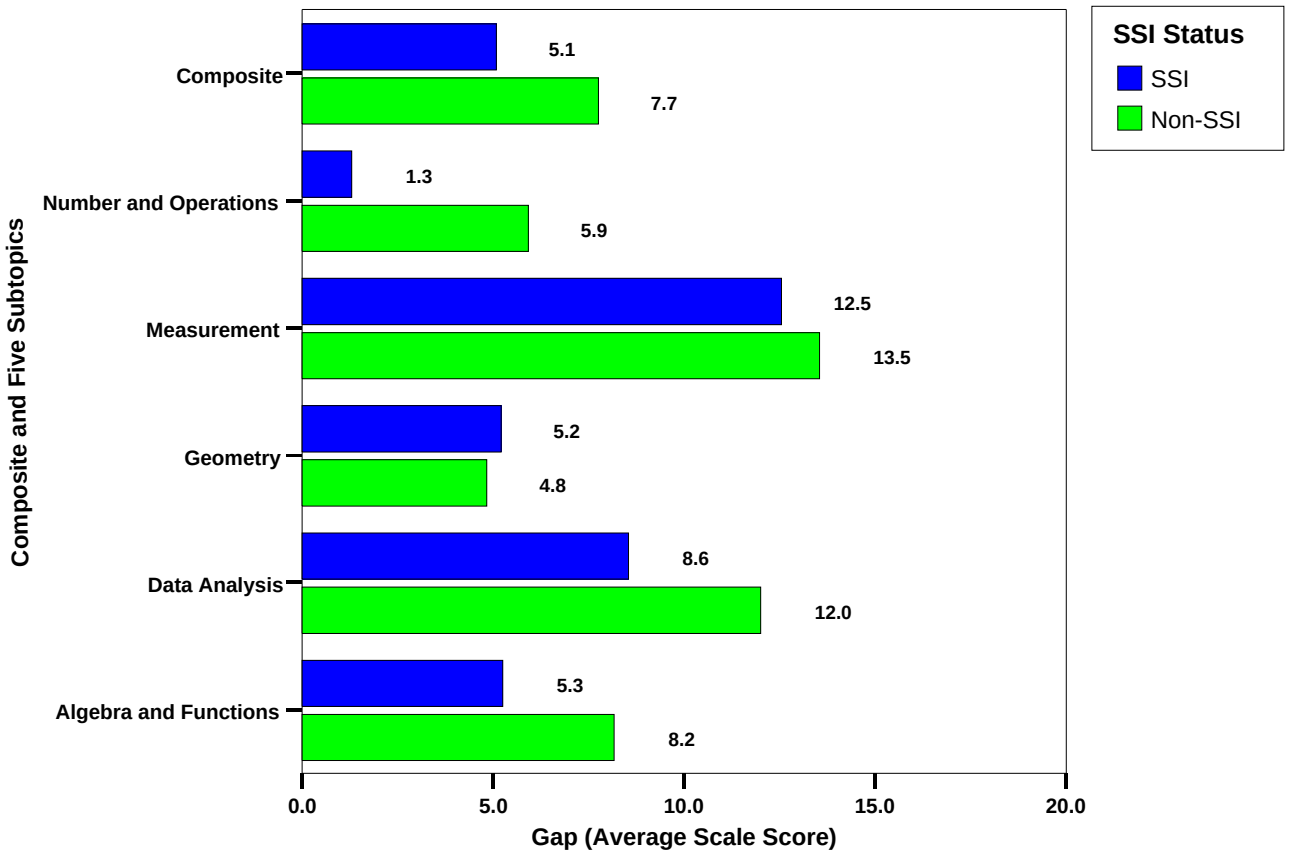
Unlike the trend of difference between White and Black students, the gaps between White and Hispanic cohort students were greater in SSI states than in non-SSI states (Figure 4.18a). In all six scale scores, the cohort growth differences were greater in SSI states: composite (9.0 for SSI, 7.0 for non-SSI); Number and Operations (6.3 for SSI, 4.7 for non-SSI); Measurement (17.7 for SSI, 13.1 for non-SSI); Geometry (6.6 for SSI, 4.3 for non-SSI); Data Analysis (14.7 for SSI, 12.7 for non-SSI); and Algebra and Functions (5.9 for SSI, 5.2 for non-SSI). In Algebra and Functions, the cohort gaps between White and Hispanic students were smaller than in other content strands.

Figure 4.18a. Differences in average scale scores between White and Hispanic students from 1992 to 1996, by SSI status: Trend Group 92-00 (14 SSI and 13 non-SSI states*).



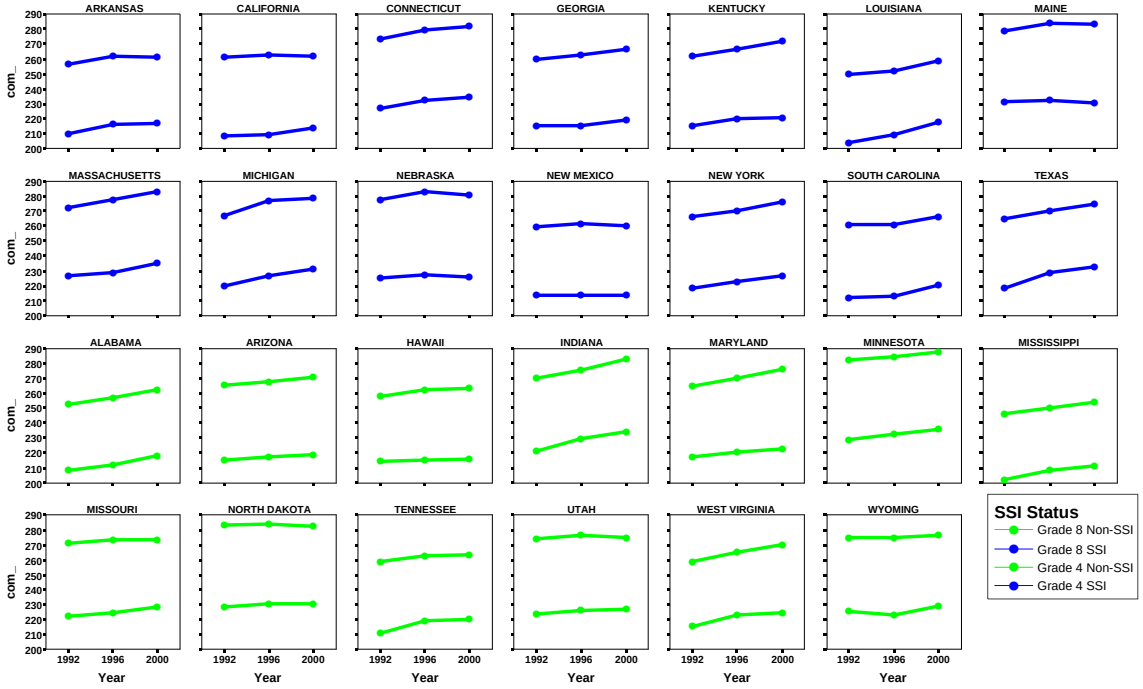
* Due to the insufficient sample size of these subgroups, results are based on 11 SSI states and 12 non-SSI states.

Figure 4.18b. Differences in average scale scores between White and Hispanic students from 1996 to 2000, by SSI status: Trend Group 92-00 (14 SSI and 13 non-SSI states*).



* Due to the insufficient sample size of these subgroups, results are based on 11 SSI states and 12 non-SSI states.

Figure 4.19. Individual state trends in average scale scores, Trend Group 92-00 (14 SSI and 13 non-SSI states).



Summary and Conclusions

This chapter presents the results from the State NAEP mathematics assessments for grades 4 and 8 in 1992, 1996, and 2000, and for two cohorts of students from SSI states and non-SSI states. We focused our analyses on the 27 states—14 SSI and 13 non-SSI—that participated in three State NAEP administrations in mathematics (See Figure 4.19). The differences between SSI and non-SSI states in the composite score and in each of the five content strands (Number and Operations; Measurement; Geometry and Spatial Sense; Data Analysis, Statistics, and Probability; and Algebra and Functions) were based on descriptive trend analyses that compared the group means of SSI states and non-SSI states across each assessment year. In general, the results revealed that substantial student gains in the mathematics composite score and in the five content strands over time were observed for grade 4, grade 8, and the growth in performance by the two cohorts in both SSI and non-SSI states. Considerable improvements were also noted for students by gender and race/ethnicity. There were some differences in performance by subgroups of students that distinguished SSI and non-SSI states over one of four-year periods. However, these differences were characteristic of one four year period and were not sustained over the eight years including in the analysis.

Summaries of performance trends for different subgroups and gaps between males and females, as well as between Whites and Blacks and between Whites and Hispanics, follow:

Trends in Average Scale Scores

- Both the 14 SSI and the 13 non-SSI states experienced an increase in the average composite scale scores of about 6 scale points from 1992 to 2000 at grades 4 and 8. In 1992, grade 8 students from SSI states scored 1.2 scale points lower than grade 8 students from non-SSI states. Otherwise, students in both grades from the two groups of states had nearly identical average scores for each of the testing times.
- Male students from SSI states had average scale scores comparable to male students from non-SSI states, at both grades 4 and 8, on the composite scale score and on each content strand. Female students from SSI states had average scale scores slightly less than female students from non-SSI states. Female students in SSI states gained at a slightly higher rate over this period than those in non-SSI states on most scale scores. The slightly lower scores by female students from SSI states generally were in the content strands of Number/Operations and Measurement.
- Slight performance differences by ethnicity between SSI and non-SSI states existed on all six scale scores at both grades 4 and 8. Regardless of SSI status, White students outperformed Black and Hispanic students on all scale scores at both grade levels. Hispanic students scored higher than Black students at both grade levels.
- White students from SSI states performed higher than White students from non-SSI states in all three years—1992, 1996, and 2000. Black students from SSI states gained more than those from non-SSI states between 1992 and 1996. Hispanic students from SSI states gained more than those from non-SSI states between 1996 and 1992.
- The difference in performance between White students and Black students in grades 4 and 8 decreased some for the SSI states, but remained rather stable or increased for students in non-SSI states over the three testing times. The Measurement content strand was the one exception at grade 8.

- The difference in performance between White students and Hispanic students was smaller for non-SSI states than for SSI states for both grades and for the three testing times. At grade 8, with the exception of Measurement, the gap between White students and Hispanic students in SSI states and non-SSI states decreased over the eight years. However, at grade 4, the gap between White students and Hispanic students in both groups of states varied by content strand. For both groups of states, the gap on the composite scale score and on Data Analysis remained nearly the same. For the non-SSI states, the gap increased on Number/Operations, Measurement, and Geometry. The gap declined on Algebra/Functions. For SSI states, the gap increased on Measurement and Geometry, stayed the same on Number/Operations, and declined on Algebra/Functions.
- Considering two cohorts of students—one in grade 4 in 1992 and grade 8 in 1996 and the other in grade 4 in 1996 and grade 8 in 2000—students in both SSI states and non-SSI states gained nearly the same over the four years, 51.7 points and 51.6 points, respectively, for the first cohort and 50.7 and 50.5, respectively, for the second cohort.
- Male students and female students in each cohort gained very nearly the same (within .7 points of each other) in SSI states and non-SSI states.
- White students in the 1992 grade 4 cohort gained about the same in both the SSI states and the non-SSI states over four years, but White students from the SSI states gained more (1.5 scale points) than White students in non-SSI states in the second cohort from 1996 to 2000.
- Black and Hispanic students in both SSI and non-SSI states gained less than White students over the four years between grade 4 and grade 8. This indicates that Black and Hispanic students continued to lose ground over these four years. However, Black students in SSI states in the first cohort (1992 to 1996) gained more from grade 4 to grade 8 (4.4 scale points) than Black students from non-SSI states. Hispanic students in SSI states in the second cohort (1996 to 2000) gained more from grade 4 to grade 8 (4.8 scale points) than Hispanic students from non-SSI states.
- The White-Black gap in the gain scores between grade 4 and grade 8 for the 1992-1996 cohort was less in SSI states than non-SSI states on all six scales. On the Number/Operations and Algebra/Functions strands, Black students in SSI states actually gained more between grade 4 and grade 8 than did White students. This result was not repeated by the next cohort of students, where the White-Black gap in the gain scores was higher for students from the SSI states than for students from the non-SSI states except for Algebra/Functions.
- The White-Hispanic gap in the gain scores between grade 4 and grade 8 for the 1992-1996 cohort was greater in SSI states than in non-SSI states on the composite score and all five content strands. However, for the next cohort, 1996 to 2000, the White-Hispanic gap in the gain scores between grade 4 and grade 8 was less in SSI states than in non-SSI states on all of the scale scores except for Geometry.

Even though the descriptive trends of average scale scores suggest that there was evidence in most cases for the differences between SSI and non-SSI states on the overall composite scale and on each of the five content strands, it is unclear whether the differences can be attributable to the relative effectiveness of SSI in those states. There are many factors involved in how students learn over the years. School structures, home environments, state educational policies, and others can affect learning. In the following chapters, we will identify some policy-relevant variables related to SSI states and their relationships with student outcomes.

Appendices

Appendix A. Numeric Tables for All Available Samples

Appendix B. Findings for Grade 4 (1992) to Grade 8 (1996) Cohort for 20 SSI States and 15 Non-SSI States (All States with NAEP Data for 1992 and 1996)

Appendix C. Narrative Summation of Data for Trend Group 90-96

Appendix A. Numeric Tables for All Available Samples

Composite Scores

Total Group

Figure 4A.1. Average scale scores, by SSI status.

	1990	1992	1996
Grade 8			
SSI	260.6	265.4	270.2
Non-SSI	263.8	266.9	272.2
Grade 4			
SSI		217.9	221.7
Non-SSI		218.3	223.0

Gender

Figure 4A.2. Average scale scores, by gender and SSI status.

	1990		1992		1996	
	Male	Female	Male	Female	Male	Female
Grade 8						
SSI	261.9	259.3	266.4	264.3	271.3	269.2
Non-SSI	265.1	262.5	267.7	266.0	272.5	272.0
Grade 4						
SSI			218.8	216.9	222.6	220.7
Non-SSI			218.7	217.9	223.7	222.2

Ethnicity

Figure 4A.3. Average scale scores, by race and SSI status.

	1990			1992			1996		
	White	Black	Hispanic	White	Black	Hispanic	White	Black	Hispanic
Grade 8									
SSI	270.0	235.6	237.9	274.8	239.2	242.4	279.4	244.7	248.7
Non-SSI	268.6	236.9	240.5	271.7	239.0	244.0	278.9	243.7	251.9
Grade 4				226.4	194.6	203.3	229.9	200.1	206.1
SSI				224.3	194.9	204.4	229.2	200.3	208.7
Non-SSI									

Subtopic Scores

Total Group

Figure 4A.4. Average scale scores in content strands, by SSI status.

		1990	1992	1996
Grade 8				
Number and Operations	**SSI**	264.8	269.7	272.1
	Non-SSI	268.3	271.0	274.7
Measurement	**SSI**	256.8	263.2	267.8
	Non-SSI	261.1	266.0	271.0
Geometry	**SSI**	258.0	260.8	267.9
	Non-SSI	261.6	262.5	269.2
Data Analysis	**SSI**	260.7	265.4	270.2
	Non-SSI	263.1	267.0	272.0
Algebra and Functions	**SSI**	259.8	264.9	271.6
	Non-SSI	262.0	265.7	273.0
Grade 4				
Number and Operations	**SSI**		214.9	218.2
	Non-SSI		215.6	219.3
Measurement	**SSI**		222.5	223.9
	Non-SSI		223.5	225.9
Geometry	**SSI**		220.8	223.6
	Non-SSI		220.7	224.7
Data Analysis	**SSI**		218.9	223.0
	Non-SSI		218.4	223.9
Algebra and Functions	**SSI**		216.4	225.1
	Non-SSI		216.6	226.5

Gender

Figure 4A.5. Average scale scores in content strands, by gender and SSI status.

		1990		1992		1996	
		Male	Female	Male	Female	Male	Female
Grade 8							
Number and Operations	**SSI**	265.6	264.1	270.1	269.4	273.1	271.2
	Non-SSI	269.4	267.1	271.3	270.6	274.9	274.5
Measurement	**SSI**	260.9	252.6	266.7	259.9	270.1	265.5
	Non-SSI	264.9	257.3	269.1	262.9	272.7	269.4
Geometry	**SSI**	259.6	256.3	262.4	259.2	269.0	266.9
	Non-SSI	262.8	260.4	263.7	261.2	269.3	269.1
Data Analysis	**SSI**	262.2	259.2	266.6	264.3	270.7	269.7
	Non-SSI	264.9	261.4	267.8	266.1	271.4	272.6
Algebra and Functions	**SSI**	259.2	260.4	264.7	265.1	272.3	270.9
	Non-SSI	261.6	262.4	265.2	266.2	273.0	273.1
Grade 4							
Number and Operations	**SSI**			215.8	214.0	218.9	217.5
	Non-SSI			215.9	215.2	219.8	218.8
Measurement	**SSI**			224.8	220.2	225.9	221.9
	Non-SSI			225.3	221.6	227.9	223.9
Geometry	**SSI**			220.7	220.9	223.2	224.0
	Non-SSI			220.0	221.4	224.1	225.4
Data Analysis	**SSI**			219.1	218.7	224.0	222.0
	Non-SSI			218.5	218.4	224.3	223.5
Algebra and Functions	**SSI**			217.1	215.7	226.7	223.4
	Non-SSI			216.4	216.7	227.8	225.0

Ethnicity

Figure 4A.6. Average scale scores in content strands, by race and SSI status.

		1990			1992			1996		
		White	Black	Hispanic	White	Black	Hispanic	White	Black	Hispanic
Grade 8										
Number and Operations	**SSI**	273.3	243.1	243.4	278.3	246.2	247.6	280.4	249.2	251.8
	Non-SSI	272.4	243.9	246.4	275.4	246.7	248.4	280.7	250.2	255.0
Measurement	**SSI**	267.5	226.5	232.9	274.8	229.1	238.6	280.3	231.5	240.6
	Non-SSI	266.2	229.9	235.6	272.1	230.0	239.3	279.8	229.9	246.9
Geometry	**SSI**	266.6	233.3	238.3	269.1	236.4	242.0	275.6	245.8	250.4
	Non-SSI	266.3	235.2	239.6	266.8	235.3	243.4	274.7	244.1	253.4
Data Analysis	**SSI**	272.3	231.5	232.6	276.6	235.9	237.2	281.4	239.9	245.2
	Non-SSI	268.9	233.3	237.7	272.8	236.2	241.6	280.5	237.7	247.3
Algebra and Functions	**SSI**	268.7	236.8	236.8	273.7	241.5	241.9	279.7	250.1	251.3
	Non-SSI	266.7	236.2	238.4	270.0	240.2	243.1	279.0	248.9	253.4
Grade 4										
Number and Operations	**SSI**				223.6	192.2	199.2	226.6	197.0	201.3
	Non-SSI				221.7	192.2	200.9	225.7	196.9	204.4
Measurement	**SSI**				231.7	196.3	208.7	233.1	198.7	207.1
	Non-SSI				229.6	197.1	209.9	232.9	199.2	211.0
Geometry	**SSI**				227.9	200.3	209.0	230.6	204.1	211.3
	Non-SSI				225.7	200.5	208.8	230.1	204.8	211.9
Data Analysis	**SSI**				227.8	194.4	204.5	231.3	201.3	208.0
	Non-SSI				225.0	194.9	204.2	230.4	201.0	208.9
Algebra and Functions	**SSI**				224.8	193.8	200.9	232.6	205.5	210.8
	Non-SSI				222.2	194.0	202.5	232.1	205.9	213.8

Gaps Between Different Groups

Gender

Figure 4A.7. Gender differences in average scale scores, by SSI status.

		1990	1992	1996
Grade 8				
Composite	**SSI**	2.6	2.1	2.1
	Non-SSI	2.6	1.7	0.5
Number and Operations	**SSI**	1.5	0.7	2.0
	Non-SSI	2.2	0.7	0.4
Measurement	**SSI**	8.3	6.8	4.6
	Non-SSI	7.5	6.2	3.4
Geometry	**SSI**	3.3	3.1	2.1
	Non-SSI	2.3	2.5	0.3
Data Analysis	**SSI**	3.0	2.4	1.0
	Non-SSI	3.5	1.6	-1.2
Algebra and Functions	**SSI**	-1.2	-0.4	1.4
	Non-SSI	-0.8	-0.9	-0.1
Grade 4				
Composite	**SSI**		1.9	1.9
	Non-SSI		0.8	1.5
Number and Operations	**SSI**		1.8	1.3
	Non-SSI		0.7	1.1
Measurement	**SSI**		4.6	4.1
	Non-SSI		3.7	4.0
Geometry	**SSI**		-0.3	-0.9
	Non-SSI		-1.4	-1.3
Data Analysis	**SSI**		0.4	2.0
	Non-SSI		0.0	0.7
Algebra and Functions	**SSI**		1.5	3.3
	Non-SSI		-0.3	2.8

Ethnicity

Figure 4A.8. Differences in average scale scores between racial subgroups, by SSI status.

		Black Gap (White – Black)			Hispanic Gap (White – Hispanic)		
		1990	1992	1996	1990	1992	1996
Grade 8							
Composite	**SSI**	34.3	35.4	33.9	32.1	32.1	31.4
	Non-SSI	31.8	31.6	34.5	29.5	28.6	27.0
Number and Operations	**SSI**	30.2	32.0	30.7	29.8	30.5	29.3
	Non-SSI	28.8	27.9	29.7	27.3	27.7	25.7
Measurement	**SSI**	40.7	45.3	47.2	34.6	35.8	40.4
	Non-SSI	36.8	40.2	49.1	32.6	33.8	32.9
Geometry	**SSI**	33.1	32.5	29.0	28.3	26.8	25.7
	Non-SSI	30.7	30.1	29.6	27.7	24.1	21.4
Data Analysis	**SSI**	40.8	40.5	41.0	39.7	39.0	37.1
	Non-SSI	35.4	35.9	42.7	32.7	32.3	33.1
Algebra and Functions	**SSI**	31.7	32.2	28.9	31.9	31.6	29.3
	Non-SSI	30.8	28.9	29.6	29.8	27.8	25.7
Grade 4							
Composite	**SSI**		31.5	29.7		23.1	23.9
	Non-SSI		29.3	29.0		19.7	20.5
Number and Operations	**SSI**		31.2	29.7		24.4	25.3
	Non-SSI		29.6	29.1		20.7	21.3
Measurement	**SSI**		35.0	34.2		23.0	26.1
	Non-SSI		32.7	33.5		19.5	21.9
Geometry	**SSI**		27.2	26.1		18.9	19.4
	Non-SSI		24.3	25.1		16.7	18.2
Data Analysis	**SSI**		33.1	30.0		23.3	23.3
	Non-SSI		29.6	29.6		20.0	21.5
Algebra and Functions	**SSI**		30.7	27.1		23.9	21.8
	Non-SSI		28.4	26.1		19.9	18.3

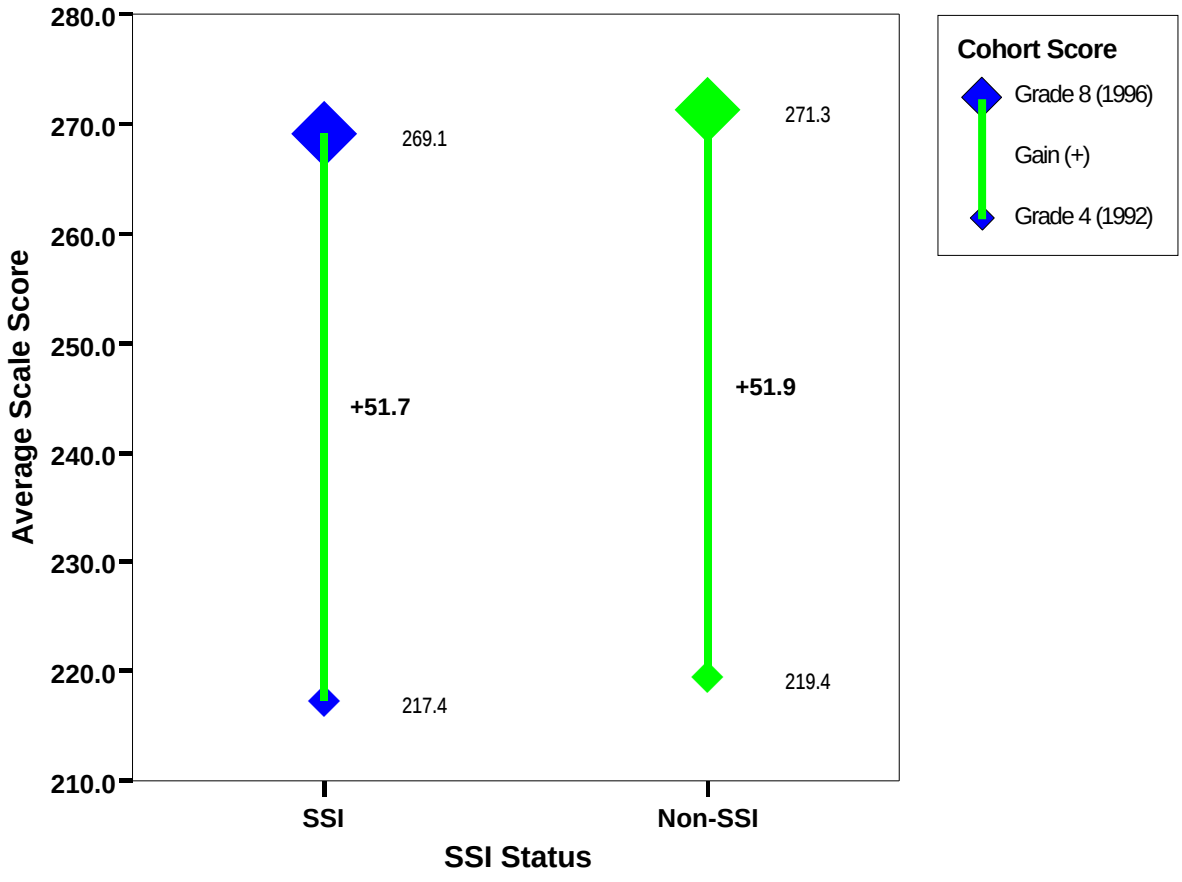
Appendix B. Findings for Grade 4 (1992) to Grade 8 (1996) Cohort for 20 SSI States and 15 Non-SSI States (All States with NAEP Data for 1992 and 1996)

Because states did not participate in every year of NAEP, the group of states included in an analysis will vary by the number of years included in the trend data. The Trend Group 92-00 included all of the 14 SSI states and the 13 non-SSI states that participated in the State NAEP for 1992, 1996, and 2000. Appendix B reports findings from the Trend Group 92-96 that includes all of the 20 SSI states and the 15 non-SSI states that participated in the State NAEP for 1992 and 1996. The Trend Group 92-00 is a subset of the Trend Group 92-96. Data in Appendix B are reported to provide some contrast in what are the findings if a larger number of states were included in the analysis for 1992 and 1996.

Composite Scores

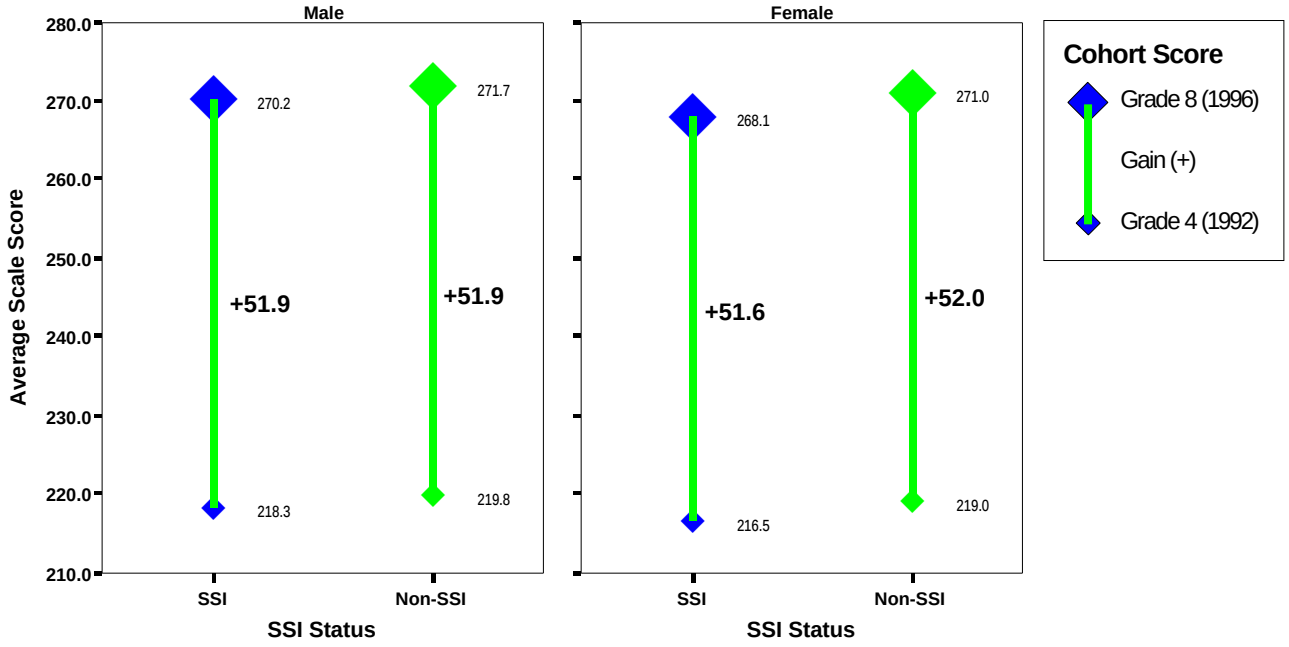
Total Group

Figure 4B.1. Cohort growth in average scale scores, by SSI status: Cohort group 92-96 (20 SSI and 15 non-SSI).



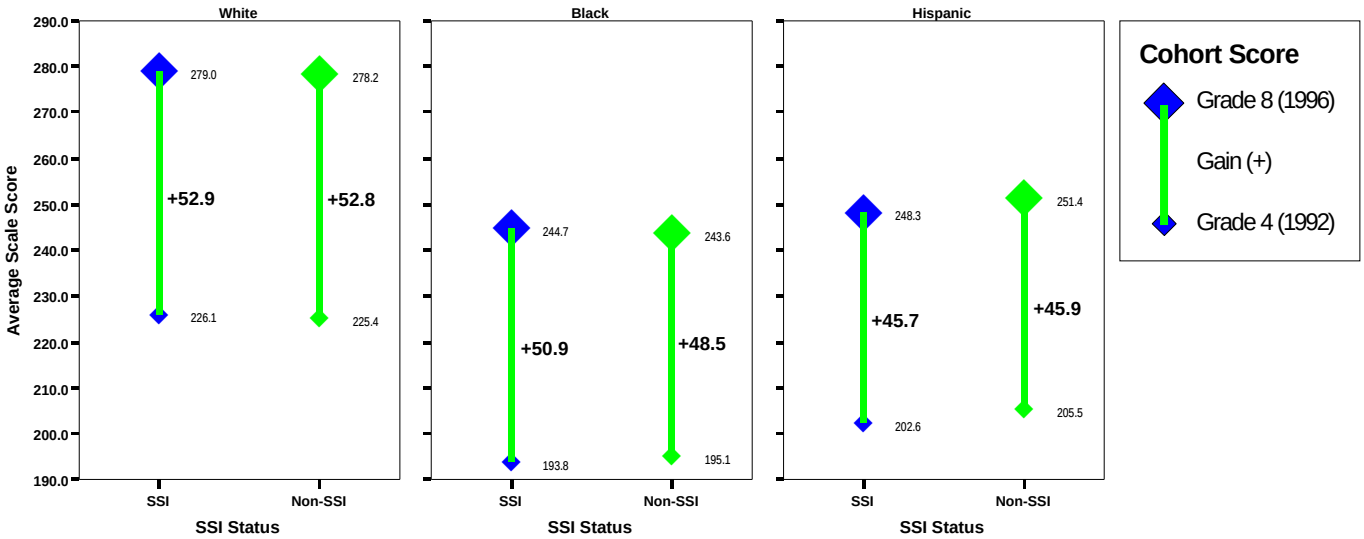
Gender

Figure 4B.2. Cohort growth in average scale scores, by gender and SSI status: Cohort group 92-96 (20 SSI and 15 non-SSI).



Ethnicity

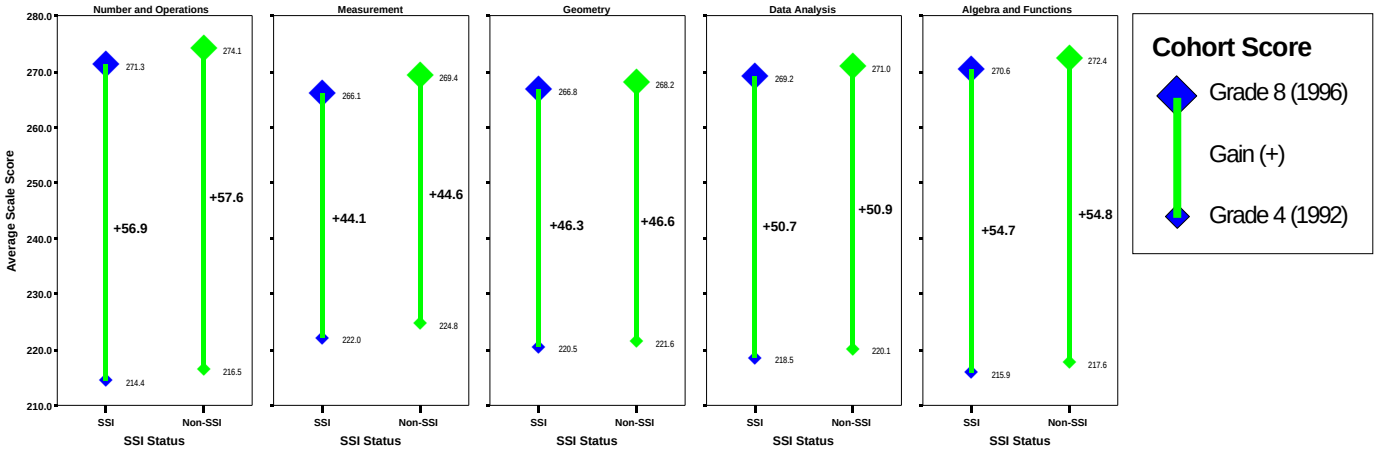
Figure 4B.3. Cohort growth in average scale scores, by race and SSI status: Cohort group 92-96 (20 SSI and 15 non-SSI).



Subtopic Scores

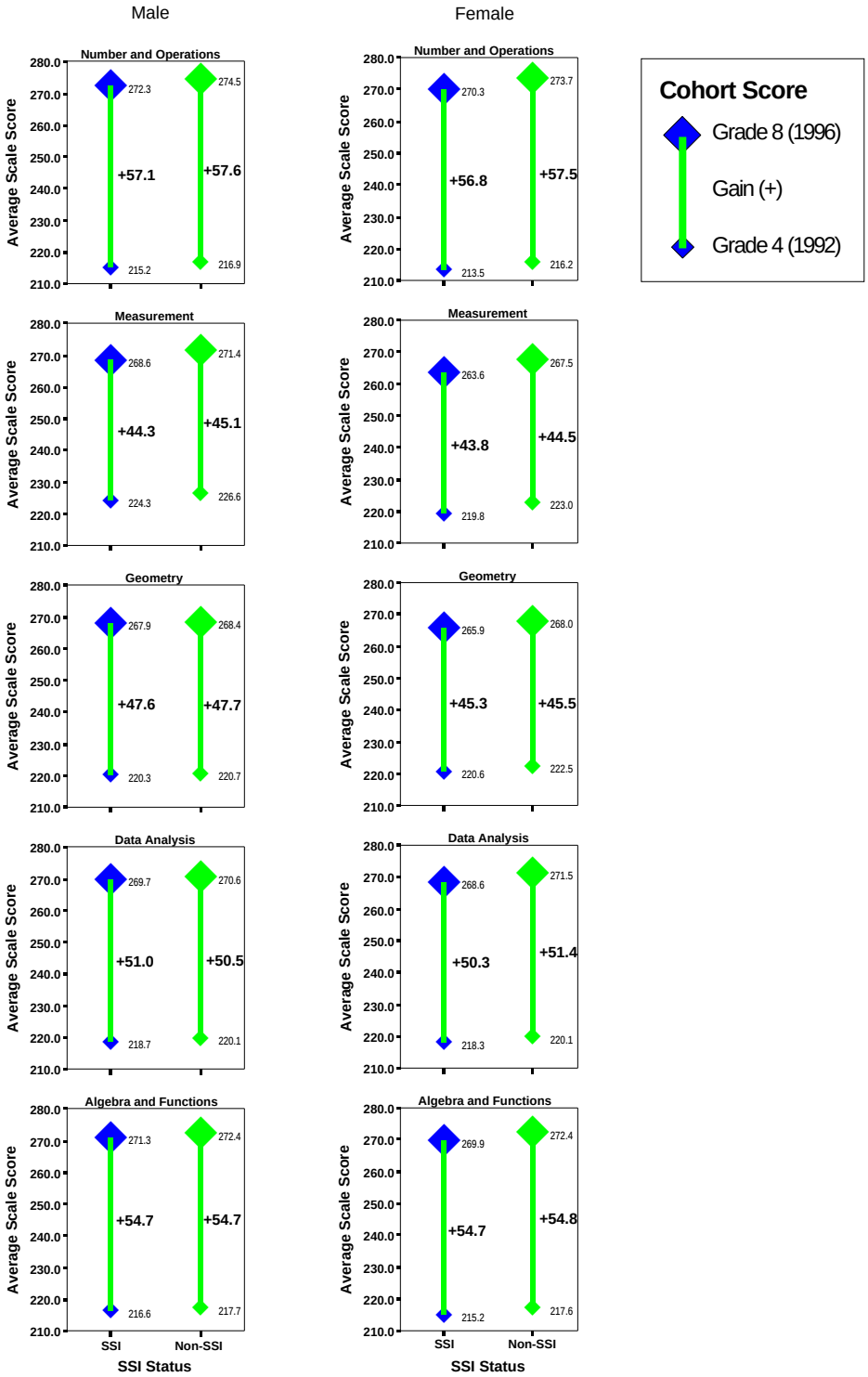
Total Group

Figure 4B.4. Cohort growth in average scale scores on content strands, by SSI status: Cohort group 92-96 (20 SSI and 15 non-SSI).



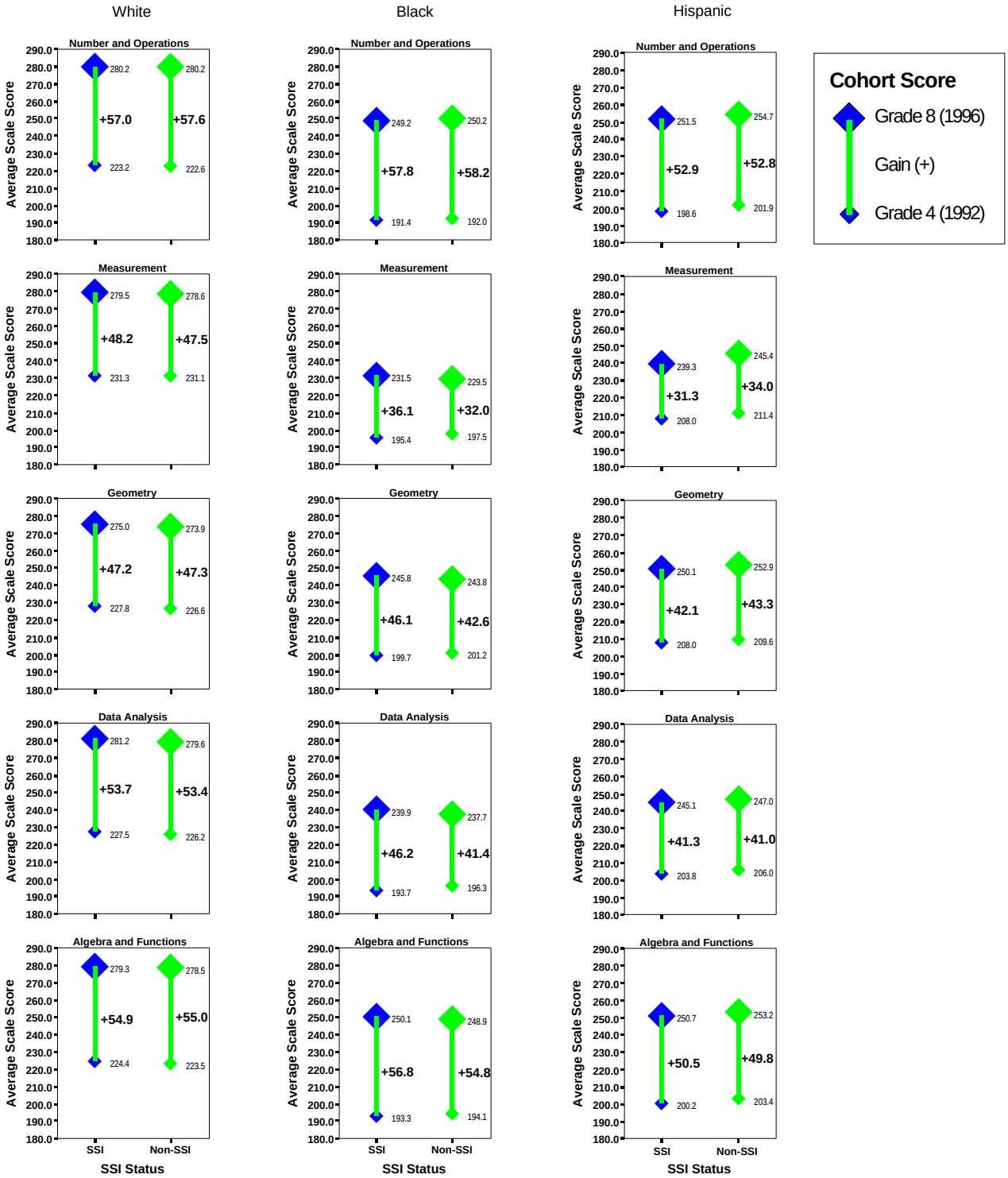
Gender

Figure 4B.5. Cohort growth in average scale scores on content strands, by gender and SSI status: Cohort group 92-96 (20 SSI and 15 non-SSI).



Ethnicity

Figure 4B.6. Cohort growth in average scale scores on content strands, by race and SSI status: Cohort group 92-96 (20 SSI and 15 non-SSI).



Gender

Figure 4B.7. Gender differences in average scale scores, by SSI status: Cohort group 92-96 (20 SSI and 15 non-SSI).

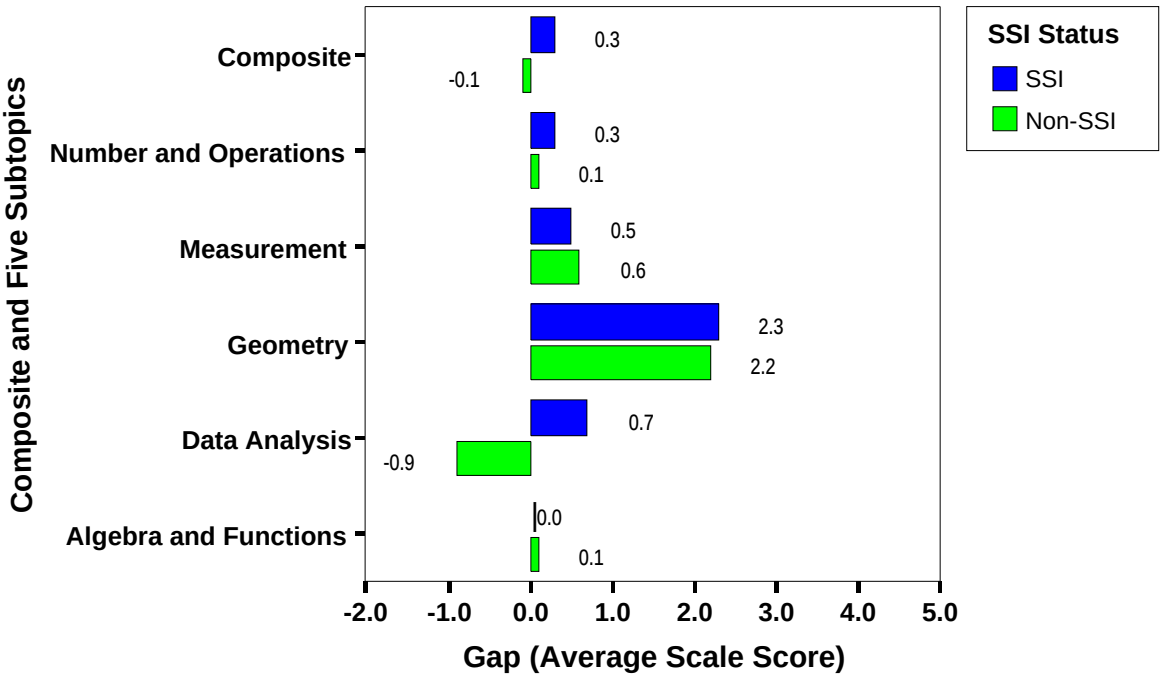


Figure 4B.8. Differences in average scale scores between White and Black students, by SSI status: Cohort group 92-96 (20 SSI and 15 non-SSI).

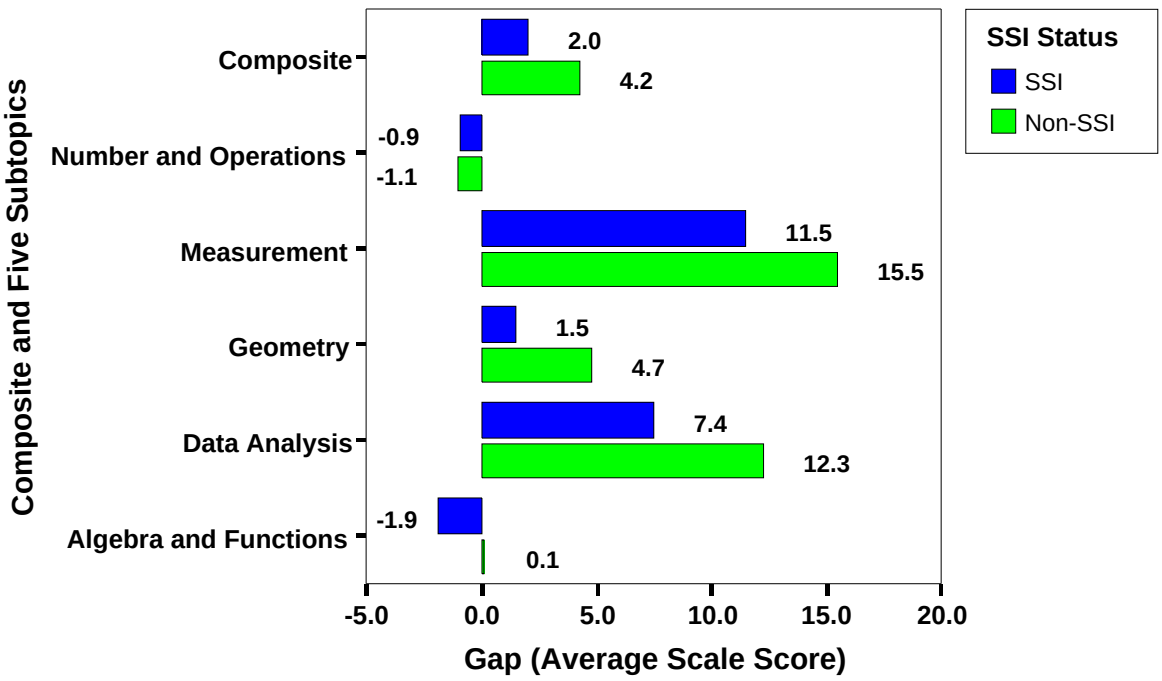
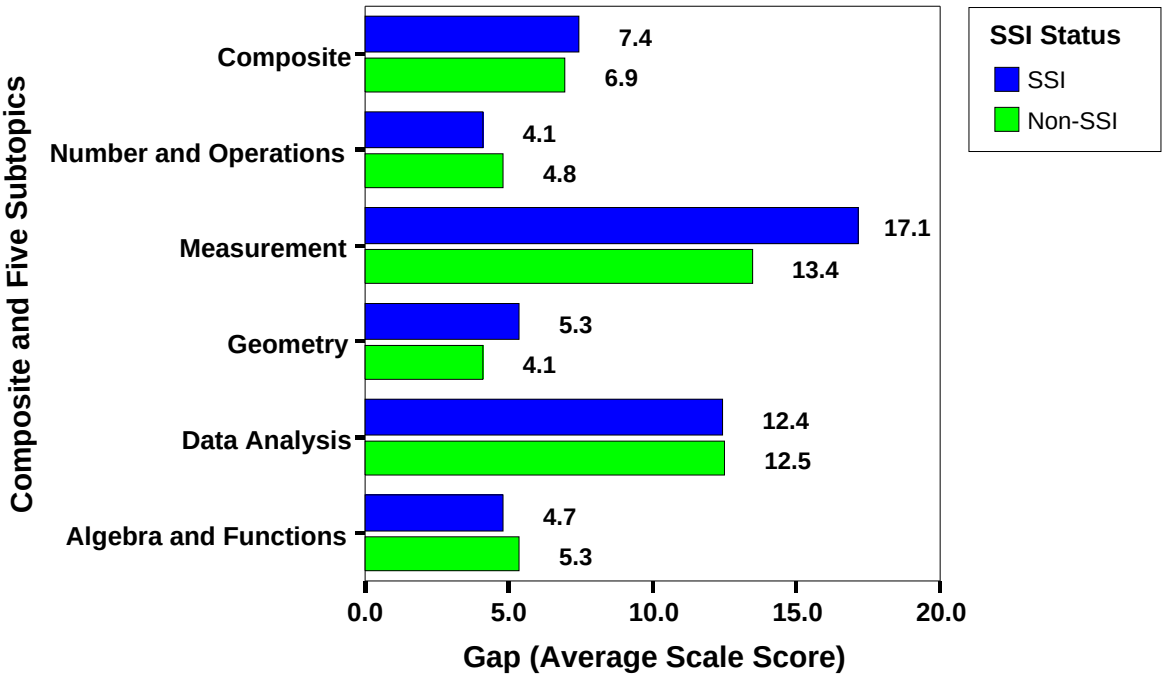


Figure 4B.9. Differences in average scale scores between White and Hispanic students, by SSI status: Cohort group 92-96 (20 SSI and 15 non-SSI).



Appendix C. Narrative Summation of Data for Trend Group 90-96

In the Technical Report by Webb, Kane, Kaufman, and Yang (2001), we reported data for all of the states that participated in the State NAEP in 1990, 1992, and 1996 (the Trend Group 90-96). Appendix C contains the findings reached from our analysis—reported in greater detail in the 2001 Technical Report (available at http://www.wcer.wisc.edu/SSI/SSI/aboutSSI/publications.htm). Data and findings are also based on the Study of the Impact of Statewide Systemic Initiatives presentation made at the Evaluation 2000 Conference in Hawaii, November, 2000, by Webb and Weiss.

NAEP Achievement Findings for the Total Group by SSI Status

1. Both the SSI states and non-SSI states gained in mean scores for grade 8 mathematics from 1990 to 1996. The average performance across the 17 SSI states was lower than the average for the 11 non-SSI states in 1990. This gap narrowed slightly by 1996. (See Webb, Kane, Kaufman, and Yang, 2001, p. 66, Figure 5.1, Trends in average scale scores, by SSI status: Trend Group 90-96.)

The mean score for grade 8 mathematics from 1990 to 1996 increased significantly for both the 17 SSI states (8.3) (F (1,16) = 103.35), $p < .01$) and the 11 non-SSI states (7.1) (F (1, 10) = 77.42), $p < .01$). The average increase in the SSI states was slightly higher than in the non-SSI states by 1.32 points. Prior to the SSI program, SSI states, on average, scored lower than non-SSI states on the NAEP grade 8 mathematics assessment. In 1990, the 17 SSI states averaged 6 points less than the 11 non-SSI states. In 1992, the difference was still about 6 points, and in 1996 it was slightly less, around 5 points.

2. Both the SSI states and non-SSI states gained in mean scores for grade 4 mathematics from 1992 to 1996. The average performance for the 17 SSI states was lower than the average for 11 non-SSI states in 1992. The SSI states gained slightly more than did the non-SSI states from 1992 to 1996. (See Webb et al., 2001, p. 66, Figure 5.1, Trends in average scale scores, by SSI status: Trend Group 90-96.)

The mean score for grade 4 mathematics increased from 1992 to 1996 for both the 17 SSI states and the 11 non-SSI states. In 1992, the SSI average was nearly 5 points lower than the non-SSI average. The gap was reduced by 1 point in 1996.

NAEP Achievement Findings by Gender

3. In grade 8, males in non-SSI states averaged 2 points higher than females in their mean mathematics composite score in 1990 and 1 point higher in 1996. In SSI states, males consistently had a 2-point advantage from 1990 to 1996. In grade 4, males in both SSI and non-SSI states averaged 2 points higher than females in 1992. In 1996, the gap was reduced to 1 point in SSI states but remained at 2 points in non-SSI states. (See Webb et al., 2001, p. 67, Figure 5.2, Trends in average scale scores, by gender and SSI status: Trend Group 90-96.)

There is little evidence from the NAEP data to indicate that the SSIs had any influence on lowering the achievement gap between male and female students. The mean mathematics composite score for female students and male students differed at most by 2 points at grades 4 and 8, at any of the testing times, and for both SSI and non-SSI states. The mean score for grades

4 and 8 for both male and female students increased over time. However, the 2-point advantage males in non-SSI states had at grade 8 in 1990 had decreased by 1 point by 1996. The 2-point gap at grade 8 in SSI-states remained the same over the three testing times. In 1992, at grade 4 the mean mathematics composite score of male and female students differed by 2 points for both SSI and non-SSI states. This gap was lowered to 1 point in SSI states in 1996, but remained the same in non-SSI states.

NAEP Achievement Findings by Ethnicity

4. SSI states, compared to non-SSI states, have proportionately larger racial and ethnic minority groups. Students in these groups have traditionally underperformed in mathematics. (See Webb et al., 2001, p. 69, Figure 5.3, Trends in average scale scores, by race and SSI status: Trend Group 90-96 grade 8 and 92-96 grade 4.)

NSF awarded SSI funds to states with relatively large minority populations. This finding indicates that NSF was successful in targeting states with the greatest need, as indicated by the proportion of students from disadvantaged groups. Using the NAEP database and weighing each state equally, the grade 8 White student populations of the 11 non-SSI states totaled 73% compared to 64% of the 17 SSI states. The 17 SSI states had a higher proportion of Black students (16%) compared to the non-SSI states (9%) and a higher proportion of Hispanic students (14%) compared to the non-SSI states (8%).

5. From 1990 to 1996, White students in the 17 SSI and the 11 non-SSI states gained in their NAEP mathematics composite scores, reflecting the gains for the state as a whole. A somewhat larger percentage of the SSI states showed statistically significant NAEP gains from 1992 to 1996 in both grades 4 and 8. (See Webb et al., 2001, p. 69, Figure 5.3, Trends in average scale scores, by race and SSI status: Trend Group 90-96 grade 8 and 92-96 grade 4.)

Grade 8 White students from both the 17 SSI states and the 11 non-SSI states gained in mathematics achievement from 1990 to 1992 and from 1992 to 1996. Grade 4 White students from both the 17 SSI states and the 11 non-SSI states gained in mathematics achievement from 1992 to 1996. In the SSI states, White students scored lower than in non-SSI states. Over time, the increase for White students in the SSI states was slightly more than the increase in non-SSI states, so the difference between SSI and non-SSI states was smaller at both grades 8 and 4 in 1996.

6. Mean scores in both grades for Black students improved more for 16 SSI states than they did for 6 non-SSI states—the states with minority populations that were large enough to report data for all three years. (See Webb et al., 2001, p. 69, Figure 5.3, Trends in average scale scores, by race and SSI status: Trend Group 90-96 grade 8 and 92-96 grade 4.)

In 1990, grade 8 Black students in 16 SSI states had a mean mathematics score 4 points below Black students in the six non-SSI states. Six years later, Black students in the 16 SSI states scored the same as those in the 6 non-SSI states—again, only those states with a sufficient percentage of Black students to make stable comparisons. The total percentage of Black students in both groups of states was about the same, 16% in the SSI states and 15% in the non-SSI states. From 1990 to 1992, the mean score for grade 8 Black students increased for both SSI states and non-SSI states. For 1992 to 1996, the mean score for grade 8 Black students increased in the SSI states only.

In grade 4, the mean mathematics score of Black students in the 16 SSI states increased by 5 points between 1992 and 1996, compared to the 4-point increase over this period by Black students in the 6 non-SSI states. In 1996, the mean score for Black students in the SSI states was 2 points lower than the mean score for Black students in non-SSI states.

NAEP Content Strand Scores for the Total Group by SSI Status

7. Very few differences are observed in the pattern of achievement among the five mathematics topics tested by NAEP in the three testing years and between SSI and non-SSI states. (See Webb et al., 2001, p. 70, Figure 5.4, Trends in average scale scores on content strands, by SSI status: Trend Group 90-96 grade 8 and 92-96 grade 4.)

SSI states had mean scores on each of the five mathematics topics below the mean scores of the non-SSI states for all testing times for both grade 4 and grade 8. However, both SSI states and non-SSI states increased or remained the same in achievement on each of the five mathematics topics. In general, grade 8 students scored higher on Number/Operations followed by Algebra/Functions, Data Analysis, Measurement, and Geometry. In general, grade 4 students scored lower on Number/Operations than on the other four topics. The greatest gains at both grade 4 and grade 8 were in Algebra/Functions. The smallest gain was in Measurement. Grade 8 students in SSI states gained slightly more than students in non-SSI states on four of the five subscales and grade 4 students in SSI states gained more on all five subscales.

NAEP Content Strand Scores and SSI emphases

8. Differences in state performance by mathematical topics may provide patterns that can be used to match variations in the emphases placed by an SSI on these topics. (See Webb & Weiss, 2000, Figures 13-15, Grade 8 changes in NAEP content subscale scores from 92-96 and SSI emphasis in each content area for Arkansas, Connecticut, and Louisiana.)

Over two NAEP testing times, from 1992 to 1996, Arkansas leaders of mathematics education indicated the SSI gave less emphasis to Number/Operations (8%) and more emphasis to the other four topics—Algebra/Functions (23%), Data Analysis (23%), Geometry (23%), and Measurement (23%). Arkansas grade 8 students in 1996 scored higher on all five topics than grade 8 students in 1992. The 1996 students increased the most on Algebra, Data Analysis, and Geometry. Arkansas students' increase on these three topics was higher than the average increase for all SSI states. Even though there was noticeable change on these topics, Arkansas students still scored below the average of the SSI states on all of the topics.

Connecticut grade 8 students in 1996 scored higher than grade 8 students in 1992 on all five topics. Both the achievement level in 1996 on each of the topics and the gain in achievement from 1992 to 1996 in Connecticut were higher than the average performance and gain by all of the SSI states tested. Connecticut state mathematics leaders reported that the state's mathematics reform over the four years gave equal emphasis to all five topics. With the exception of Number/Operations, the increase in scores from 1992 to 1996 by Connecticut students on the mathematics topics was similar, ranging from a 5- to 8-point gain.

Mathematics education leaders in Louisiana could not report on the different emphasis given to the five topics during 1992 to 1996. The increase in scores by grade 8 students in 1996 compared to grade 8 students in 1992 varied more in Louisiana than in the other two states included in this report. There was no increase in scores for Data Analysis and little increase in scores for

Measurement and Number/Operations. The increase on these three topics was less than the average increase for the SSI states. Louisiana students achieved about the average increase for SSI states for Algebra/Functions and Data Analysis, noticeably higher than their increase on any of the other three topics.

The differentiation in patterns among the SSI states included in this analysis suggests that one way of tracking the impact of an SSI is by comparing differences in the emphasis given by a reform to specific mathematical topics. We have used the report of knowledgeable people to estimate the different emphasis placed on the five topics. But other sources could also be references, such as the state assessments and curriculum frameworks. The differences in patterns observed in analyzing these three states clearly indicate that states vary in the increase in performance by topics.

NAEP Analysis of Achievement Gap between White Students and Black Students

9. SSI states varied in reducing the achievement gap between White and Black students. (See Webb & Weiss, 2000, Figures 16-21.)

Data from three states are reported as an illustration of how SSI states varied in reducing the differences in performance of Black students and White students. Of the Arkansas students who participated in the 1996 NAEP, about 70% were White students and 20% were Black students (Figure 16, State profile: Arkansas). White students in Arkansas, in both grade 4 and grade 8, generally scored from 30 to 35 points higher than Black students. The mean achievement for both White and Black students in Arkansas was lower than the mean achievement for each respective group across the SSI states (Figure 17, NAEP achievement for Black and White students, Arkansas). The achievement gap in grade 4 between White students and Black students increased by about two points from 1992 to 1996. Whereas the achievement gap was below the average gap for the SSI states in 1992, the gap in Arkansas increased, becoming the same as the average gap for the SSI states in 1996. In grade 8, the achievement gap between White and Black students also increased about two points over time, from 1990 to 1996. The achievement gap between White students and Black students in grade 8 was below the SSI average gap in 1990 and 1992, but the Arkansas average gap exceeded the SSI average gap in 1996.

In 1996, over 70% of Connecticut's grade 4 and grade 8 students were White and about 10% of the state's students were Black (Figure 18, State profile: Connecticut). The gap between White students and Black students was larger in Connecticut than in Arkansas. White students in Connecticut generally scored from 33 to 45 points higher than Black students (Figure 19, NAEP achievement for Black and White students, Connecticut). The mean achievement for both White and Black students in Connecticut was equal to or higher than the mean achievement for the respective group across the SSI states. The difference in the mean achievement between Connecticut grade 4 White students and grade 4 Black students exceeded the difference in the mean for SSI states both in 1992 and 1996. However, Connecticut did make progress by lowering the gap between White and Black students in 1996 from the gap in 1992 in grade 4. This was the result of an increase in achievement by Connecticut grade 4 Black students that exceeded that of Connecticut White students, as well as the SSI average for Black students. In contrast, the gap between Connecticut's White students and Black students in grade 8 increased from 1990 to 1996. Over this period of time, scores of Black students stayed nearly the same while scores by White students increased. The widening gap between White students and Black students at grade 8 in Connecticut exceeded the average gap for SSI states.

In Louisiana, in 1996, the percentage of White students, about 50%, nearly equaled the percentage of Black students, about 40% (Figure 20, State profile, Louisiana). Differently from the other two states, the gap between White students and Black students was generally 30 points or less. Louisiana students, both White and Black, scored below the mean score of their respective groups for the SSI states (Figure 21, NAEP achievement for Black and White students, Louisiana). From 1992 to 1996, the difference in the mean between White students and Black students in grade 4 decreased by about two points. Although achievement of grade 4 white students increased over this period of time, about the same as the average for SSI states, the achievement of grade 4 Black students increased more. In grade 8, the gap between White students and Black students steadily increased from 1990 to 1996, but still remained less than the mean gap for the SSI states. At grade 8 in Louisiana, the rate of increase in achievement from 1992 to 1996 was lower than the average increase for the SSI states for both Whites and Blacks. However, Black students in Louisiana increased their achievement at a lower rate, thus widening the gap with White students.

Many factors influence the difference in performance between groups of students. The gap in the average scores of White students and Black students across SSI states remained consistent over two testing periods at grade 4 (31 points) and over three testing periods at grade 8 (35 points). When the results are disaggregated by states, more interesting findings are evident. In both Connecticut and Louisiana at grade 4, the gap between White students and Black students decreased from 1992 to 1996. In Connecticut, Louisiana, and Arkansas, at grade 8, the gap between White students and Black students increased.

CHAPTER 5

REFORM-RELATED CHANGES IN EDUCATIONAL PRACTICES IN SSI AND NON-SSI STATES[1]

Introduction

The National Science Foundation (NSF) instituted the Statewide Systemic Initiatives (SSIs) in 1991 to promote systemic educational change based on high academic standards. The *Curriculum and Evaluation Standards for School Mathematics* (National Council of Teachers of Mathematics, 1989) defined the kind of standards-based curricula NSF encouraged. While states were not required to adopt the NCTM *Standards*, the expectation was that the policies and practices of the SSI states would at least be consistent with them.

Higher mathematics achievement for all students, including those historically underserved, is the ultimate measure of the success of an SSI. NSF considered the achievement goals as the outcome drivers of educational system reform. In addition, NSF identified several process drivers, or policies and practices, in support of high student achievement. This chapter examines the process drivers, using items from the State NAEP teacher questionnaires to create indicators of the drivers.

Since we are using items from the State NAEP, the same measures can be used with all states. In any NAEP year, all states that received SSI funding can be compared to all other states. Change over time can be assessed when the State NAEP teacher questionnaires includes the same or similar items from one year to the next.

The method section of this chapter describes the development of reform-related indicators based on State NAEP teacher questionnaires, and the samples used to study the effects of the SSI program. Differences between SSI and non-SSI states are examined via cross-sectional and longitudinal analyses. Multiple regression modeling is used to describe relationships among the indicators and student achievement. Models were developed from the 1992 and 1996 State NAEP data, with the expectation of testing the models on the 2000 data. However, the 2000 indicators could not be examined because the 2000 State NAEP teacher questionnaire contained very few items compared to the previous years.

Method

Developing Indicators from the NAEP Teacher Questionnaires

The State NAEP, begun in 1990 at grade 8, is designed to estimate parameters for an individual state. Besides the achievement test items, State NAEP includes teacher, student, and

[1] An earlier version of this paper was presented as part of a symposium on Analyzing Statewide Change in Mathematics through NAEP Analyses Linked with Qualitative Analyses, presented at the American Educational Research Association Annual Meeting, New Orleans, 2002.

school questionnaires. Since 1992, the mathematics portion of the State NAEP has been administered every four years.

The research reported here uses items from the State NAEP teacher questionnaires to describe characteristics of the SSI and non-SSI states. The questionnaires requested information about the teachers' backgrounds, general training, and their instructional practices.

Teacher questionnaire item results are often reported in terms of the proportion of students whose teachers selected a specific response option (see, for example, Shaughnessy, Nelson, & Norris, 1997). Analyses are limited to nonparametric approaches that compare two or more groups on the proportion of responses in each category. Statistical models using questionnaire items frequently create dummy variables, collapsing the response categories into a dichotomous variable and, consequently, reducing the information content of the measure. As an alternative approach, we created scales by combining responses to related items. With a scale, random error is reduced and true score variability is increased. A scale simplifies reporting because the responses to several items are combined into a single measure. Scale scores allow the use of parametric statistics when the distribution of scale scores approximates a normal distribution.

We began with an examination of the teacher questionnaires in order to identify items indicative of the goals of the Statewide Systemic Initiatives. We used a model of systemic reform (Clune, 1998) to categorize the items and then reviewed the responses to each selected item. As a result of this review, a few categorized items were eliminated because almost all respondents chose the same response option, usually the highest or lowest. Either there is an extremely high level of teacher agreement, or the items are not sensitive to differences among teachers.

We then reviewed the individual items of the 1996 grade 8 teacher questionnaire and discussed the "best" answer, from the perspective of mathematics reform. For most of the items in the teacher questionnaire, response options ranged from a low of "Never" or "None" to a high of "Almost Every Day" or "A Lot." For most items, responses in the NAEP data set were coded from 1 to N, with N as the number of response options. In our analyses, we reversed the scales when necessary, so the highest value represented the most frequent occurrence. In discussions, project staff generally agreed that with successful statewide systemic initiatives, reform-related practices would increase, but that traditional practices focused on mastering facts, concepts, and routine procedures would also have a major role. We had concerns about a simple scale where "more" of something was considered to be "better" and explored assigning the greater number of points to response options that described a moderate frequency of occurrence. The alternative scales were evaluated using Cronbach's coefficient alpha, a measure of internal consistency (Cronbach, 1951). None of the proposed scoring systems improved on the original 1 to N coding, where 1 indicated the lowest frequency and N the highest.

The extensive review and analysis of the State NAEP 1996 teacher questionnaire items resulted in six indicators of mathematics reform:

I(RC), Relative Emphasis on Reasoning and Communication—how much reasoning and communication were addressed, relative to facts and procedures.

I(MD), Mathematical Discourse—a scale of students' opportunities to discuss, present, and write about mathematical ideas based on nine items in 1996 and seven items in 1992
I(MD4)—a scale based on four items from the I(MD) scale that were exactly the same in 1992 and 1996.

I(C), Calculator Use—a scale of the extent to which students used calculators in the classroom and on tests.

***I(S), NCTM* Standards**—a single item that asked about teachers' knowledge of the NCTM *Standards*.

I(PD), Last Year's Professional Development—a single item that asked how much time teachers spent in professional development in mathematics or mathematics education during the last year.

I(RT), Reform-Related Topics Studied—a count of the number of reform-related topics teachers have studied out of the seven topics listed in the NAEP questionnaire.

Additional details about the indicators are included in the project's technical report (Webb, Kane, Kaufman, & Yang, 2001).

The 1996 State NAEP teacher questionnaire was not the same as the questionnaire administered in 1992. Several items were added, particularly items related to curricular reform. Wording of some items was modified and, for some items, the number and labels of the response options were changed. Despite these differences, the similarities of the questionnaire items in 1992 and 1996 provided a means for comparing SSI and non-SSI states across time.

Samples

Twenty-five states and Puerto Rico received funding through NSF's SSI program, and 25 states did not. NSF discontinued funding early for four states, resulting in 21 states with the full five years of funding. Under the SSI program, awards were made in three cohorts—the first in 1991, the second in 1992, and the third in 1993.

Not all states participated in State NAEP in any given year. In this report, analyses and conclusions about the effects of the SSI program are limited to those states that chose to participate in State NAEP. While State NAEP also included data from the jurisdictions of Guam, Puerto Rico, the Virgin Islands, Washington, DC, and Department of Defense Schools, only state data were used for this study.

Yearly samples—1992 and 1996. For each year of the State NAEP, comparisons can be made between all participating SSI states and non-SSI states, using all of the available data in a given year. Table 5.1 presents the number and percentage of SSI and non-SSI states participating in State NAEP each year at each grade level.

Table 5.1
Number and Percentage of SSI and Non-SSI States in Each Yearly Sample

	SSI States *n* = 21		Non-SSI States *n* = 25	
	n	%	*n*	%
1992				
Grade 8	18	86%	19	76%
Grade 4	18	86%	19	76%
1996				
Grade 8	18	86%	18	72%
Grade 4	19	90%	20	80%

Trend sample, 1992-2000. The trend sample is comprised of states that participated in three consecutive State NAEP administrations: 1992, 1996, and 2000. (See Table 5.2.) Fourteen SSI states (67% of all SSI states) and 13 non-SSI states (52% of all non-SSI states) are in the trend sample. By 1996, the first SSI cohort was completing its fifth year, and others were well into their third or fourth years.

While the 1992 measure provides a baseline for the 1996 measure, it is not necessarily independent of SSI. Since the first round of NSF funding began in 1991, some of the states had been funded for a time. More importantly, some of the states had extensive prior experience with reform initiatives, positioning them to be interested in and selected for NSF's Statewide Systemic Initiatives Program.

Table 5.2
Trend-Sample States

SSI States *n* = 14	Non-SSI States *n* = 13
Arkansas	Alabama
California	Arizona
Connecticut	Hawaii
Georgia	Indiana
Kentucky	Maryland
Louisiana	Minnesota
Maine	Mississippi
Massachusetts	Missouri
Michigan	North Dakota
Nebraska	Tennessee
New Mexico	Utah
New York	West Virginia
South Carolina	Wyoming
Texas	

Unit of Analysis

State NAEP is designed to provide information about each state as a whole. The student is the basic data unit, and teachers' responses are matched with each of their students to define one record in the data file. Each student has an associated weight, based on the sampling plan, and state means are computed using weighted values (Allen, Jenkins, Kulick, & Zelenak, 1997). In the analyses reported here, the focus is on the state means and the variability among the means, rather than on the within-state variability.

With the state as the unit of analysis, SSI states are grouped together as replications receiving the treatment (e.g., the SSI program), and non-SSI states are grouped as replications not receiving the treatment. The SSI states used many, and varied, approaches to systemic reform. However, grouping the states together assumes that each belongs to a general category, despite their differences. The statistical comparisons allow conclusions about whether something is more or less likely to occur in one group than another. There is no claim that all states in one group will share a characteristic, or that the characteristic will not be present in any of the states in the comparison group(s).

Another caution in interpreting the results of these analyses is that, unlike experimental research, the SSI treatment was not randomly assigned to the states. States participating in the SSI program had to submit a proposal, and NSF selected which proposals it would fund.

With states as the unit of analysis, the sample size is fairly small. In order to reject the null hypothesis, differences have to be fairly large. For comparisons between SSI and non-SSI states, we used an alpha level of .10 (Grissmer, Flanagan, Kawata, & Williamson, 2000).

Controlling Sources of Variability

Socioeconomic status. The NAEP Toolkit includes procedures for creating a socioeconomic status (SES) variable using a student's answers to six items, including mother's and father's education and the presence of educational materials in the child's home. The SES variable is computed by taking the mean of a set of z-scores. In the analyses reported here, the SES variable is used as a covariate to adjust for sources of variability unrelated to a state's SSI status.

State testing program. During the 1990s, many states focused on educational reform. The Goals 2000 program provided resources to states that were developing state standards and/or frameworks, and states were increasingly implementing assessment and accountability programs, with the goal of raising student achievement.

Earlier work on the indicators found a relationship between I(C), Calculator Use, and whether students were permitted to use calculators on the state test (see Webb et al., 2001, p. 211-237). In this chapter, the analyses include a factor indicating whether or not states had assessment programs based on criterion-referenced tests in grades 3-8. As the next chapter describes, our study of the 14 SSI states in the trend sample found that criterion-referenced tests

seemed to be associated with achievement gains. By using it as a factor in these analyses, we hoped to account for sources of variability that were not directly related to the SSI program.

Information about state testing programs was obtained from the Fall, 1996, *Annual Survey of State Student Assessment Programs* (Roeber, Bond, & Braskamp, 1997). Each state was classified in one of two groups, based on whether it had criterion-referenced tests in mathematics in two or more grades below high school, as reported in *Survey* Tables 3.02 and 3.10. Table 5.3 lists the states on the basis of whether they used criterion-referenced tests (CRTs) in mathematics in 1996. (Information for each state is listed in Appendix 5.1.)

Table 5.3
Listing of SSI and Non-SSI States that Used or Did Not Use Criterion-Referenced Tests in Mathematics at Two or More Grades Levels Below High School in 1996

	SSI States	Non-SSI States
CRT States	Connecticut[a] Georgia[a] Louisiana[a] Massachusetts[a] Michigan[a] New York[a] Ohio South Carolina[a] Texas[a] Vermont	Indiana[a] Illinois Maryland[a] Missouri[a] New Hampshire Oklahoma Oregon Tennessee[a] Utah[a] West Virginia[a]
Non-CRT States	Arkansas[a] California[a] Colorado Delaware Kentucky[a,b] Maine[a] Montana Nebraska[a] New Jersey New Mexico[a]	Alabama[a] Alaska Arizona[a] Hawaii[a] Idaho Iowa Minnesota[a] Mississippi[a] North Dakota[a] Pennsylvania Washington Wisconsin Wyoming[a]

[a]Trend sample state
[b]Kentucky was the most difficult to classify, because the assessments are described as criterion-referenced in some places. We decided to base the coding on the information in the selected tables of the annual report. However, all analyses were repeated with Kentucky as a CRT states, as a check on the findings. Results were comparable.
Note: States with less than four years of SSI funding were not included in the analyses.

Within each SSI group, the number of states with criterion-referenced student assessments was similar to the number without such assessments, as shown in Table 5.3. A 2 x 2 analysis of variance (ANOVA), with SSI status crossed with CRT status, was used to examine differences between groups on socioeconomic status (SES) at grade 4 and grade 8 in 1992 and 1996. No significant differences in SES were found.

Results

Cross-Sectional Comparisons of SSI and Non-SSI States on Indicators of Mathematics Reform

We examined the effect of the SSI program by comparing all SSI and non-SSI states in a given year. As explained above, the socioeconomic status variance was used as a covariate, and use of CRT-based accountability was used as a design factor. Because no significant effects for CRT were found in these analyses, Table 5.4 reports the results for SSI only.

In 1992, close to the beginning of the SSI program, the SES variable was significantly related to I(C), Calculator Use, at grade 8 (multivariate $F = 6.117$, $p < .05$) and to I(C) and I(PD), Time in Professional Development, at grade 4 (multivariate $F = 2.37$, $p < .10$). States with higher values for SES averaged higher in I(C) at both grade levels. SES was inversely related to I(PD) at grade 4, with low SES states averaging higher on the professional development indicator. The SSI effect was not statistically significant.

In 1996, the SES covariate was significantly related to I(C) and I(S) at both grades, as well as to I(RC) at grade 8 ($F = 10.99$, $p < .01$) and I(PD) at grade 4 ($F = 5.20$, $p < .01$). SES was positively associated with the indicators, except for I(PD), where states low in SES tended to have higher levels of professional development, replicating the 1992 result. In addition, the multivariate F for SSI was statistically significant at both grade 8 ($F = 2.41$, $p < .10$) and grade 4 ($F = 3.41$, $p < .05$). At both grade levels, SSI states averaged higher than non-SSI states on I(RC), Relative Emphasis on Reasoning and Communication, and I(MD), Use of Mathematical Discourse. At grade 8, SSI states also averaged significantly higher on I(S), Teachers' Knowledge of the NCTM *Standards.* At grade 4, SSI states also averaged higher on I(PD), Time in Professional Development in the Last Year.

Table 5.4
Indicator Means for all SSI and Non-SSI States that Participated in State NAEP Each Year

	SSI states		Non-SSI States			
1992[a]	*M*	*SE*	*M*	*SE*	$F_{(SSI)}$	$F_{(SES)}$
I(RC)						
Grade 8	43.50	0.50	42.76	0.49	1.08	1.72
Grade 4	40.20	0.41	39.12	0.40	3.47	0.32
I(MD)						
Grade 8	15.62	0.17	15.25	0.17	2.27	0.04
Grade 4	15.81	0.19	15.30	0.18	3.62	0.14
I(C)						
Grade 8	11.45	0.23	11.12	0.23	0.95	6.12*
Grade 4	7.31	0.14	7.18	0.14	0.42	3.10^{+}
I(PD)						
Grade 8	3.25	0.08	3.26	0.05	0.01	0.55
Grade 4	2.57	0.05	2.60	0.05	0.18	3.06^{+}
I(RT)						
Grade 8	4.76	0.08	4.70	0.08	0.29	0.97
Grade 4	4.72	0.06	4.73	0.06	0.01	0.04
1996[b]						
I(RC)						
Grade 8	46.03	0.33	44.44	0.33	11.58*	7.54*
Grade 4	44.04	0.30	42.39	0.30	15.28*	0.31
I(MD)						
Grade 8	23.29	0.27	22.33	0.27	6.18*	0.01
Grade 4	23.88	0.27	23.00	0.27	5.37*	0.08
I(C)						
Grade 8	9.99	0.14	9.79	0.14	1.07	18.85*
Grade 4	8.11	0.10	7.90	0.10	2.15	10.27*
I(S)						
Grade 8	2.74	0.04	2.57	0.04	11.03*	18.48*
Grade 4	1.98	0.04	1.89	0.04	2.93^{+}	6.78*
I(PD)						
Grade 8	3.47	0.07	3.30	0.07	2.49	0.42
Grade 4	2.88	0.06	2.70	0.06	5.01*	6.29*
I(RT)						
Grade 8	5.18	0.73	5.02	0.73	2.33	0.06
Grade 4	4.87	0.06	4.81	0.06	0.45	2.58

[a]The 1992 State NAEP sample had 18 SSI states and 19 non-SSI states at both grades 4 and 8.
[b]The 1996 State NAEP sample had 18 SSI states and 18 non-SSI states at grade 8 and 19 states in each group at grade 4.

*$p < .05$, $^{+}p < .10$

The finding that SSI states averaged higher than the non-SSI states on several indicators of mathematics reform in 1996 and not in 1992 is evidence for the effectiveness of the SSI program. However, with cross-sectional studies, the varying sample from one year to the next is a possible confounding. Since a large percentage of the sample participated each year, this seems like a relatively small threat to external validity. In this study, though, we can use the trend sample to examine change for the same set of states.

Longitudinal Comparisons of SSI and Non-SSI States on Indicators of Mathematics Reform

A longitudinal design has the potential to provide stronger evidence for the effectiveness of the SSI program because change over time can be identified. However, the analyses have somewhat reduced power because of the smaller sample size: longitudinal comparisons were limited to those states that consistently participated in State NAEP—that is, the trend sample of 14 SSI states and 13 non-SSI states listed in Table 5.2.

In the longitudinal comparisons, state means were compared across 1992 and 1996. We had hoped to extend the analyses through 2000, but the 2000 State NAEP teacher questionnaire did not include the items from prior years

Representativeness of the trend sample. The representativeness of the trend sample was examined by comparing the means of the trend sample with the means of the other states participating in State NAEP in a given year (Table 5.5). In general, the mean for the trend sample is not significantly different from the mean of other states participating in State NAEP. Exceptions include the SES variable in all four comparisons and the highly correlated Calculator Use indicator in 1996 at grade 8. For these measures, the trend-sample states average significantly lower than the other states. When the SES variable is used to adjust I(C(96)), there is no significant difference between the trend-sample states and other states. In summary, states in the trend sample average slightly lower in socioeconomic status than states that did not participate consistently in State NAEP. Measures of reform-related indicators seem comparable for the two groups.

Table 5.5

Means for the Trend-Sample States and Other States on the SES Variable and Reform-Related Indicators.

	Trend Sample[a]		Other States[b]		
1992	*M*	*SE*	*M*	*SE*	$F_{(Trend\ vs.\ Others)}$
Grade 8					
SES	-0.03	0.03	0.05	0.02	4.15*
I(RC)	42.78	1.98	44.03	2.28	2.71
I(MD)	15.42	0.70	15.47	0.81	0.47
I(C)	11.24	1.10	11.38	0.88	0.13
I(PD)	3.26	0.22	3.24	0.22	0.12
I(RT)	4.69	0.37	4.82	0.25	1.06
Grade 4					
SES	-0.03	0.02	0.05	0.03	4.23*
I(RC)	39.62	1.99	39.70	1.18	0.02
I(MD)	15.59	0.88	15.44	0.57	0.25
I(C)	7.23	0.65	7.27	0.50	0.02
I(PD)	2.60	0.23	2.53	0.15	0.84
I(RT)	4.70	0.27	4.80	0.17	1.19
1996					
Grade 8[c]					
SES	-0.05	0.03	0.10	0.06	4.80*
I(RC)	45.02	1.55	45.87	1.86	1.86
I(MD)	22.69	1.30	23.16	0.98	0.94
I(C)	9.75	0.74	10.30	0.30	4.71*
I(S)	2.63	0.20	2.73	0.18	1.78
I(PD)	3.41	0.32	3.31	0.30	0.75
I(RT)	5.10	0.31	5.10	0.32	0.00
Grade 4[d]					
SES	-0.03	0.02	0.04	0.03	3.17+
I(RC)	43.19	1.42	42.22	1.75	0.00
I(MD)	23.38	1.28	23.51	1.08	0.09
I(C)	7.95	0.57	8.12	0.30	0.98
I(S)	1.90	0.18	2.00	0.20	2.39
I(PD)	2.80	0.27	2.74	0.27	0.48
I(RT)	4.83	0.26	4.82	0.31	0.00

[a] $n = 27$

[b] Other states are those that participated in State NAEP in a given year but were not part of the trend sample. In 1992, grade 8 $n = 10$ and grade 4 $n = 9$; in 1996, grade 8 $n = 9$ and grade 4 $n = 11$.

*$p < .05$, +$p < .10$

Trend analyses. For the trend sample, two different approaches were used to analyze change from 1992 to 1996 on the classroom practice indicators. Both used SSI status and CRT

accountability as factors in the analysis, along with the SES variable as a covariate. For indicators that had the same scale in 1992 and 1996, a repeated measures analysis of variance was used to examine absolute change over time, as well as the interactions of time, SSI status, and use of criterion-referenced tests. For indicators with different scales in 1992 and 1996, the 1992 measure was used as a covariate along with the SES variable.

Most of the 1992 indicators have strong positive correlations with the corresponding 1996 indicators, as shown in Table 5.6. The high correlations suggest that a specific culture of education that is sustained across the years. Even when means increase or decrease over time, states tend to keep their relative position on the indicators. Correlations for grade 4 are slightly higher than for grade 8, suggesting more continuity in the measures in the lower grade.

Table 5.6
Correlations Between 1992 and 1996 Reform-Related Indicators at Grade 8 and Grade 4

	Correlation of 1992 and 1996 measures		Correlation of Grade 4 and Grade 8	
Indicator[a]	Grade 8	Grade 4	1992	1996
I(RC)	.67*	.78*	.63*	.64*
I(MD)	.66*	.74*	.75*	.80*
I(C)	.71*	.77*	.73*	.61*
I(S) [a]				.76*
I(PD)	.23	.56*	.48*	.73*
I(RT)	.53*	.62*	.69*	.75*

[a]I(S) is was not part of the 1992 teacher questionnaire.

*$p < .05$

Covariance analyses. Table 5.7 lists the results from the covariance analyses for the trend sample. All but one of the 1992 indicators were significantly related to the corresponding 1996 indicator. The one exception was I(PD) at grade 8. In addition, the SES variable was significantly related to I(C) and I(S) at both grades and I(RC) at grade 8, replicating the findings from the cross-sectional analyses. The trend analyses found that SSI status was related to I(RC(96)) at both grades, I(S(96)) at grade 8, and I(C(96)) and I(PD(96)) at grade 4. These results replicate the findings from the cross-sectional analyses. Unlike the findings from the yearly samples, SSI was not related to I(MD(96)) in the trend sample.

Table 5.7
1996 Indicator Means for Trend-Sample States, Adjusted for 1992 Values and Socioeconomic Status.

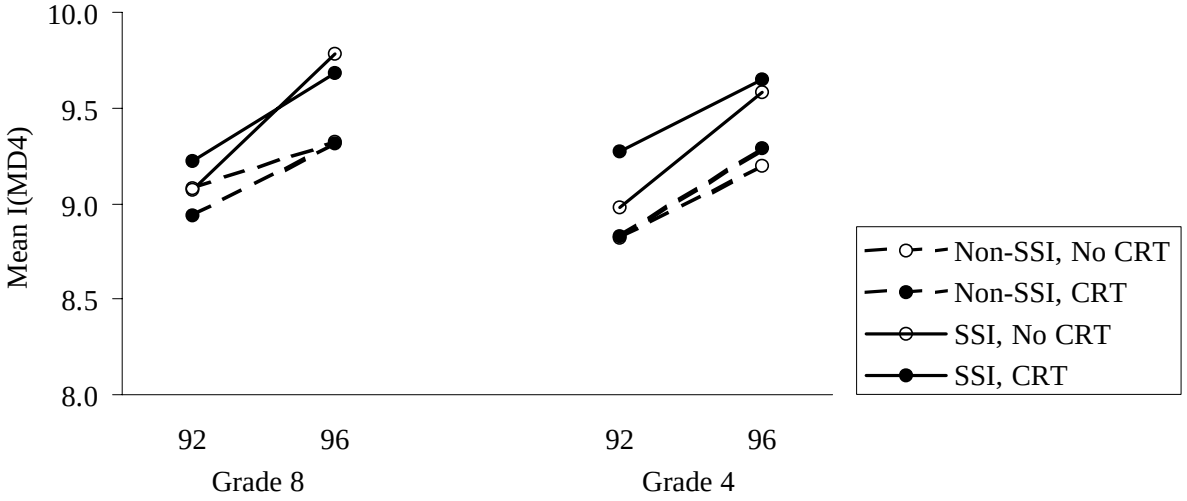
	SSI states		Non-SSI States			
	M	*SE*	*M*	*SE*	$F_{(SSI)}$	$F_{(SES)}$
I(RC)						
Grade 8	45.57	0.27	44.45	0.28	7.97*	3.83^{+}
Grade 4	43.59	0.20	42.78	0.21	7.11*	0.01
I(MD)						
Grade 8	22.88	0.28	22.53	0.28	0.69	0.01
Grade 4	23.49	0.23	23.29	0.24	0.82	0.35
I(C)						
Grade 8	9.86	0.14	9.64	0.15	1.04	5.59*
Grade 4[a]	8.08	0.08	7.85	0.08	4.16^{+}	7.78*
I(S)						
Grade 8	2.70	0.05	2.56	0.05	4.25^{+}	10.52*
Grade 4	1.95	0.05	1.85	0.05	2.09	8.32*
I(PD)						
Grade 8	3.44	0.09	3.38	0.09	0.22	0.16
Grade 4	2.92	0.06	2.68	0.06	8.33*	1.84
I(RT)						
Grade 8	5.17	0.07	5.04	0.72	1.68	0.42
Grade 4	4.86	0.06	4.80	0.06	0.43	0.89

[a]At grade 4, the interaction of SSI and CRT was significant, F = 8.63, $p < .05$. See Figure 5.2 below.
*$p < .05$, $^{+}p < .10$

Repeated measures analyses. Two of the indicators, I(PD), professional development, and I(RT), reform-related topics, have the same scale in 1992 and 1996, along with I(C) for grade 4. In addition, I(MD4), based on the four mathematical discourse items common to I(MD(92)) and I(MD(96)), can be used to examine absolute change across the four years.

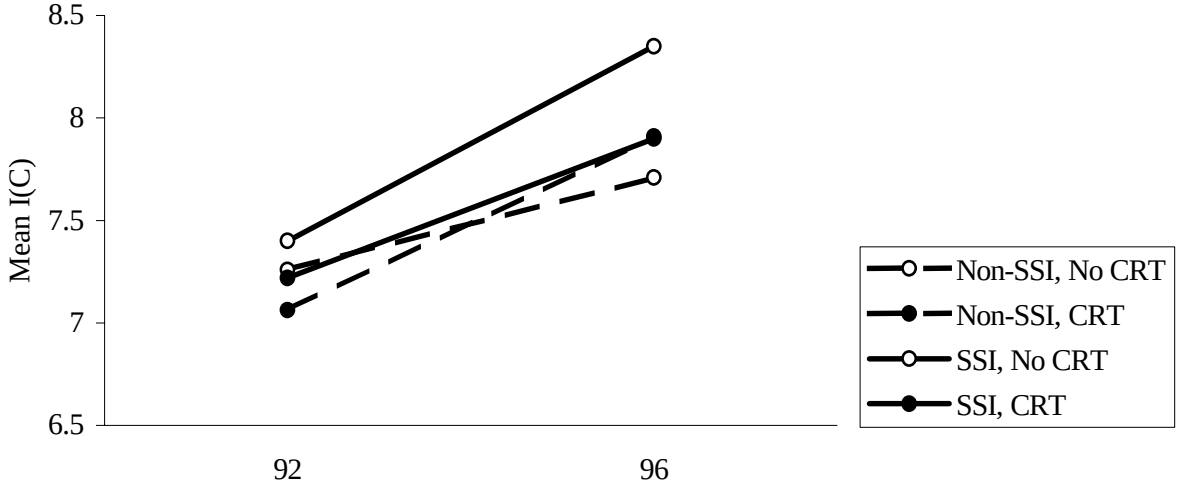
The results for I(MD4) are graphed in Figure 5.1. I(MD4) increased from 1992 to 1996 at both grades (for grade 8, $F = 26.81$, $p < .01$; for grade 4, $F = 44.83$, $p < .01$). In addition, the main effect for SSI was statistically significant (for grade 8, $F = 2.94$, $p < .10$; for grade 4, $F = 3.08$, $p < .10$), but the Year x SSI interaction was not, indicating that SSI states in the trend sample averaged higher on I(MD4) in both years.

Figure 5.1. Change in I(MD4) from 1992 to 1996 as a function of SSI status and the use of criterion-referenced state testing, adjusted for socioeconomic status.



At grade 4, I(C), an indicator of students' opportunities for calculator use, also increased significantly from 1992 to 1996 ($F = 98.87$, $p < .01$). In addition, there was a significant three-way interaction of Year x SSI status x CRT ($F = 4.98$, $p < .05$). Figure 5.2 illustrates the interaction, with values adjusted for SES ($F = 4.18$, $p < .10$). As the graph shows, all groups increased in students' use of calculators from 1992 to 1996, but non-SSI states without CRT programs had the smallest increase.

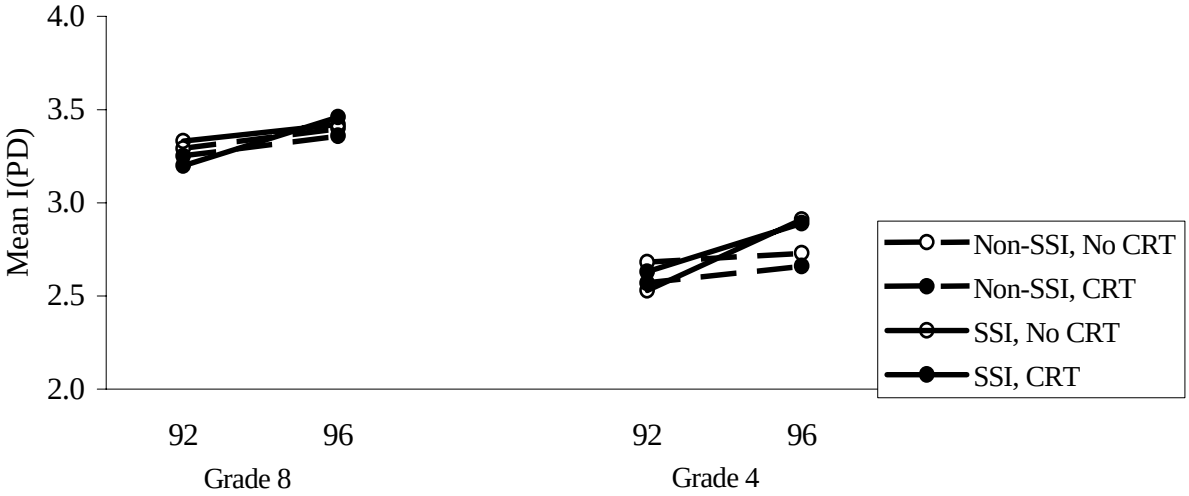
Figure 5.2. Grade 4 change in I(C) from 1992 to 1996 as a function of SSI status and the use of criterion-referenced state testing, adjusted for socioeconomic status.



Repeated measures analyses of I(PD), an indicator of time in professional development, found a significant increase from 1992 to 1996 at grade 4 ($F = 18.63$, $p < .01$) along with a significant interaction of Year x SSI status ($F = 8.39$, $p < .01$). As Figure 5.3. shows, the SSI

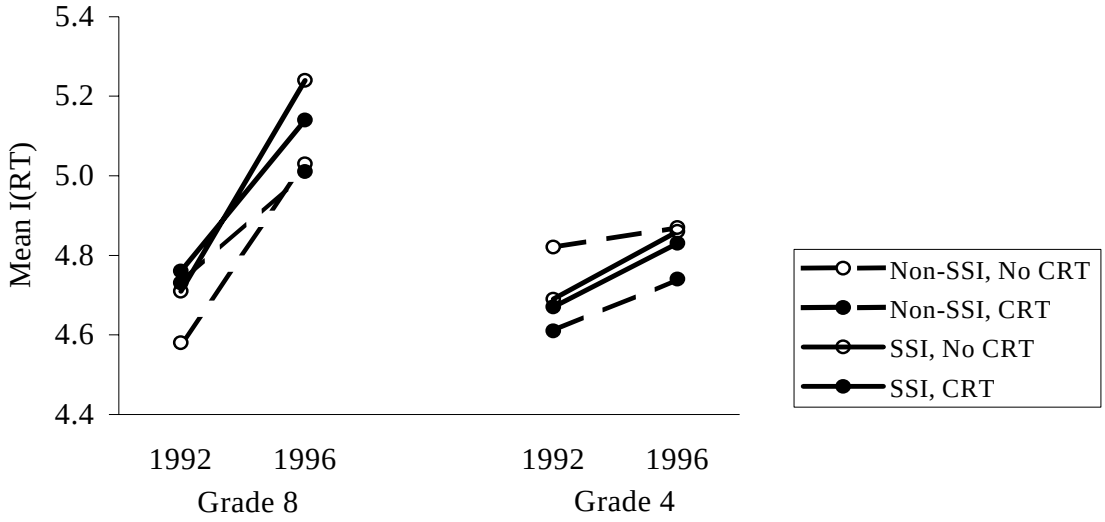
states had a larger universe in I(PD) than the non-SSI states. I(PD) also increased at grade 8 (F = 3.92, $p < .10$) for all states.

Figure 5.3. Change in I(PD) from 1992 to 1996 as a function of SSI status and the use of criterion-referenced state testing, adjusted for socioeconomic status.



I(RT), reform-related topics studies by teachers, also increased significantly from 1992 to 1996 at both grade 8 (F = 46.93, $p < .01$) and grade 4 (F = 6.62, $p < .05$), as shown in Figure 5.4.

Figure 5.4. Change in I(RT) from 1992 to 1996 as a function of SSI status and the use of criterion-referenced state testing, adjusted for socioeconomic status.



The repeated measures analyses show that I(MD4), I(C), I(PD), and I(RT) all increased significantly from 1992 to 1996. The significant effect for SSI in the I(MD4) analyses helps to

explain the lack of significant results with the trend sample. As Figure 5.1 shows, in the trend sample, SSI states averaged higher than non-SSI states in 1992. Consequently, adjusting for I(MD(92)) would reduce the differences between SSI and non-SSI states.

Summary. Six potential indicators of educational reform were developed from the State NAEP teacher questionnaire. Cross-sectional comparisons for 1996 found that SSI states averaged significantly higher than non-SSI states on three of the six indicators at both grade levels: I(RC), emphasis on Reasoning and Communication; I(MD), students' opportunities for Mathematical Discourse, and I(S), teachers' knowledge of the NCTM *Standards*. In addition, SSI states averaged higher than non-SSI states on I(PD), time in Professional Development, at grade 4. Longitudinal analyses, using a smaller trend sample of 27 states, found that SSI states increased more than non-SSI states on I(RC) at both grade levels. At grade 4, SSI states also increased more in I(PD), the time teachers' spent in mathematics-related staff development during the last year and I(C), Calculator Use. SSI states in the trend sample averaged higher than non-SSI states on I(MD4) in both 1992 and 1996.

Correlations of the indicators across years and across grade levels within the same year were relatively high. This pattern indicates that states differ in their educational practices and that these differences are endured across the four years. It seems that changes over the four years were gradual, building from educators' current practices. Only I(PD), time in Professional Development, seemed to fluctuate substantially from year to year.

Relationships Between the Indicators and Student Achievement: Multiple Linear Regression Modeling

The relationships of the indicators to student achievement were examined using multiple regression modeling. This work was exploratory and directed to model development. The goal was to develop hypotheses to be tested with the 2000 State NAEP data. Because the 2000 State NAEP used only a limited teacher questionnaire, we were not able to do the model testing.

Figure 5.5 illustrates relationships among the six indicators in the model. The indicators fall into three groups, related causally. The two indicators on the left, I(PD) and I(RT), are enclosed in a broken line box to represent teachers' opportunity learn, that is, time spent in professional development in mathematics over that last year, and reform-related topics they have studied at some time during their career. The next group in the model is represented by only one indicator, I(S). This represents the teachers' knowledge and skills, or what they have learned through their studies and professional development. In this model, the indicator refers specifically to teachers' knowledge of the NCTM *Standards*. The next set of three indicators represents what teachers actually do while teaching. The indicators include and include instructional goals as well as teaching practices. These indicators are expected to be most directly related to student achievement and to result from the education and training teachers have completed. The SES variable is one component of "Other" along the bottom of the figure.

Figure 5.5. Indicators of mathematics curricular reform and their relationship to student achievement.

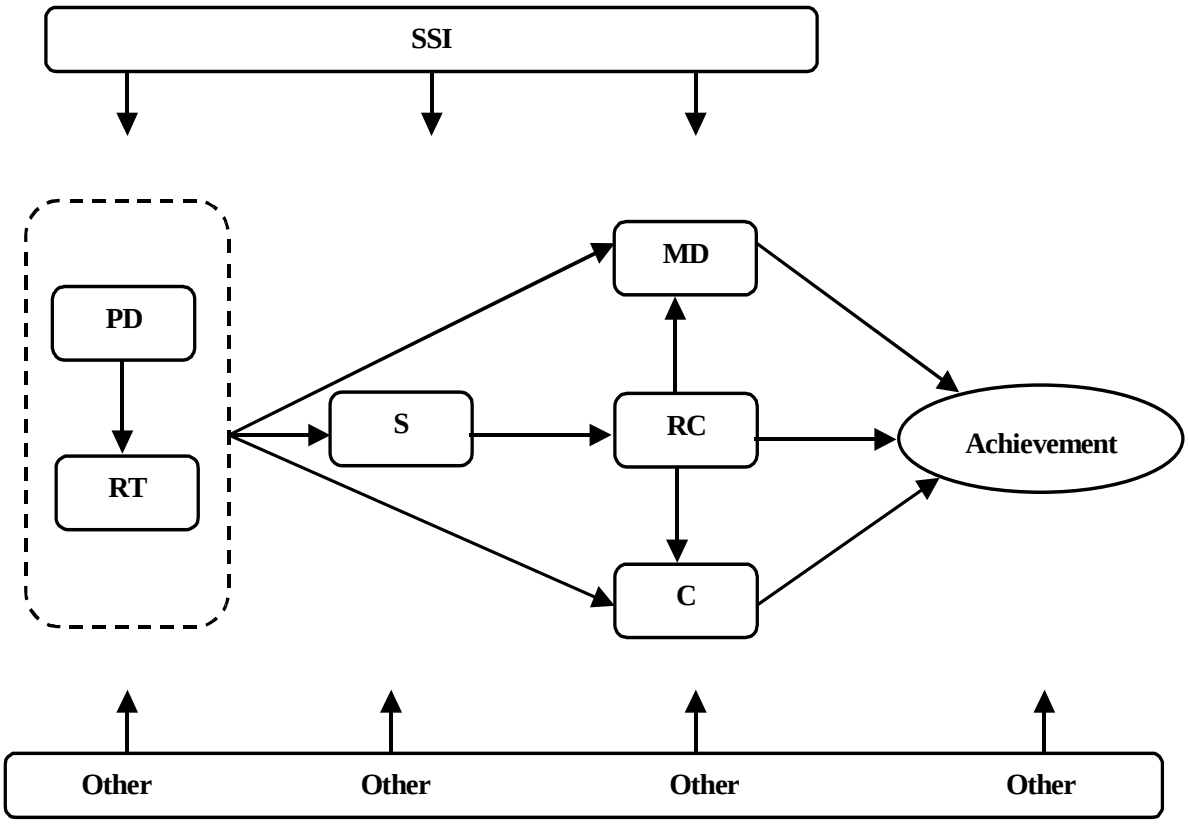


Table 5.8 presents the intercorrelations of the six indicators, the SES variable, and the state mean mathematics composite scores for both grade 4 and grade 8 in 1996. As the table shows, the SES variable accounts for much of the variability in the mean State NAEP mathematics composite. In addition, SES is significantly related to two indicators in both grades: I(C) and I(S).

The three classroom practice indicators are strongly interrelated, with the highest correlation between I(RC) and I(MD) at both grade 8 and grade 4. I(S), the teacher's knowledge of the NCTM *Standards*, is correlated with the three classroom practice indicators at both grades 4 and grade 8. At grade 8, knowledge of the NCTM *Standards* is related to time spent in mathematics-related staff development during the last year, but not at grade 4. In addition, I(PD) is positively correlated with I(PD) at both grade 8 and grade 4.

Correlations of the indicators with the mathematics composite differ somewhat between grades 8 and 4. At both grades, the mean mathematics composite is significantly related to the calculator use indicator, and it is not significantly correlated with the mathematical discourse indicator. At grade 8, two additional indicators are also positively related to the mathematics composite: the relative emphasis on reasoning and communication and teachers' knowledge of

the standards. At grade 4, two indicators are negatively related to achievement—amount of professional development during the last year and the number of reform topics studied.

Table 5.8
Intercorrelations Among the Six Indicators of Mathematics Reform and the State NAEP Mathematics Composite at Grade 8 and Grade 4, 1996

	Composite	SES	I(RC)	I(MD)	I(C)	I(S)	I(PD)
Grade 8 (n = 36)							
SES	.87*						
I(RC)	.32^{+}	.32^{+}					
I(MD)	-.05	-.05	.68*				
I(C)	.64*	.59*	.54*	.48*			
I(S)	.31^{+}	.50*	.73*	.57*	.58*		
I(PD)	-.20	-.15	.41*	.48*	.18	.43*	
I(RT)	-.01	-.08	.28^{+}	.49*	.33*	.18	.39*
Grade 4 (n = 38)							
SES	.85*						
I(RC)	-.17	-.10					
I(MD)	-.16	-.06	.87*				
I(C)	.38*	.46*	.26	.46*			
I(S)	.11	.38*	.41*	.50*	.50*		
I(PD)	-.35*	-.38*	.51*	.44*	.01	.21	
I(RT)	-.32*	-.26	.35*	.49*	.18	.13	.43*

*$p < .05$, $^{+}p < .10$

Using multiple linear regression to assess the model in Figure 5.5 presents some problems because of the relatively high correlations among several predictors. Because multiple regression assumes the predictors are independent, relationships among predictors raise issues about how to estimate the model parameters. In part, this problem can be solved by the model specification. If the model must include all predictors, an analytic method that will divide the shared variance among the predictors is often used.

According to the model in Figure 5.5, variation in I(S), I(PD), and I(RT) is expected to be reflected in variation in the classroom practice indicators, I(RC), I(MD), and I(C). Socioeconomic status is expected to influence the model at many different points.

Grade 8, 1996. At grade 8, the parsimonious model for predicting Y, the state's mean mathematics composite, was

Y = .85*SES + .21*I(RC) + .27*I(C) -.42*I(S).

With just SES as a predictor, the model R^2 was .73 (F = 97.25, $p < .05$). Adding the additional indicators increased the adjusted R^2 to .82. Including I(RC), I(C) and I(S) along with SES significantly improves the prediction of the 1996 mathematics composite, compared to a model

based on SES alone. In the equation, I(S) has a negative coefficient, indicating that the variability in I(S) that is unrelated to the other measures is inversely related to the mean mathematics composite.

Grade 4, 1996. At grade 4,the state mean mathematics composite was predicted by:

Y = .91*SES - .29*I(S).

The adjusted R^2 for the model was .73 (F = 40.66, $p < .05$). As in the model for grade 8, I(S), an indicator of teachers' knowledge of the NCTM *Standards*, had a negative coefficient, even though I(S) was positively related to student achievement, as shown in Table 5.8.

To provide perspective on the 1996 models, data from the 1992 State NAEP was also used to evaluate the model in Figure 5.5. Table 5.9 lists the intercorrelations for the 1992 data.

Table 5.9
Intercorrelations Among the Six Indicators of Mathematics Reform and the State NAEP Mathematics Composite at Grade 8 and Grade 4, 1992

	Composite	SES	I(RC)	I(MD)	I(C)	I(PD)
Grade 8 (n = 37)						
SES	.85*					
I(RC)	.26	.18				
I(MD)	-.00	-.09	.74*			
I(C)	.57*	.36*	.49*	.50*		
I(PD)	-.06	-.12	.32^{+}	.43*	.26	
I(RT)	-.13	-.08	.43*	.53*	.10	.42*
Grade 4 (n = 37)						
SES	.87*					
I(RC)	-.15	-.14				
I(MD)	-.07	-.11	.91*			
I(C)	.26	.28	.44*	.47*		
I(PD)	-.35*	-.28^{+}	.38*	.25	.34*	
I(RT)	-.13	.04	.41*	.37*	.46*	.36*

*$p < .05$, $^{+}p < .10$

Grade 8, 1992. For grade 8 in 1992, the mean state mathematics composite was predicted by the SES variable and I(C). Adding I(C) to the model increased the adjusted R^2 from .71 to .78 (F = 13.03, $p < .01$):

Y = .74*SES + .30*I(C)

As in 1996, the SES variable is strongly related to the mean mathematics composite. As in 1996, adding I(C), an indicator of calculator use, adds to the prediction of mean mathematics achievement.

Grade 4, 1992. In 1992, the state mean mathematics composite was predicted by SES and I(RT), an indicator of the number of reform-related topics teachers had studied:

Y = .87*SES -.16*I(RT).

The adjusted R^2 was .77 ($F = 62.68$, $p < .01$). As in 1996, the coefficient for the indicator of teachers' knowledge is negative.

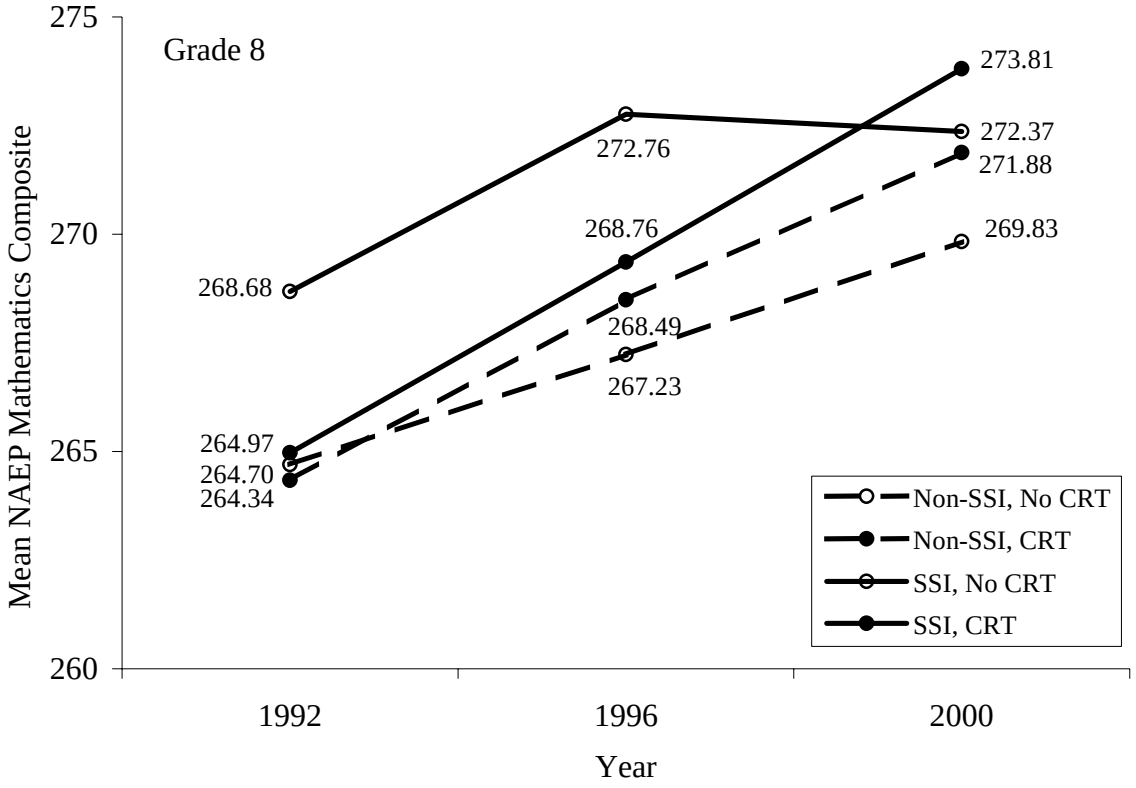
In all these models, the SES variable was strongly related to the state's mathematics achievement. At grade 8, there is some indication that instructional practices are also related to mathematics achievement, apart from their relationship with SES. In 1992, just one indicator added to the prediction of mathematics achievement. By 1996, three were included in the model.

In 1996, both regression models included I(S), and the indicator had a negative coefficient. In 1992, another indicator of teachers' knowledge of reform-related topics, I(RT), also had a negative coefficient. It is possible be that states with relatively low mean achievement had focused on professional development, but that teachers had not yet had time to make changes in their instructional practices based on their knowledge.

State NAEP Mean Mathematics Composite as a Function of SSI Status and Use of Criterion-Referenced Tests

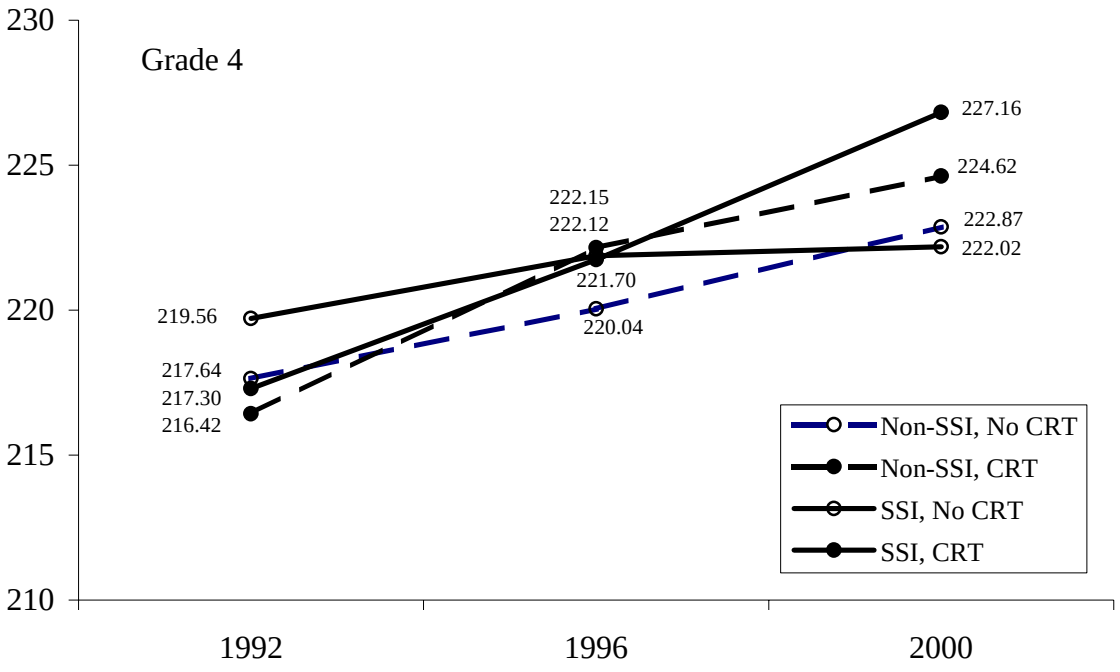
Previous sections of this chapter focused on data from State NAEP in 1992 and 1996. Information about student achievement in 2000 is also available from State NAEP. In this section, the effect of SSI status and CRT Use on student achievement is examined with a 2 x 2 repeated measures analysis of variance, adjusting for SES. Figure 5.6 shows the mean scores at grade 8 for each group, and Figure 5.7 shows the mean scores at grade 4. Year was a significant effect in both analyses, with grade 8 increasing an average of 6.30 points across the eight years ($F = 44.17$, $p < .01$), and grade 4 increasing by 6.35 points ($F = 47.52$, $p < .01$). The interaction of Year and CRT Use was also statistically significant at both grades; states using CRT tests gained significantly more than the other states. At grade 8, the eight-year gain for states with CRT tests averaged 8.19 compared to 4.41 for the other states ($F = 4.31$, $p < .05$). At grade 4, the gains averaged 8.87 points for states using criterion-referenced tests compared to 3.85 points for the other states ($F = 11.61$, $p < .01$). Finally, the three-way-interaction of Year x CRT Use x SSI Status was also statistically significant at both grade levels, with a significant quadratic component (Grade 8: $F = 4.89$, $p < .05$; Grade 4: $F = 4.29$, $p < .05$). Figures 5.6 and 5.7 show that SSI states with CRT tests continued to increased steadily from 1992 to 2000, while SSI states without CRT tests had comparable gains from 1992 to 1996 but did not sustain the rate of gain over the next four years. (Analyses without the SES covariate found the same significant effects at both grade levels.)

Figure 5.6. Grade 8 mean NAEP mathematics composite in 1992, 1996, and 2000 by SSI status and CRT use.



The slopes of the lines in Figures 5.6 illustrate the rate of gain in the mean State NAEP mathematics composite across each four-year interval. At grade 8, states using criterion-referenced tests had the greatest rate of gain. Among the states without CRTs, the states in the SSI program increased as much as the states using CRTs from 1992 to 1996. Between 1996 and 2000, when the SSI program ended, mathematics achievement did not continue to increase for this group. From 1992 to 1996, states with neither SSI nor CRTs had the lowest rate of gain, but these states were able to sustain the rate of gain from 1996 to 2000.

Figure 5.7. Grade 4 mean NAEP mathematics composite in 1992, 1996, and 2000 by SSI status and CRT use.



Results for grade 4 are similar to those for grade 8. The SSI states with CRTs had the fastest gain across the eight years (9.84 points), and the Non-SSI states with CRTs were second (6.98 points). Both of these groups were able to sustain the rate of increase across the two four-year intervals. In contrast, the SSI states without CRTs increased an average of 2.56 points from 1992 to 1996, and then leveled off from 1996 to 2000. Since achievement gain slowed when SSI funding was discontinued, the finding supports the conclusion that the SSI spurred gains in these states. Although the rate of gain did not continue once funding ending, the achievement gains were sustained, suggesting that the changes resulting from the SSI continued through the next four years.

Summary and Conclusions

By using State NAEP, we were able to compare almost all of the SSI states with other states using common measures. This chapter describes six reform-related indicators developed from the State NAEP teacher questionnaire, including three classroom practice indicators:

- I(RC), Relative Emphasis on Reasoning and Communication.
- I(MD), Student's Opportunities for Mathematical Discourse
- I(C), Students' Calculator Use

and three indicators of teachers' knowledge and professional development:

- I(S), Teachers' Knowledge of the NCTM *Standards*
- I(PD), Time in Professional Development Last Year
- I(RT), Number of Reform-Related Topics Teachers Studied

Both cross-sectional and longitudinal analyses found that SSI states averaged significantly higher than non-SSI states on I(RC), an indicator of the relative emphasis on reasoning and communication, at both grade 4 and grade 8 in 1996. Furthermore, a regression model for grade 8 found that I(RC) added to the prediction of the mean State NAEP mathematics composite along with SES. These findings support the conclusion that the SSI program promoted challenging curriculum standards for all students, particularly at grade 8.

Cross-sectional analyses also found that SSI states averaged higher than non-SSI states in I(MD), an indicator of students' opportunities for mathematical discourse, at both grades 4 and 8 in 1996, but not in 1992. In the trend sample, repeated measures analyses of I(MD4) found that SSI states averaged higher than non-SSI states across 1992 and 1996, and I(MD4) increased significantly at both grade levels.

Changes in I(C), an indicator of calculator use, were significantly related to SES at both grades 4 and 8. With the trend sample, there was a significant interaction at grade 4. Among the SSI states, the increase in calculators use was largest for SSI states without criterion-referenced tests and smallest for non-SSI states without criterion-referenced tests.

I(S), an indicator of teachers' knowledge of the NCTM *Standards*, was part of the teacher questionnaire in 1996 but not in 1992. Both SES and SSI status were related to I(S) in the yearly sample, with SSI states averaging higher than non-SSI states. Results of the trend sample supported the findings at grade 8, but at grade 4 the SSI effect interacted with the use of criterion-referenced tests, with the largest increase in SSI states without criterion-referenced tests.

The trend sample makes it possible to evaluate change in those indicators with the same scale from 1992 to 1996. On most measures, trend-sample states were not significantly different from other states that participated in State NAEP but were not part of the trend sample. The one significant difference was that the trend-sample states averaged lower in SES. Conclusions based on the trend sample may be limited to states whose mean SES is below the national average.

Repeated measures analyses found significant increases from 1992 to 1996 on all measures. These results document the extent of educational improvements from 1992 to 1996 in all states—the SSI states as well as the non-SSI states. Because State NAEP provides a comparison group, we can differentiate change tied to the SSI program from change from other national efforts.

Regression modeling found that SES was the primary predictor of a states' mean mathematics composite in both 1992 and 1996. In several models, one of the professional development indicators added to the prediction, but with a negative coefficient. This result may

indicate that states with relatively low achievement scores invested in professional development. The 1996 model for grade 8 included I(RC) in addition to the other predictors, providing some evidence that emphasizing reasoning and communication is related to mathematics achievement gains.

The high correlations of the indicators across years, and between grade 4 and 8 for the same year, indicate the consistency of the instructional practices within each state. States seem to have distinctive profiles. Documenting educational practices is essential to evaluating the effectiveness of reform efforts. Equally important, though, is careful consideration of the practices prior to the reform efforts, so current practices can be considered in light of previous practices.

In addition to the states' SSI status, the analyses reported in this chapter examined whether states used criterion-referenced tests (CRT) in at least two of the grades 3 through 8. In the cross-sectional analyses, CRT was not a significant factor. But it was significant in the trend analyses of I(C) at grade 4.

Analyses of student achievement across three State NAEP administrations showed that both SSI status and CRT were related to achievement gains across time. Gains were largest among SSI states with criterion-referenced tests. In both grade 8 and grade 4, achievement gains from 1996 to 2000 were smallest for states that had been in the SSI program but did not have criterion referenced tests.

Appendix 5-1
Type of Mathematics Assessment Below High School in 1995-96
for SSI and Non-SSI States Participating in State NAEP[a]

State	Type[b]	Grades	Test Name
Alabama	NRT	3 - 8	Stanford Achievement Test
Alaska	NRT	4, 8	California Test of Basic Skills
Arkansas	CRT	4	Criterion Referenced Test
	NRT	5, 7	Stanford Achievement Test
Arizona	NRT	4, 7	Iowa Test of Basic Skills
California	None		
Colorado	None		
Connecticut	CRT	4, 6, 8	Connecticut Mastery Test
Delaware	None		
Georgia	CRT	3, 5, 8	Curriculum-based assessments
	NRT	3, 5, 8	Iowa Test of Basic Skills
Hawaii	NRT	3, 6	Stanford Achievement Test
Idaho	NRT	3 - 8	Iowa Test of Basic Skills
Illinois	NRT & CRT	3, 6, 8	Illinois Goal Assessment Program
Indiana	NRT & CRT	3, 6	Iowa Test of Basic Skills
Iowa	None		Each district required to develop an assessment system to monitor progress toward district goals
Kentucky	Performance assessment	5, 8	Kentucky Instructional Results Information System
Louisiana	CRT	3, 5, 7	Louisiana Educational Assessment Program
	NRT	4, 6	California Test of Basic Skills
Massachusetts	CRT	4, 8	Massachusetts Educational Assessment Program
Maryland	CRT	3, 5, 8	Maryland School Performance Assessment Program
Maine	Neither	4, 8	Maine Educational Assessment
Michigan	CRT	4, 7	Michigan Educational Assessment Program
Minnesota	None		
Missouri	CRT	3, 6, 8	Missouri Mastery and Achievement Test
Mississippi	NRT	4 - 8	Iowa Test of Basic Skills
Montana	NRT	4, 8	Districts select tests from state list
Nebraska	None		
North Dakota	NRT	3, 6, 8	California Test of Basic Skills
New Hampshire	CRT	3, 6	New Hampshire Educational Improvement and Assessment Program
New Jersey	CRT	8	Grade 8 Early Warning Test[c]
New Mexico	NRT	3, 5, 8	New Mexico Achievement Assessment
New York	CRT	3, 6	Pupil Evaluation Program Tests
Ohio	CRT	4, 6	Fourth and Sixth Grade Proficiency Testing

Oklahoma	CRT	5, 6, 8	Oklahoma Core Curriculum Tests
	NRT	3, 7	Iowa Test of Basic Skills
Oregon	CRT	3, 5, 8	Oregon Mathematics Assessment
Pennsylvania	NRT	5, 8	Pennsylvania Mathematics Assessment
South Carolina	CRT	3, 8	Basic Skills Assessment Program
	NRT	4, 5, 7	Metropolitan Achievement Tests
Tennessee	CRT	2 - 8	TCAP Achievement Test – CRT
	NRT	2 - 8	TCAP Achievement Test – NRT
Texas	CRT	3 - 8	Texas Assessment of Academic Skills
Utah	CRT	1 - 8	Core Curriculum Assessment Program
	NRT	5, 8	Stanford Acheivement Test
Vermont	CRT	4, 8	New Standards Math
Washington	NRT	4, 8	California Test of Basic Skills
West Virginia	CRT	1 - 8	West Virginia - STEP
	NRT	3, 6	California Test of Basic Skills
Wisconsin	NRT	4, 8	Wisconsin Student Assessment System
Wyoming	None		

[a]Information from Roeber, Bond, and Braskamp, 1997, Parts 1.01, 3.10, and 3.14A,B.
[b]NRT stands for norm-referenced test; CRT stands for criterion-referenced test.
[c]New Jersey was included in the non-CRT group because only one grade below high school was tested with a criterion-referenced test.

CHAPTER 6

USING SUPPLEMENTARY INFORMATION IN EVALUATING STATEWIDE SYSTEMIC REFORM[1]

The State Assessment Program of the National Assessment of Educational Progress (State NAEP) provides common measures of student achievement for each participating state. State NAEP was first administered in 1990 at grade 9 and was expanded to include grade 4 in 1992. Although the SSI program addressed both mathematics and science education, only mathematics achievement is addressed in this study.

The 14 SSI states in the trend sample are the focus of this chapter. Among these states, some had substantial increases in their NAEP mathematics composite from 1992 through 2000, while others had relatively small increases. Through the use of supplementary information from a variety of sources, hypotheses about the features of systemic reform associated with relatively large statewide gains in mathematics achievement are developed and refined.

For each state, grade 4 and grade 8 state means for the mathematics composite score were examined across 1992, 1996, and 2000. Based on mean achievement gains, the states were categorized into three groups. (See Table 6.1.) A gain of 3.5 points was used as the reference point because the national gain averaged between 3 and 4 points for each four-year interval. The groups were:

Steady Increase: The mean mathematics composite increased by more than 3.5 points from 1992 to 1996 and from 1996 to 2000 at grade 4 and/or grade 8.

Some Increase: The mean mathematics composite increased by more than 3.5 points at both grade levels for one of the four-year intervals.

Little/No Change: The mean mathematics score increased by less than 3.5 points for both four-year intervals at grade 4 and/or grade 8.

Using a variety of information sources, this chapter addresses three questions about the Statewide Systemic Initiatives:

1. What features differentiate SSI states with steady increases in mathematics achievement at grades 4 and/or 8 from the SSI states with little or no increases? Are these features shared by SSI states with some increases in mathematics achievement?
2. Can statewide achievement gains be attributed to a state's SSI?
3. When is statewide achievement gain an appropriate measure of systemic reform efforts?

[1] This chapter is based on an earlier paper, Using Information from National Databases to Evaluate Systemic Educational Reform, presented at the annual meeting of the American Education Association, Arlington, VA, November, 2002.

Table 6.1
State Groups Based on Mean NAEP Mathematics Composite Gain

	Gain from 1992 to 1996		Gain from 1996 to 2000		Total Gain	
	Grade 4	Grade 8	Grade 4	Grade 8	Grade 4	Grade 8
Steady Increase						
Texas*	10.79	5.61	3.96	4.65	14.75	10.26
New York	4.18	3.81	3.93	6.03	8.11	9.84
Michigan	6.38	9.52	4.63	1.58	11.01	11.10
Kentucky	4.94	4.35	1.00	4.97	5.94	9.32
Massachusetts*	2.37	4.79	5.99	5.55	8.36	10.34
Louisiana*	4.88	2.40	8.94	6.60	13.82	9.00
Some Increase						
South Carolina*	0.69	0.01	7.23	5.57	7.92	5.58
Georgia	–0.13	3.11	4.10	3.86	3.97	6.97
Connecticut*	5.23	5.85	2.21	2.41	7.44	8.16
Arkansas	5.64	5.34	1.21	0.71	6.85	5.05
Little/No Increase						
Maine	0.57	6.62	–1.64	–0.42	–1.07	5.00
California	0.73	1.88	4.44	–0.60	5.17	1.28
Nebraska	2.21	5.12	–1.59	–2.15	0.62	2.97
New Mexico	0.54	2.36	0.03	–2.13	0.57	0.24

*Phase II states

Information Sources

Several sources of information were used to compare and contrast the three groups of SSI states. The sources are described in the following paragraphs.

State NAEP—Achievement data. The State NAEP mathematics achievement test provides a statewide mathematics composite, as well as five subscores. The NAEP Mathematics Framework includes five content strands: 1) Number Sense, Properties, and Operations, 2) Measurement; 3) Geometry and Spatial Sense; 4) Data Analysis, Statistics, and Probability; and, 5) Algebra and Functions. Calculator use is permitted on approximately one-third of the test questions. Results reported here are based on data from public school students under conditions that did not offer accommodations to special needs students.

State NAEP—Reform indicators. State NAEP also includes a teacher questionnaire, with items about teachers' preparation and instructional practices. The State NAEP teacher questionnaire included a wide variety of questions prior to 2000, but the number of questions was kept to a minimum in 2000.

Items from the teacher questionnaire were used to create six indicators of reform-related practices, as described in the previous chapter (Webb, Kane, Kaufman, & Yang, 2001, pp. 107-237). The indicators are listed below.

I(RC), Relative Emphasis on Reasoning and Communication—extent to which reasoning and communication were addressed, relative to facts and procedures.

I(MD), Mathematical Discourse—a scale of students' opportunities to discuss, present, and write about mathematical ideas.

I(C), Calculator Use—a scale of the extent to which students used calculators in the classroom and on tests.

***I(S), NCTM* Standards**—a single item that asked about teachers' knowledge of the NCTM *Standards*.

I(PD), Last Year's Professional Development—a single item that asked how much time teachers spent in professional development in mathematics or mathematics education during the last year.

I(RT), Reform-Related Topics Studied—a count of the number of reform-related topics teachers have studied out of the seven topics listed in the NAEP questionnaire.

The first three indicators describe teachers' classroom practices, while the others ask about what teachers know, how much time they spent in mathematics staff development during the last year, and which topics they have studied.

Data from all states participating in the State NAEP in a given year were used to standardize the indicators to a scale with a mean of 0 and a standard deviation of 1. For each state, comparisons between 1992 and 1996 were used to identify relative changes in the indicators. Comparisons with 2000 were not possible because the items were not administered to teachers.

State reports. Project staff compiled reports on many of the SSI states, based on interviews with SSI leaders and documents about state reform efforts. State reports were completed in 12 of the 14 states in the trend sample, in all but South Carolina and New Mexico. The reports focused on reform efforts from 1990 through 1996 and include background on SSI and non-SSI reforms, the target population of the SSI, saturation, form and systemicness, and the nature of mathematics. Specifics of the interview protocol are summarized below:

Saturation: SSI leaders were asked to estimate the percentage of teachers and/or students who were reached by the reform in 1990, 1992, and 1996.

Form and Systemicness: SSI leaders were asked to rate the emphasis given to four components of reform: Policy, curriculum, instruction, and incentives. Ratings were made for 1990, 1992, and 1996, using a scale from 1 (Low) to 5 (High).

Nature of Mathematics: SSI leaders reported the relative emphasis on each of the five content topics, using 100% for the total. When available, state tests

> and/or state standards were examined and the percent of items or standards for each of the five content topics was computed.

Analyses of state SSIs. Clune and his colleagues proposed a model of systemic reform, as follows (Clune, 1998; Clune, Osthoff, & White, 2000):

> Systemic reform, through purposeful activities, leads to systemic policy which leads to a rigorous implemented curriculum for all students, which leads to measured high student achievement in the curriculum as taught. (Clune, 1998, p. 2)

Based on this model, the researchers developed an analytic approach to describe the breadth and depth of the components, both pre-SSI and near the end of Phase I funding. To date, 16 SSI states have been examined, including 11 of the 14 in the trend sample (Osthoff, 2002). The three missing states are South Carolina, Nebraska, and New Mexico.

Evaluations of the SSI program. The National Science Foundation commissioned evaluations of the SSI program as a whole, as well as of specific program components (e.g., Zucker, Shields, Adelman, Corcoran, & Goertz, 1998; Corcoran, Shields, & Zucker, 1998; LaGuarda, Goldstein, Adelman, & Zucker, 1998).

Annual surveys of state student assessment programs. The Council of Chief State School Officers (CCSSO) provides extensive data on statewide student assessment programs, based on surveys mailed to states each Fall for the prior year's program. In this study, information from CCSSO's annual surveys of state assessments for 1995–96 and 1999–2000 was used.

Findings

This section presents selected characteristics of the SSI state data from the sources listed above. The data sources provided a wide variety of information about each state. The information in the tables was selected to illustrate differences between the three groups of SSI states in the trend sample.

SSI leaders' ratings on form and systemicness. Table 6.2 presents selected ratings from state SSI leaders. They were asked to rate the effort directed to each of four reform components (policy, curriculum, instruction, and accountability), using a scale from 1 to 5, with 1 representing Low Effort and 5 High Effort. The ratings for Policy and Instruction are reported in Table 6.2. In rating Instruction, SSI leaders were asked to consider both direct services to schools and teachers as well as the creation of an infrastructure for capacity building.

The first two columns in Table 6.2 are ratings for 1990, prior to the start of the SSI. The next two columns are 1996 ratings, well into the SSI. Differences between the

Table 6.2

Ratings by State SSI Leaders of the Relative Effort Directed to Selected Components of Systemic Reform—Scale of 1 to 5, with Anchor Points of 1 (Low Effort) and 5 (High Effort)

	1990		1996		Change		Difference
	Policy	Instruction	Policy	Instruction	Policy	Instruction	in 1996 Ratings
Steady Increase							
Texas	5.0	3.0	4.5	4.0	–0.5	1.0	0.5
New York	NA	NA	NA	4.0	NA	NA	NA
Michigan	4.0	4.0	4.0	4.0	0.0	0.0	0.0
Kentucky	4.0	4.0	5.0	5.0	1.0	1.0	0.0
Massachusetts	1.0	1.0	4.0	4.0	3.0	3.0	0.0
Louisiana	1.0	1.0	3.5	3.0	2.5	2.0	0.5
Some Increase							
South Carolina	*Ratings not available*						
Georgia	1.0	1.0	4.0	4.0	3.0	3.0	0.0
Connecticut	2.5	1.5	3.5	3.0	1.0	1.5	0.0
Arkansas	2.0	1.0	5.0	5.0	3.0	4.0	0.0
Little/No Change							
Maine	1.0	2.0	4.0	5.0	3.0	3.0	-1.0
California	2.5	1.0	1.5	2.5	–1.0	1.5	-1.0
Nebraska	NA	NA	3.0	5.0	NA	NA	-2.0
New Mexico	*Ratings not available*						

Note: Values are comparable within site but not across sites, since raters were not trained to the same standard.
NA – Not available

two sets of ratings are listed in the next two columns. A value of 0.0 indicates that the ratings did not change from 1990 to 1996, values greater than 0.0 means the emphasis increased during the SSI, and values less than 0.0 indicate that the emphasis decreased. The last column provides information about how the emphases on Policy and Instruction compared in 1996. A value of 0.0 indicates that Policy and Instruction were emphasized equally, while values greater than 0.0 mean that Policy was emphasized more than Instruction, and values less than 0.0 mean Instruction was emphasized more than Policy.

SSI leaders were not trained to the same standard, so ratings across states are not comparable, but ratings within a state indicate where resources were directed. The 1990 ratings indicate the extent of systemic reform efforts prior to SSI funding, and the 1996 ratings indicate efforts in the third, fourth, or fifth year of the program, depending on whether the state joined the SSI program in 1991, 1992, or 1993.

As shown in the first two columns of Table 6.2, three of the 14 states in the trend sample had relatively high ratings in 1990, prior to the start of the SSI and continued to be high in 1996. All three had steady increases in students' mathematics achievement from 1992 to 2000.

The last column of Table 6.2 shows that states with little or no statewide change in mathematics achievement had SSIs that put greater emphasis on instruction rather than policy in 1996.

Ratings of components in a theoretical model of systemic reform. Clune (Clune, 1998) and his associates have developed criteria and benchmarks for describing systemic reform (Clune, Osthoff, & White, 2000). The model has two Pre-SSI components, Policy and Infrastructure, along with four SSI components, Leadership and Change Strategy, Policy, Infrastructure, and Standards-Based Instructional Reform. Detailed descriptions of the components are available in Clune (1998). Policy consists of five areas: Standards/Frameworks, Professional Development, Assessment, Accountability, and Other Instructional Guidance Policies; and Infrastructure has one: Networks. Standards-Based Instructional Reform includes three categories, Individual Capacity Building, Organizational Capacity Building, and Classroom Practice. The Instructional Reform component indicates the extent to which teachers, classrooms, and schools implement standards-based instruction. All components are rated for both Breadth and Depth. Table 6.3 presents mean ratings for Policy and Infrastructure both Pre-SSI and in 1997, along with mean ratings for Instructional Reform. The breadth and depth ratings were averaged to get the values in Table 6.3. The values used were obtained through personal communication (Osthoff, 2002).

Table 6.3 also lists changes in Policy and Infrastructure ratings from Pre-SSI to 1997; values above 0.0 indicate that the ratings increased, and values below 0.0 indicate the ratings decreased. Finally, the last column in the table compares the 1997 ratings for Policy and Standards-Based Instructional Reform. The value is comparable to the last column of Table 6.2.

Table 6.3
Mean Ratings of Selected Components from a Model of Systemic Reform

	Pre-SSI Ratings		1997 Ratings[a]			Change during SSI		Difference Between Policy & Instructional Reform (1997)
	Policy	Infra-structure	Policy	Infra-structure	Instructional Reform	Policy	Infra-structure	
Steady Increase								
Texas	3.30	3.00	3.80	3.50	2.50	0.50	0.50	1.30
New York	3.00	2.50	3.40	2.50	1.50	0.40	0.00	1.90
Michigan	2.70	3.00	3.10	3.50	2.33	0.40	0.50	0.77
Kentucky	3.60	3.50	3.90	3.50	2.83	0.30	0.00	1.07
Massachusetts	1.70	2.00	3.40	3.50	2.67	1.70	1.50	0.73
Louisiana	2.10	2.00	3.10	4.00	3.50	1.00	2.00	-0.40
Some Increase								
South Carolina	- *Ratings not available* -							
Georgia	2.90	2.00	3.40	3.00	2.50	0.50	1.00	0.90
Connecticut	2.40	4.00	3.20	4.00	3.50	0.80	0.00	-0.30
Arkansas	2.40	2.00	3.30	3.50	2.83	0.90	1.50	0.47
Little/No Change								
Maine	2.10	2.50	3.10	4.00	4.00	1.00	1.50	-0.90
California	3.20	3.00	2.70	2.50	3.85	-0.50	-0.50	-1.15
Nebraska	- *Ratings not available* -							
New Mexico	- *Ratings not available* -							

[a]Interviews were conducted in Spring, 1997. Cohort I had completed five years of Phase I funding, Cohort II was completing the fifth year and Cohort III was in the fourth year.

Because ratings were tied to criteria and benchmarks, comparisons across states were possible. Four of the five states with the highest Pre-SSI ratings all had steady increases in their mean NAEP mathematics composite. For these four states, Policy and/or Infrastructure ratings increased during the SSI. The one exception was California, the only state where ratings declined during the SSI program.

State assessment programs. Tables 6.4a and 6.4b list selected characteristics of state mathematics assessment programs in 1996 and 2000, coinciding with the years the State NAEP in which mathematics was administered. The information was compiled from the annual surveys of CCSSO (Council of Chief State School Officers, 2001; Roeber, Bond, & Braskamp, 1996).

The first column indicates whether the assessment was criterion-referenced or norm-referenced. The terms are defined in the glossary (Council of Chief State School Officers, 2001, p. 33-35).

Criterion-referenced test An assessment on which the student's performance is compared to a standard or objective, and the score indicates the extent to which the student achieved the standard or set of objectives.

Norm-referenced test A test on which a student's score is compared to the performance of a norm group, and the score indicates the proportion of students in the norm group that the student outscored.

In 1996, Kentucky's state mathematics assessment was based on open-response items, performance events, and portfolios. Maine's innovative state assessment was comprised primarily of open-ended items, with the tests containing different sets of items selected via matrix sampling to provide for broad content coverage.

The second column lists the grades in which the mathematics assessment was administered. For states with more than one type of assessment, the grades for each type are listed. The third column lists the time of year during which the statewide assessments were administered. Almost all states had a spring testing, except for Connecticut, where tests were administered in the Fall.

The fourth column lists consequences that were linked to the results of the state assessments. The list includes negative consequences for schools, because these seemed associated with achievement gains, at least in 1996. The last column reports on one kind of consequence for students—whether state policy required students to pass a graduation test to receive a high school diploma.

Descriptions of state assessment programs from 1996–2000. Table 6.5 provides a brief summary of the statewide assessment programs in mathematics for the 14 states in the trend sample, to complement the information in Table 6.4a and 6.4b. Major changes to the assessment between 1995 and 2000 are noted.

Table 6.4a
Characteristics of State Assessment Programs in Mathematics and Accountability Policies, 1996

	Type Of Test	Grades Tested	Time of Testing	Negative Consequences for Schools	High School Graduation Test
Steady Increase					
Texas	CRT	3–8;10–12	Oct, Spring	Wrn,PWL,TO,Dis	Yes
New York	CRT	3,6	Spring	Wrn,PWL,TO	Yes
Michigan	CRT	4,8,11	Sept, Oct, Mar	Wrn,PWL,TO,Dis	No
Kentucky	PA	5,8,11	Spring	Wrn,PWL,TO,Dis	No
Massachusetts	CRT	4,8	Spring	None	No
Louisiana	CRT/NRT	3,5,7/4,6	Spring/Spring	None	Yes
Some Increase					
South Carolina	CRT/NRT	3,6,8,10/4,5,7,9,11	Spring/Spring	TO	Yes
Georgia	CRT–MS/NRT	3,5,8,11/3,5,8,11	Spring/Spring	None	Yes
Connecticut	CRT	4,6,8,10	Fall	None	No
Arkansas	NRT/CRT	5,8,11/4	Fall/Spring	None	No
Little/No Change					
Maine	MS	4,8,11	Spring	None	No
California	- - - - - - - - - *Statewide assessment program limited to end-of-course high school exams*- - - - - - - - -				
Nebraska	- *No statewide assessment program* - - - - - - - - - - - - - - - - - -				
New Mexico	NRT/CRT	3,5,8/10	Spring	None	Yes

Type of Test:
- CRT – Criterion-referenced test
- NRT – Norm-referenced test
- MS – Matrix-sampled test
- PA – Performance assessment

Possible Negative Consequences for Schools:
- Wrn – Give warnings to schools
- PWL – Put on probation or watch list
- TO – Take over schools
- Dis – Dissolve schools

Table 6.4b
Selected Characteristics of State Assessment Programs in Mathematics and Accountability Policies, 2000

	Type Of Test	Grades Tested	Time of Testing	Negative Consequences for Schools	High School Graduation Test
Steady Increase					
Texas	CRT	3–8;10–12	Various	Wrn,PWL,TO,Dis	Yes
New York	CRT	4,8	Spring	Wrn,PWL	Yes
Michigan	CRT	4,7	Spring	Wrn,PWL	No
Kentucky	CRT/NRT	5,8,11/3,6,9	Spring/Spring	None	No
Massachusetts	CRT	4,8,10	Spring	None	No
Louisiana	CRT/NRT	4,8/3,5,6,7,9	Spring/Spring	None	Yes
Some Increase					
South Carolina	CRT/NRT	3-8/5,8,11	Spring/Spring	Wrn,PWL,TO[a]	Yes
Georgia	CRT/NRT	4,6,8/3,5,8	Spring/Spring	Wrn,PWL[a]	Yes
Connecticut	CRT	4,6,8,10	Fall	Wrn,PWL	No
Arkansas	NRT/CRT	5,7,10/4,8	Fall/Spring	Wrn,PWL[b]	No
Little/No Change					
Maine	CRT–MS	4,8,11	Fall,Spring	None	No
California	NRT	2–11	Spring	None	No
Nebraska	*No statewide assessment program*				
New Mexico	NRT	3–9	Spring	PWL	Yes

[a]Consequences based on the results of the criterion-referenced tests.
[b]Consequences based on the results of the norm-referenced tests.

Type of Test
- CRT – Criterion-referenced test
- NRT – Norm-referenced test
- MS – Matrix-sampled test

Possible Negative Consequences for Schools:
- Wrn – Give warnings to schools
- PWL – Put on probation or watch list
- TO – Take over schools
- Dis – Dissolve schools

Table 6.5
Brief Descriptions of Selected State Assessment Programs at the End of the 1990s

Steady Increase States

- New York's state testing program is the oldest in the nation, first administered in 1865. Tests are based on the learning standards, and results provide a level of accountability for state schools. In 1998–99, the Board of Regents adopted new standards and approved the development of new tests based on the standards.
- The Texas Assessment of Academic Skills (TAAS), a criterion-referenced program that assesses mathematics in grades 3–8, began in 1990. New standards, Texas Essential Knowledge and Skills (TEKS), were adopted in 1997 and changes were made to TAAS so it would be aligned with TEKS by 1999-2000. Legislation in 1999 mandated a new testing program for 2002–2003.
- In Michigan, the next generation of Michigan Educational Assessment Program (MEAP) tests was under development, based on the new curriculum content standards approved in 1995.
- In Kentucky, the Kentucky Instructional Results Information System (KIRIS) was replaced with the Commonwealth Accountability Testing System (CATS), required by legislation passed in 1998.
- The Massachusetts Assessment Program (MAP) was first administered in 1986 and then in 1990, 1992, 1994, and 1996. A new state assessment system was authorized by legislation in 1993, and tests based on the new curriculum frameworks were first implemented in 1997–98.
- In 1996, the Louisiana Assessment Program had criterion-referenced tests in grades 3, 5, and 7 and norm-referenced tests at grades 4 and 6. New content standards were adopted, and criterion-referenced tests based on the mathematics standards were implemented at grades 4 and 8 in 1998–99, along with norm-referenced tests at grades 3, 5, 6, 7, and 9.

Some Increase States

- Connecticut had criterion-referenced tests, with implementation of the third generation of tests scheduled for 2000–2001. In Connecticut, tests are administered in the Fall.
- Arkansas, which had a norm-referenced test in 1996, expanded its assessment program to include criterion-referenced tests, along with the norm-referenced tests, in 2000.
- Georgia administered both criterion- and norm-referenced tests in 1996, but moved to norm-referenced tests in 2000.
- South Carolina had criterion-referenced tests in grades 3, 8, 10, and 11 and norm-referenced tests at the other grades in 1996. By 2000, criterion-referenced testing was expanded to grades 3 through 8 and 10 to 12, and norm-referenced testing of a sample of students was continued at grades 5, 8, and 11.

Little/No Change States

- California's criterion-referenced testing program, the California Learning Assessment Program (CLAS), was discontinued in 1994–95 as a result of the

governor's veto. Local districts were encouraged to select their own standardized tests. Legislation in 1995 and 1996 required development of new standards in the major subject areas and a statewide pupil assessment program. By 2000, California was using the Stanford Achievement Test, Ninth Edition, in grades 2 to 11, supplemented with standards-based test items.

New Mexico administered the Iowa Test of Basic Skills to students in grades 3, 5, and 8 in 1996. In 2000, the state used a different norm-referenced standardized test, the California Test of Basic Skills 5/TerraNova, supplemented with items linked to state standards, in grades 3 through 9.

Nebraska had no statewide testing program. Local districts were required to select a test for their reporting requirements.

The Maine Educational Assessment (MEA) began in 1985 and was redesigned in 1998–99, based on the new state standards. The test includes both common and matrix-sampled items.

State mathematics standards/frameworks and assessments and SSI goals. Table 6.6 presents information about the alignment of the SSI goals and the state assessment, as reported in Laguarda et al., (1994) and Zucker et al., (1998, p. 24), as well as some background information on the state mathematics framework (Zucker et al., 1998, p. 26). The State NAEP includes items from five content areas: Number and Operations; Measurement; Geometry; Data Analysis, Statistics, and Probability; and, Algebra and Functions, with roughly 30 to 40% of the items from Number and Operations depending on the grade level and year. The last two columns of Table 6.6 list the percentage of state assessment items that can be classified as Number and Operations items, providing an indication of the breadth of content coverage on the state assessment.

Research Questions

Features That Differentiate Steady Increase States from Other SSI States

Several characteristics were identified that were associated with the groups of SSI states in the trend sample. The evidence related to each characteristic is reviewed in the next five sections.

1. State-level policy. In their overall evaluation of the SSI program, Zucker et al., (1998, pp. 9–16) identified eight strategies used by the SSIs and classified them in terms of their focus:

Strategies focused on teachers, classrooms, and schools:

1. Supporting teacher development
2. Developing, disseminating, or adopting instructional materials
3. Supporting model schools

Strategies focused on districts, regions, and states

4. Aligning state policy
5. Creating an infrastructure for capacity building

Table 6.6
Alignment of SSI Goals, State Assessments, and State Standards

	Alignment of State Test And SSI Goals		Year 1997 Framework Was Adopted	Were SSI funds used to develop the framework?	1996 Emphasis on State Assessment for Number & Operations	
	1994	1997			Grade 4	Grade 8
Steady Increase						
Texas	No	No	Revised, 1997	Yes	66	56
New York	No	Developing	1994	No	62[1]	58[1]
Michigan	Yes	Yes	Revised, 1995	No	60	47
Kentucky	Yes	Yes	1993	No	31[2]	33[2]
Massachusetts	Yes	Yes	1994	Yes	42	48
Louisiana	No	No	1996	Yes	63[3]	60[3]
Some Increase						
South Carolina	No	Developing	1993	Yes	NA	NA
Georgia	No	Revising	Revising, 1997	Yes	NA	63
Connecticut	No	Yes	Revising, 1997	No	55	57
Arkansas	No	No	1993	No	20[4]	20[4]
Little/No Change						
Maine	Yes	Yes	1996	No	45	30
California	Yes[a]	Developing	Revising, 1997	No	NA	NA
Nebraska	NA	NA	1994	No	NA	NA
New Mexico	No	No	1997	Yes	NA	NA

[a]The rating applied to the CALS program, which was discontinued after one year.
[1]The grade 3 and grade 6 New York state mathematics tests were analyzed in 1996.
[2]Based on categorization of the released KIRIS items (open response from 1995–1996 and multiple choice for 1996–1997).
[3]The grade 5 and grade 7 tests described in *Teachers' Guide to LEAP Tests* were used.
[4]Based on the state mathematics framework rather than an analysis of the assessment.

NA – Not available

6. Funding local systemic initiatives
7. Reforming higher education and the preparation of teachers
8. Mobilizing public and professional opinion

The authors emphasized that no strategy was used in isolation and that SSIs typically employed four or more. However, some SSIs primarily focused on strategies close to the classroom, others focused on state system and infrastructure, and others did both (balanced focus). Table 6.7 below cross-tabulates the SSI focus with the statewide groupings based on achievement gains.

Table 6.7
Focus of State SSI and Statewide Achievement Gains from 1992 to 2000

	State System and Infrastructure	Balanced	Close to the Classroom
Steady Increase	Michigan Texas	Louisiana Massachusetts	Kentucky New York
Some Increase	Connecticut Georgia South Carolina	Arkansas	
Little/No Change		Maine	California Nebraska New Mexico

As Table 6.7 indicates, all of the states with a focus on "State System and Infrastructure" had some or a steady increase in mathematics achievement, and all but one of the states with a "Balanced" focus also increased in mathematics achievement. Of the five states with a focus "Close to the Classroom," only two showed gains.

Data from other sources also support the association between statewide policies and statewide achievement gains. In states with some or steady statewide achievement gains, SSI leaders rated the emphasis on state policy as high or higher than the emphasis on instruction (Table 6.2, last column). By contrast, in the three states with little or no gains, SSI leaders reported more emphasis on instruction than on policy.

Table 6.3 provides additional support for the association between statewide policy and statewide achievement gain. In Clune's model of systemic reform (1998), ratings for statewide infrastructure are separate from rating for standards-based classroom practices

(i.e., Instructional Reform). The two states with little or no change in mathematics achievement had the highest ratings for Instructional Reform but the lowest ratings for Policy. Although reform strategies in these states influenced many individual teachers, schools, and classrooms, the effect was not evident in statewide achievement gains.

Unlike the ratings in Table 6.2, those in Table 6.3 can be compared across states because they are based on a common set of criteria and benchmarks. Four of the six states with steady increases had relatively high ratings for Policy and Infrastructure prior to the SSI. The other two states with steady increases, Massachusetts and Louisiana, had some of the lowest pre-SSI ratings, but had among the largest increases during the SSI.

In most states, both Policy and Infrastructure ratings increased together. It seems that in most states, these two components were coordinated. Three states were exceptions: New York, Kentucky, and Connecticut. In these states, Policy ratings increased, but Infrastructure ratings did not change, perhaps indicating that in these states the components of systemic reform were not closely connected.

Ratings for Policy and Infrastructure in California actually decreased during the SSI. This result is a reminder that systemic reform is not necessarily self-sustaining.

In summary, information from three different sources supports the conclusion that statewide achievement gains across four years are more likely to be evident when reform efforts address state policy as much as or more than teachers and classroom practices. When reform efforts primarily focus closer to the classroom, four years may not be adequate to influence enough teachers and students to result in statewide achievement gains. Moreover, without state policies to support instructional reform, teachers may be less likely to substantially change their classroom practices.

2. Tests tied to standards. Proponents of state testing programs expect that assessment will have a substantial influence on increasing student achievement. All of the states showing some gain or steady gains in mathematics achievement had testing programs in place in 1996, while only two of the other states had testing programs at grades 4 and/or 8. (See Table 6.4a.)

Table 6.4a also shows that, in 1996, five of the six states with steady increases in mathematics achievement used criterion-referenced tests either exclusively, or more than, norm-referenced tests. Because criterion-referenced items are designed to be mapped back to the knowledge and skills assessed, test results can be used to identify areas of strength and weaknesses for individual students as well as for an instructional program. With norm-referenced tests, results for individuals or groups are reported in terms of a reference group and are less informative as a basis for remediation or program improvement.

For the states with some increase in mathematics achievement, three of the four used a mix of norm-referenced and other kinds of tests. The one exception, Connecticut, used only criterion-referenced tests, and Connecticut had the largest overall gains in this

group. Unlike the states with steady increases, however, Connecticut administered its assessments in the Fall rather than the Spring, emphasizing their use for instructional planning.

Besides criterion- and norm-referenced tests, some states used matrix sampling to assess a broad set of items. With matrix sampling, different students receive different subsets of items. Comparing individuals is challenging, since each student completes a unique set of items. Results for schools or districts can provide information on aggregated student performance on extensive sets of content. Georgia used matrix sampling on its 1996 state assessment and changed to criterion-referenced tests by 2000. Maine used matrix sampling in both 1996 and 2000. The states using matrix-sampled tests generally had below-average achievement gains, except for grade 8 in Maine from 1992 to 1996.

Comparisons between Table 6.4a and 6.4b, along with the details in 6.6, provide information about how state testing programs changed from 1996 to 2000. States with steady achievement increases generally continued their criterion-referenced testing programs, though some changed the grades that were assessed.

- New York, for example, assessed mathematics at grades 3 and 6 in 1996 and changed to grade 8 by 2000. As Table 6.1 shows, grade 8 gains in New York were 3.81 points from 1992 to 1996, compared to 6.03 points from 1996 to 2000. At grade 4, the gain was about the same for both four-year intervals.
- In Louisiana in 1996, criterion-referenced tests were administered in grades 3, 5, and 7; by 2000, they were used in grades 4 and 8. Gains from 1996 to 2000 were 8.94 in grade 4 and 6.60 in grade 8—much larger than the prior gains of 4.88 and 2.40, respectively.
- Results from Michigan indicate that gains decreased when testing at a specific grade level was discontinued. In 1996, Michigan's statewide mathematics assessment was administered at grade 8: by 2000, the assessment had moved to grade 7. From 1992 to 1996, grade 8 gains were 9.52; from 1996 to 2000, they were 1.58.

Among the states with some increase in mathematics achievement, the link between grade tested and gains was not as strong as in the states with steady increases. Perhaps the effect is stronger with criterion-referenced tests than with the norm-referenced tests. South Carolina, Georgia, and Arkansas all used a combination of norm-referenced and criterion-referenced tests from 1996 to 2000. However, South Carolina and Georgia put more weight on the results of the criterion-referenced tests by using them to determine consequences for schools, while Arkansas emphasized the results of the norm-referenced tests.

In summary, statewide assessment policies and practices seem to be important components of systemic reform. The existence of a state assessment program seems to be related to statewide achievement gains, particularly when criterion-referenced tests are used.

3. State assessments aligned with SSI goals. Laguarda et al. (1994) and Zucker et al. (1998) both report on the alignment of state assessments and the SSI goals, as summarized in Table 6.6. Only four of the 14 states in the trend sample had aligned assessments in both 1994 and 1997, and three of the four—i.e., Michigan, Kentucky, and Massachusetts—had steady increases on the State NAEP. The one exception was Maine.

Content area coverage on the state assessment is an important aspect of alignment. In analyses of state reform efforts, items from each state's 1996 assessments were categorized by the five NAEP content strands. Tests aligned with the SSI goals would be expected to have a substantial number of items from all content strands rather than a concentration in one or two of the strands. The last two columns on Table 6.6 list the percentage of items from the Number and Operations area for grades 4 and 8. At grade 8, all states with aligned assessments had fewer than 50% of items from Number and Operations.

Three of the states among the states with steady increases had assessments that were not aligned with SSI goals—i.e., Texas, New York, and Louisiana. In Texas and Louisiana, the SSI was directly involved in ongoing work to develop the state's mathematics framework as a foundation for a new generation of assessments. Only New York had assessments and frameworks that were unrelated to the SSI goals.

Among the states with some increase in student achievement, both South Carolina and Georgia had larger gains from 1996 to 2000 than from 1992 to 1996. In these states, the SSI played a role in developing the mathematics framework and, by 1997, the state assessments were under revision as a result of that effort. In Connecticut and Arkansas, the SSI did not address the alignment of the state assessment and SSI goals. Both states had strong gains from 1992 to 1996, but they were not able to sustain the rate of gain over the next four years.

These results illustrate the challenges faced by statewide SSI programs. When state policies are not supportive of SSI goals, reform efforts may be compromised or even undermined.

4. Accountability policies. State accountability policies generally accompany state assessments. A wide variety of features of these policies were reviewed, using annual surveys of assessment practices by CCSSO (Roeber, E., Bond, L., & Braskamp, D. 1996; Council of Chief State School Officers, 2001). Few features were useful in differentiating states with steady increases from the others. Some features were common to almost all states, such as using test results for performance reporting; others were shared by only a few states, such as using test results for promotion decisions (see Roeber, Bond, & Braskamp, 1996, p. 21; Council of Chief State School Officers, 2001, pp. 189-224).

Tables 6.4a and 6.4b include information on whether test results could lead to negative consequences for schools and whether students had to pass a test in order to graduate from high school. These features were shared by some, but not most, of the states.

In 1996, accountability policies with high stakes for schools seemed to be associated with a steady increase in achievement, as shown in Table 6.4a. In four of the states with steady increases, accountability policies produced several negative consequences based on assessment results. Only one of the other states in the trend sample used assessment results that had negative consequences for schools in 1996.

By 2000, Texas had retained all of the negative consequences for schools, New York and Michigan had reduced them a bit, and Kentucky had eliminated them. In addition, Massachusetts and Louisiana made steady gains without negative consequences for schools. Among the four states with some increase, all added negative consequences for schools between 1996 and 2000, but only South Carolina and Georgia had above-average achievement gains. The association between negative consequences and achievement gains was less clear in 2000 than in 1996.

States having a high-stakes graduation test also did not differentiate between the achievement gain groupings. In each group, roughly half of the states had a high-stakes graduation test. Comparing Tables 6.4a and 6.4b with Table 6.6 shows that in no states with high-stakes graduation tests were state tests aligned with SSI goals.

The general conclusion derived from these findings is that there is no clear relationship between specific accountability policies and statewide achievement gains in mathematics for the 14 SSI states in the trend sample.

5. Statewide change in classroom practices. Tables 6.8a and 6.8b present comparisons of selected reform-related indicators developed from the State NAEP teacher questionnaires in 1992 and 1996. Comparisons with 2000 are not possible because of the shortened teacher questionnaire in the State NAEP. The tables include:

I(RC), a measure of the extent to which teachers emphasize reasoning and communication relative to facts and procedures;
I(MD), a measure of students' opportunities for mathematical discourse, such as talking with other students, making class presentations, or writing about a solution; and,
I(PD). a measure of the amount of time teachers spent in professional development during the last year.

For each year, measures were standardized with a mean of 0 and a standard deviation of 1. Consequently, change from 1992 to 1996 was relative to change for the total group of states. An increase on an indicator means that the increase for the particular state was larger than the increase for all participating states; a "decrease" may mean that the state did not increase as much as the other states.

Tables 6.8 shows that two classroom practice indicators, I(RC) and I(MD), changed substantially in four of the six states with steady increases in mathematics achievement, and in only one of the other eight states (i.e., California), particularly at grade 8. But the direction of change varied. The indicators increased in Kentucky at grade

8 and in Massachusetts at both grades; they decreased in Louisiana at grade 4 and I(MD) decreased in Texas at both grades. Except for Louisiana, all of these relatively large changes in classroom practice indicators were accompanied by large increases in I(PD).

Comparisons between Tables 6.8 and the first column of Table 6.6 show that the two states with large increases in reform-related classroom practices had state assessments aligned with the SSI goals, while the two states with large decreases had state assessments that were not aligned with the SSI goals.

In New York and Michigan, the two other states with steady increases in mathematics achievement, reform-related indicators did not change much from 1992 to 1996, except for a decrease in I(PD) at grade 4 in Michigan. The two states were similar at grade 4, with indicator means near 0. At grade 8, I(MD) was below average for New York, and above average for Michigan. Since New York's assessment was not aligned with the SSI goals while Michigan's was, this difference in I(MD) supports the conclusion that the use of reform-related practices was related to features of the state assessment.

Among the states with some increase in achievement, Connecticut had the largest increase in I(RC) and I(MD). As Table 6.6 indicates, Connecticut was the only state to change from unaligned to aligned tests.

Four additional states had relatively large changes in I(PD) (i.e., increases at grade 4 for Georgia, Arkansas, California, and New Mexico, and a decrease at grade 8 for New Mexico), but these changes were not accompanied by notable changes in other indicators. None of these states were using assessments aligned with SSI goals.

The general conclusion from these findings is that statewide changes in instructional practices are associated with steady achievement gains. However, the direction of the change is relative to the content of the state assessment. When assessments were aligned with the goals of the SSI, reform-related instructional practices increased; when they were not aligned, reform-related instructional practices decreased.

Can Statewide Achievement Gains be Attributed to a State's SSI?

The previous section identified several characteristics that differentiated SSI states making steady achievement gains from other SSI states. This section focuses on the observed changes in each state, in light of the focus of each state's SSI. The section ends with a comparison of states that received Phase II funding to other SSI states, as well as to the non-SSI states in the trend sample.

Table 6.8a

Means of Selected Standardized Reform-Related Indicators and Relative Change from 1992 to 1996 for Grade 8

	1992 Standardized Indicators			1996 Standardized Indicators			Relative change from 1992		
	I(RC)	I(MD)	I(PD)	I(RC)	I(MD)	I(PD)	I(RC)	I(MD)	I(PD)
Steady Increase									
Texas	.07	.35	.16	.04	–.10	.34	–.03	–.45	.18
New York	–.09	–.19	–.27	–.02	–.33	–.18	.07	–.14	.09
Michigan	.21	.22	–.07	.06	.24	–.07	–.15	.02	.00
Kentucky	–.05	.18	–.15	.22	.44	.50	.27	.26	.65
Massachusetts	–.09	–.32	–.35	.18	.10	.48	.27	.42	.83
Louisiana	–.05	.00	–.07	–.24	–.13	–.21	–.19	–.13	–.14
Some Increase									
South Carolina	.14	.09	.11	.04	.15	.06	–.10	.06	–.05
Georgia	.06	.11	–.19	.12	.25	.00	.06	.14	.19
Connecticut	.11	.02	.20	.27	.23	.04	.16	.25	–.16
Arkansas	–.34	–.32	.27	–.26	–.49	.10	.08	–.17	–.17
Little/No Change									
Maine	.11	.17	.07	.09	.13	–.12	–.02	–.04	–.19
California	.12	.16	.38	.23	.56	.53	.11	.40	.15
Nebraska	–.18	–.03	.03	.00	–.02	–.27	.18	.01	–.30
New Mexico	–.15	–.04	–.19	–.14	–.06	–.54	.01	–.02	–.35

Table 6.8b
Means of Selected Standardized Reform-Related Indicators and Relative Change from 1992 to 1996 for Grade 4

	1992 Standardized Indicators			1996 Standardized Indicators			Relative change from 1992		
	I(RC)	I(MD)	I(PD)	I(RC)	I(MD)	I(PD)	I(RC)	I(MD)	I(PD)
Steady Increase									
Texas	.26	.37	.21	.34	.11	.53	.08	–.26	.32
New York	.01	.09	–.26	.07	–.10	–.19	.06	–.19	.07
Michigan	.15	.04	.25	–.02	.03	–.07	–.17	–.01	–.32
Kentucky	.19	.38	.04	.19	.40	.26	.00	.02	.22
Massachusetts	–.06	–.04	.04	.06	.14	.31	.12	.18	.27
Louisiana	.25	.25	.09	–.01	–.15	.15	–.26	–.40	.06
Some Increase									
South Carolina	.09	.07	.13	.09	.05	.11	.00	-.02	–.02
Georgia	.09	.07	–.26	.16	.19	.05	.07	.12	.31
Connecticut	.16	.11	.09	.21	.29	–.10	.05	.18	–.19
Arkansas	–.37	–.46	.05	–.24	–.48	.32	.13	–.02	.27
Little/No Change									
Maine	.03	.16	.09	.20	.33	.03	.17	.17	–.06
California	.10	.26	.19	.12	.31	.44	.02	.05	.25
Nebraska	–.10	–.09	–.25	–.12	–.08	–.14	–.02	.01	.11
New Mexico	–.13	.06	–.34	–.02	.00	–.11	.11	–.06	.23

Close-to-Classroom Focus: New York, Kentucky, California, Nebraska, and New Mexico. For these states, it is unlikely that the SSI program had a statewide impact on student achievement. The design of the SSI was to work with a relatively small proportion of the schools throughout the state. Information from Tables 6.2 and 6.3 shows little change in ratings of statewide components of systemic reform during the SSI. (Ratings for Nebraska and New Mexico were not available.) Three of these five states, California, Nebraska, and New Mexico, had no or little statewide gains in student achievement (Table 6.1). However, two states, New York and Kentucky, had steady increases. Information from Tables 6.2 and 6.3 shows that these two states were fairly high in systemic reform prior to the SSI, increasing slightly during the SSI. For these states, the steady increase in achievement may have resulted from the system that was in place prior to SSI funding.

State System and Infrastructure Focus: Texas, Michigan, South Carolina, Georgia, and Connecticut. Among these states, two showed steady increases in student mathematics achievement and three showed some increases. The two states with steady increases, Texas and Michigan, differed somewhat on pre-SSI policy ratings. Texas started with the second highest mean policy rating (3.3) and ended with the second highest (3.8), while Michigan began near the group average (2.7) and increased to 3.1, still lower than the point at which Texas started. On infrastructure, both states had relatively high average ratings prior to the SSI (3.0 for each), and these ratings increased by 0.5 during the SSI.

Among the states with some increase, achievement gains in Connecticut kept pace with those of the Steady Increase states from 1992 to 1996. During the SSI, Connecticut's mean policy rating increased from a relatively low 2.4 to 3.2, while the strong infrastructure rating of 4.0 continued. At the end of Phase I funding, Connecticut seemed ready to sustain the rate of achievement gains, but it slowed from 1996 to 2000. In Georgia, mean statewide policy ratings increased from 2.9 to 3.4 and statewide infrastructure increased from 2.0 to 3.0. While these increases are notable, Georgia's infrastructure rating was still one of the lowest. Gains from 1996 to 2000 were larger than in previous years, but not as large as in states with higher ratings for systemic reform components. Ratings for South Carolina were not available.

For this group of states, it seems reasonable to conclude that the SSI contributed to achievement gains.

Balanced Focus: Massachusetts, Louisiana, Arkansas, and Maine. As a group, these four states had pre-SSI ratings among the lowest in the trend sample. By the end of Phase I, all had made substantial gains in both policy and infrastructure, with increases close to a full point or more. Two of the four states in this group, Louisiana and Massachusetts, received Phase II funding, and had achievement gains from 1996 to 2000 that were larger than those from the previous four years.

In contrast, neither Arkansas nor Maine received Phase II funding and had very small achievement gains from 1996 to 2000.

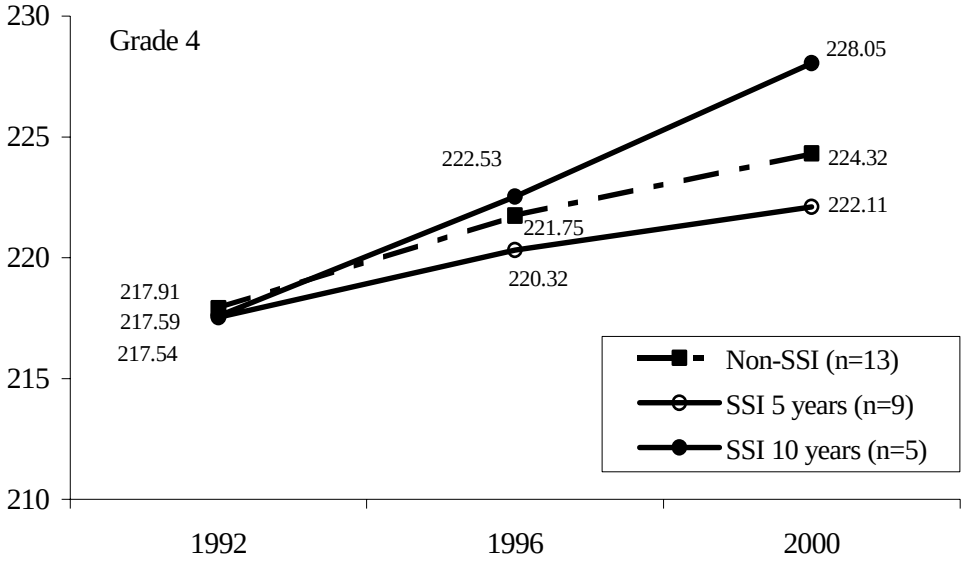
The results for this group raise the question of how long it takes to create systemic reform policy and infrastructure that is self-sustaining. With Phase II funding, Louisiana and Massachusetts were able to achieve above-average gains from 1996 to 2000. The substantial accomplishments made in Arkansas and Maine may have been too new to be able to continue without ongoing support and resources from the SSI program.

In summary, all but one of the SSI states had increases in ratings of selected components of systemic reform over the course of the SSI. States with relatively strong ratings pre-SSI generally had steady gains in student achievement. States with large increases in ratings during the SSI were able to steadily increase achievement with Phase II funding, but achievement gains did not occur in states that did not receive Phase II funding. In all states, the alignment of state frameworks and assessment with the SSI goals was an important influence on statewide student achievement. California, the only state with a decline in pre-SSI ratings, illustrates the substantial role external forces play in shaping the success of the SSI.

States with Phase II funding. Five of the 14 SSI states in the trend sample received Phase II funding: Connecticut, Louisiana, Massachusetts, South Carolina and Texas. Two of the Phase II SSIs had a balanced focus (Louisiana and Massachusetts) and three had a focus on State System and Infrastructure (Texas, Connecticut, and South Carolina). Massachusetts and Louisiana started with some of the lowest ratings for Policy and Infrastructure and had the largest gains during the Phase I, as shown in Table 6.3.

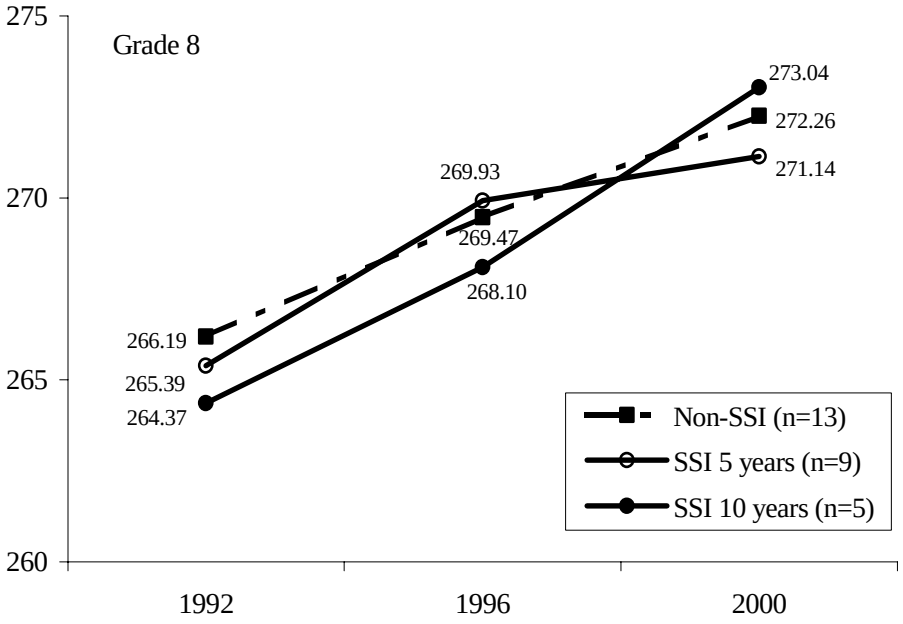
Figures 6.1 and 6.2 show how the mean achievement gains in Phase II SSI states compare to gains in other SSI states as well as to gains in non-SSI states. At grade 4 in 1992, the mean NAEP mathematics composite was almost identical for the three groups of states. In 1996, means were noticeably different, but the differences were relatively small. By 2000, the five states that had received Phase II funding averaged 4 to 6 points higher than the other two groups.

Figure 6.1. Mean NAEP mathematics composite as a function of SSI status for grade 4.



At grade 8, the states receiving Phase II funding made substantial and steady increases across each four-year interval. The non-SSI states also had steady gains in the mean NAEP mathematics composite, though at a somewhat slower rate than the SSI states with Phase II funding. SSI states with only Phase I funding had the largest gain from 1992 to 1996, but the smallest gain from 1996 to 2000.

Figure 6.2. Mean NAEP mathematics composite as a function of SSI status for grade 8.



The graphs show that all groups of states had achievement gains in both four-year intervals. By 1996, there are differences between the groups, but they are relatively small. However, by 2000 the differences are substantial, with Phase II states having the largest gains. It may take several years to see evidence that a particular approach to systemic reform is more effective than other ongoing efforts.

The graphs also show that from 1996 to 2000 the rate of gain slowed considerably in SSI states with only Phase I funding. Since the rate decreased when the SSI program ended, this result provides evidence for the claim that the SSI resources were related to the rate of gain from 1992 to 1996. This also suggests that the reforms were not sufficiently institutionalized to be self-sustaining after the first five years of funding.

When is statewide achievement gain an appropriate measure of systemic reform efforts?

The SSI program was designed to raise student achievement statewide. In some states, substantial and steady gain is evident, but in others the mean mathematics composite scores changed very little across the eight years from 1992 to 2000. The analyses above suggest that statewide achievement gains are likely to be evident under certain conditions:

1. The statewide assessment is aligned with the SSI goals.
2. The assessment is criterion-referenced, based on the state standards or frameworks.
3. Statewide policy and infrastructure are relatively strong (mean rating near 3.0) and well established.

If these conditions are not in place, it may be unreasonable to expect statewide achievement gains. Rather, program evaluators may want to focus on progress toward those conditions that support systemic reform. Analytic models of systemic reform with specific criteria and benchmarks, like the one developed by Clune (1998) and Clune and his colleagues (2000), may be helpful in this effort.

Some SSIs directed their focus “close to the classroom.” While there is a place for such initiatives, their evaluation design should probably differ from that of SSIs that address state policy and infrastructure. Effects in a relatively small number of schools are unlikely to result in a noticeable change in a statewide mean. However, even when the SSI focus is close to the classroom, the alignment of the SSI goals and state assessments is likely to be an important consideration. Unless there is a reasonable match between at least some components of the state assessment program and SSI goals, reform efforts may be compromised.

Discussion

NSF's Statewide Systemic Initiatives program took the role of the state seriously as NSF partnered with each participating state to plan and implement statewide educational improvements. While much was accomplished in many states during the SSI program, the results did not necessarily lead to above average increases in statewide mathematics achievement. However, SSI states with notable gains seemed to share common features, including:

- State curriculum standards in mathematics,
- Tests based on the state standards,
- Test results used for school and/or student accountability,
- Alignment of SSI goals and state assessments, and
- Effective statewide infrastructure for staff development.

Smith and O'Day (1991) outlined the concept of systemic reform more than 10 years ago. They focused on the important role of the state in promoting, supporting, and sustaining educational reform:

> [T]he states are in a unique position to provide coherent leadership, resources, and support to the reform efforts in the schools. States not only have the constitutional responsibility for education of our youth, but they are the only level of the system that can influence all parts of the K-12 system: the curriculum and curriculum materials, teacher training and licensure, assessment and accountability. (p. 246)

Smith and O'Day proposed a coherent system of instructional guidance, emphasizing three aspects:

> curriculum frameworks and school curricula,
> professional development, both preservice and inservice, and
> accountability assessment.

The findings of this chapter fit well with Smith and O'Day's proposal for a coherent system of instructional guidance. But it is clear that coordinating and aligning the many components of a statewide educational system is neither easy nor quick. In their article, Smith and O'Day described the fragmented, complex, multilayered educational policy system in the United States:

> On the formal policy side, school personnel are daily confronted with mandates, guidelines, incentives, sanctions, and programs constructed by a half-dozen different federal congressional committees, at least that many federal departments and independent agencies, and the federal courts; state school administrators, legislative committees, boards, commissions and courts; regional or county offices in most states; district level administrators and school boards in 14,000 school districts (with multiple

> boards and administrative structures in large systems); and local school building administrators, teachers and committees of interested parents. Every level and many different agencies within levels attempt to influence the curriculum and curricular materials, teacher in-service and pre-service professional development, assessment, student policies such as attendance and promotion, and the special services that schools provide to handicapped, limited English-proficient and low-achieving students. (Smith & O'Day, 1991, p. 237)

Effective systemic reform addresses all of the policy layers. While a reform may focus on a subset of the policy layers, the context provided by other governance layers is extremely important. The 14 SSI states described in this chapter illustrate the challenges of coordinating all elements of the system. However, those with above-average gains show that when the major components work together, notable student achievement gains result.

The conclusions of this chapter also fit well with those of Volume I, *Lessons Learned About Designing, Implementing, and Evaluating Statewide Systemic Reform.* The researchers identify two key tasks of systemic reform:

1. Establish the kinds of interventions needed to increase the knowledge and skills of the participating teachers, administrators, and students.
2. Foster the professional, political, and public support to put the needed interventions in place on a large scale.

On the basis of results reported in the present chapter, the second task seems crucial for realizing achievement gains statewide. Absent a supportive policy context, it is questionable whether any reform can impact schools statewide.

The findings of this chapter also have implications for the evaluation of systemic initiatives, whether for an entire state or for a subset of districts or schools within a state. Increased student achievement for all students is an important focus of evaluation efforts. But features of the reform context are equally important for understanding how gains were achieved, or why they were not achieved. Examining policy coherence—in assessment and accountability, in professional development, and in instructional practices—can inform our understanding of effective systemic reform.

The results highlight the role of state assessment and accountability policies in reform efforts. When these policies are aligned with the goals of a systemic initiative, and when the statewide infrastructure is sufficiently strong to support teachers and schools as they change their practices, systemic reform can result in substantial achievement gains in a relatively short time.

Reform efforts may be compromised when statewide policies are not aligned with the goals of systemic change. Evaluation efforts that address the extent of the match over time will provide a means of measuring progress toward the coherent instructional

guidance system Smith and O'Day envisioned. Work by Clune and his colleagues provides a model for assessing the many and varied components of systemic reforms (Clune et al., 2001; Clune, 1998).

The SSI program was designed to catalyze educational reform efforts in participating states. From 1992 to 1996, student achievement in states with SSI funding increased slightly more than in states without SSI funding, providing evidence for the catalytic role of the SSI. During Phase II, states with continued funding experienced even greater achievement increases than from 1992 to 1996. By contrast, states that did not continue on in Phase II maintained the gains they had made but did not sustain their *rate* of achievement gain, especially at grade 8. This finding also supports the catalytic role of the SSI program, since the rate of gain slowed for those states where funding was discontinued.

These results point to how important it is to sustain systemic reform efforts once external funding has ended. Whether the states with Phase II funding will be able to maintain their rate of change after 10 years of the SSI program is not known at this time. Understanding how externally funded programs can be continued when external resources are no longer available is a central issue for both funding agencies and program evaluators.

CHAPTER 7

CONTRAST OF STATE NAEP WITH THREE STATE ASSESSMENTS

Introduction

In addition to its analysis of State NAEP data, the Study of the Impact of Statewide Systemic Initiatives investigated test results for three statewide assessment programs. While State NAEP mathematics content represents a national consensus of valued knowledge and skills and offers uniform information across states that allows for comparisons between aggregates of SSI and non-SSI jurisdictions, it has limitations. First in importance is the variance of state mathematics frameworks from those used for the State NAEP. A second limitation is that, because of its matrix sampling design, the State NAEP assessment allows only for state level inferences from its results. A third, often overlooked limitation, is that, because it provides no student-level results, the degree to which students are motivated to perform well is questionable. In addition to their potential to compensate for these limitations, state assessments provide complementary data that can be used to corroborate NAEP achievement levels and trends. Data from multiple measures improve confidence over single-instance information. With regard to trends, the additional data points, frequently available because state assessments are usually administered more regularly than the State NAEP, serve to verify NAEP trends across years when there are no NAEP data available. Similarly, state assessment data may support extrapolation of NAEP trends to years beyond those in which NAEP has been administered but for which state assessment information is available.[1]

When there is information about differential levels of SSI intervention among schools within a state, the school-level information from state assessments provides a unique opportunity to determine whether achievement levels are impacted by the SSI efforts. Such information about the differential level of SSI activities among schools was available for two of the three states for which state assessment results were investigated in this research.

It is worth noting that State NAEP results can be used to improve the understanding of state reform efforts. For example, Klein et al. (2000) and Linn (2000) suggest that improvements in performance on state tests may result from teaching to tests, thus narrowing curriculum. If the content of state assessments is more narrow than that of the State NAEP, comparison of NAEP and state assessment data should show whether improvements in mathematics achievement is limited to the specific content of the state

[1] Initially the Study of the Impact of the Statewide Systemic Initiatives was funded to analyze State NAEP data from 1990 through 1996 and state assessment data was to be explored as a means of extrapolating NAEP results to subsequent years. In 2000, funding became available to add 2000 State NAEP data to the study.

measures, or whether it also extends to the broader content that is measured by State NAEP.

It is important to keep in mind that state assessments differ substantially. Feuer et al. (1999) noted the following dimensions of dissimilarity among state programs: content coverage, item format, test administration procedures, intended use and purpose, and stakes or the consequences linked to their results. Laguarda, Breckenridge, and Hightower (1994) classified statewide programs on the basis of grades tested, item types, scoring (criterion-referenced, norm-referenced, or performance level), and reporting level (district, school, student). Proper interpretation and, especially, comparison of results from state assessments requires a through understanding of the characteristics of the overall program.

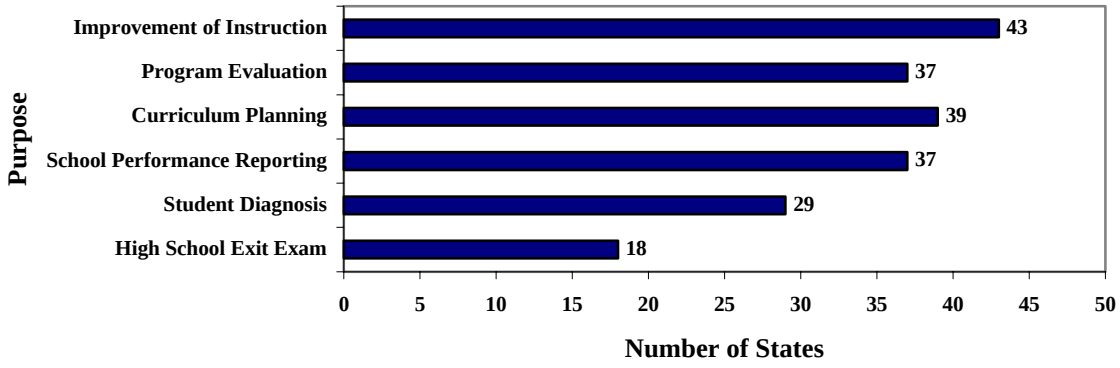
Description of State Assessment Programs

The assessment programs of the three states used to study the impact of the Systemic Initiatives were analyzed in terms of six categories: purposes, test and program design specifications, test domain, technical characteristics, data, and documentation. The categories are described and discussed in the following eight sections.

Purposes

Testing purposes influence program design and the importance and use of results, as well as student motivation to reflect their full capabilities in their performance. In order to interpret a program's outcomes and compare them with other indicators, it is essential the characteristics of the assessment system be fully articulated. Figure 7.1 shows seven purposes that Feuer et al. (1999) identified for state assessment in the 1996–1997 school year. In the latter half of the 1990s, as states worked to become compliant with the *Improving America's Schools Act* (*IASA*) (U.S. Department of Education, 1994), "school performance reporting" has become synonymous with accountability and increased stakes for teachers, schools, and districts. Of course, high school graduation requirements establish very high stakes for students. While program evaluation, curriculum planning, and improvement of instruction yielding higher student achievement levels are likely purposes associated with any state assessment, high-stakes purposes most often take precedence in design decisions regarding state measures.

Figure 7.1. Most frequently reported assessment purposes.



Differences in stakes and in level of awareness of results, as well as the frequency of reporting results, are considerations when comparing state assessment and State NAEP results. Such differences between measures stem from the differences associated with their intent or purposes. The State NAEP is intended to allow volunteering states to compare their students' mathematics achievement to national performance standards as well as to the performance of students in other states (Allen et al., 1997). There are, however, no stakes associated with either set of comparisons. In looking at state assessments and State NAEP results, one should consider how the uses of the data as dictated by purposes may influence the results themselves. For example, where teachers and school administrators have incentives to attain high student performance, students are likely to be influenced to perform well. Furthermore, over time, students may become familiar with pertinent content and acquire test-taking skills essential for particular item types and formats, thus influencing not only annual scores but score trends across years. Certainly, any student stakes, such as graduation or promotion requirements, associated with a state measure increase the level of motivation compared to the motivation for answering NAEP questions. This factor may confuse the comparisons of results on the two assessments.

Test and Program Design Specifications

For the purposes of this report, test and program design specifications are classified into eight categories. These categories are: 1) test source; 2) interpretive model; 3) item type; 4) item distribution; 5) test length; 6) population; 7) score type; and 8) reporting. Each of these design specification categories are described and discussed briefly.

Test source. Typically, tests used for statewide programs are either commercial or customized. Commercial tests are available to schools, districts, or states from publishing companies and are typically based on a national consensus content framework. In contrast, customized instruments are developed for individual states, based on the particular state's test framework, or content and performance standards. Hybrid programs may add items to a commercial instrument in order to bring assessment into alignment

with their frameworks. With the increased focus on test alignment[2] that has occurred since the implementation of *IASA* (U.S. Department of Education, 1994), customization is becoming essential if states are to meet federal requirements.

Interpretive model. It has become customary to classify educational achievement tests as norm- or criterion-referenced. However, in practice, data from administration of a single instrument is *interpreted* in terms of a pre-existing standard, as well as in terms of a defined group's performance. State NAEP is an example of this practice. Because norm- or criterion-referencing is not a test characteristic per se, for the sake of clarity, it seems best to ascribe these terms to the interpretation of test data rather than to the test itself.

Even though a single test's results can be interpreted either way, design specifications should differ depending on the desired interpretation of its results: for example, if a test is to be used primarily for norm-referenced interpretation, it should be designed to discriminate best around the test's mean. For criterion-referenced uses involving decisions such as pass/fail or classification into performance categories, tests should discriminate best at critical cut scores.

Item type. There are two major classifications for achievement test items; these are selected- and constructed-response. Within the category of selected-response are: multiple-choice, true/false, and matching. Constructed-response items vary on continua of the length and complexity expected of test-taker responses. Item-type differences between tests complicate the comparison of test results. Because differing item formats may require variant cognition, even when two test blueprints call for similar content, the extent of comparability for measures based on two such blueprints remains unknown. Also, comparability may be compromised because different item types may require varying levels of engagement or motivation on the part of students. Not only do these problems cause difficulty when comparing results from distinct testing programs, but they also interfere with within-program trend analysis if item types are not constant across years.

Item distribution. Typically, instruments used in large-scale assessment programs either administer a common set of items to all subjects or a sample of items to students. In the latter case, a large domain of items is covered by the assessment with the individual student being exposed to only a subset of the domain. The advantage of the sampling approach is that it allows, within a given time limit for testing, the coverage of a broader domain of subject-matter content. For state assessments, the sampling approach has a significant shortcoming. In order to make defensible inferences about the entire content domain, within a reporting category (e.g., a school or district), a substantial number of subjects take each item in a domain. This is problematic because every state has a number of small schools in which the above requirement cannot be met. In the past, some states (e.g., Maine and Massachusetts) have tried to get around this problem by administering a common set of items to all students and then have applied statistical procedures to estimate individual student performance over the entire content domain.

[2] See above, pp. 22–30 and 74–77 for a good definition of alignment and the criteria for its evaluation.

Such procedures would be unlikely to pass the acceptable criteria (see U.S. Department of Education, 1999) requirement for alignment.

Test length. Test length is a function of time spent taking the test as well as the number of items administered. Length of engagement is a consideration in student performance on an instrument. As tests differ in length, the comparability of results may be affected. This may be a cause for concern across programs, or from year to year within programs.

Population. The federal *Individuals with Disabilities Act* (*IDEA*) and *IASA* require that nearly all disabled and LEP students under the purview of these statutes be included in statewide tests. Because, at least in the past, inclusion rates varied among states, districts within states, and schools within districts, whenever comparisons are made or trends studied, it is important to pay attention to the inclusion rates for the entities and years under consideration. Level of absenteeism is another, often overlooked, issue that may affect the character of populations considered for comparison.

Score types. Raw, scale, proficiency, national or state percentile, grade equivalent, national stanine, and normal-curve equivalents (NCE) are the types of scores typically available from educational assessment systems. While proficiency scores yield criterion-referenced interpretations, the others are norm-referenced. Because proficiency standard setting procedures are unique to assessment programs, these scales are of little value when comparing states or when comparing states with State NAEP. Within systems, across years, the percentages of students in various proficiency categories can serve as good indicators of distributional changes.

Reporting. The levels and categories of disaggregation are the primary issues for reporting. Typical levels for state assessments are: school, district, and state. *IASA* required that states report student results by gender, race, socio-economic status (SES), migrant status, disabilities, and LEP. Free or reduced-cost lunch is the SES indicator most often used by states. *IASA* also requires that results at the school and district level be reported separately for students in attendance for a full academic year. Prior to the time that state assessment systems were brought into compliance with *IASA*, disaggregation was limited mostly to gender, race, and SES, with even the latter being frequently omitted.

Large-scale assessments frequently divide major domains, such as mathematics, into subclassifications. For example, NAEP uses the subcategories of Number Sense, Properties, and Operations; Measurement; Geometry and Spatial Sense; Data Analysis, Statistics, and Probability; and, Algebra and Functions. However, probably due to the lower reliability of such subscores, particularly at the individual level, a limited number of scale types are offered by state assessment designs. Percent correct seems to be the most often used subscore indicator.

Test Domain

Content frameworks or performance standards typically delineate test domains. Broad disciplines such as mathematics are usually broken down into classification categories such as strands (e.g., NAEP), or objectives (Texas' TAAS). Some frameworks or standards structures —the State NAEP, for example—mix disciplinary and cognitive representations of content. Frameworks or standards that appear to define similar content or cognition may, in fact, yield test items that are considerably different in the scope and/or complexity of the cognitive demand required for adequate performance. Even though instruments may be designed for the same content area and grade level, differences between the test's domains, or the inability to adequately evaluate the test's domains, may severely limit the comparability of results from two measures.

While few assessment programs analyze the cognitive demands of their tests, judgments about the comparability of results requires confidence about the equivalence of thinking skills required for performance on two measures. Kenney and Silver (1998) devised a process for evaluating the cognitive demand of mathematics items and used it to compare State NAEP and a statewide mathematics test. These researchers found that panelists reached consensus by rating the cognitive demand of items as either high or low. Typically, problem solving, reasoning, communication, and connections were rated as high for cognitive demand, while recall of facts, routine procedures, and estimation were rated as low.

Technical Characteristics

Table 7.1 shows the categories for judging the technical quality of statewide assessment programs that have been developed by the U.S. Department of Education (1999). These criteria were developed to represent a consensus [see American Educational Research Association, American Psychological, National Council on Measurement in Education, (1999)] of principles and criteria for technical quality and served as the foundation of the USDE's evaluation of a state's compliance with the assessment technical quality requirements of *IASA*.

Table 7.1
Dimensions of Technical Quality

Dimensions	Description
Validity	• Is there evidence for the inferences intended for the instrument? • Since inferences are about constructs, validity study requires gathering evidence about test: content; relationship to other variables; respondents' cognition; and internal structure.
Reliability	• What is the extent of error reflected in the measures results? • Reliability studies focus on the sources of unintended variation in results. • Reliability is specific to inferences and depends on the level of results reported as well as the nature of reporting.

Fairness/accessibility	• With respect to the instruments domain, do the items allow all students equal opportunity to demonstrate their level of knowledge and skill? • Have all students had equal opportunity to learn the knowledges and skills necessary to perform the tasks required by items?
Comparability of results	• Are inferences about a single domain equivalent across different forms and administrations? • Are the constructs measured the same for different test forms? • Are test administration specifications and procedures the same for different administrations?
Administration, scoring, analysis, and reporting	• Are there standardized procedures? • Is the use of the standardized procedures insured?
Interpretation and use	• Is necessary knowledge present for users to make the intended use of results?

Validity. Within the context of large-scale assessments, judgments about validity depend on the evidence available to support intended inferences. Messick (1989) noted that "Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment." As such, while decisions about the adequacy of validity may involve quantitative criteria, they are, typically, qualitative in nature. The author's use of the term "integrated" raises the importance of expert judgment in analysis and decision-making about validity.

Historically, the analysis of validity has included three types: content; construct; and predictive and concurrent, criterion-related. However, over the past 20 years, a unified theory has evolved that views test content and criterion-related data as sources of evidence for construct validity (see Messick, 1989; American Educational Research Association, 1999). Thus, current thinking about the validity of educational achievement measures requires that we judge a test's adequacy based, in part, on how well its results represent our conceptions of the underlying domain of interest. This requirement implies that we have a clear and comprehensive understanding, not only of a domain's knowledge and skills, but also of the cognitions that support their application to test performance.

The *IASA* Peer Review Guidance (U.S. Department of Education, 1999) discusses four types of evidence for test validity: first is information about a test's contents alignment with the domain's content specifications; second is the relationship of assessments results with those of other indicators; third is evidence about the alignment of test takers' cognitions with those laid out in the domain specifications, and, fourth is evidence of the internal consistency of the measure.

The inference-based conception of validity requires a clear, complete, unambiguous description of the referent for the inference. Such attention to the distinguishing features of domains would be helpful, not only for pointing to evidence

necessary for judging validity, but also for informing decisions about the comparability of measures.

In addition to evaluating whether inferences are valid, measurement experts increasingly stress that thorough evaluation of validity requires study of the consequences of assessment (Messick, 1989; American Educational Research Association, 1999). One perspective of such study would examine whether the original purpose for assessment is accomplished. For example, where a purpose underlying assessment for accountability is that student achievement will be improved, the study of validity should include a component to judge whether this goal is accomplished. In designing such a component, there would need to be a clear theory of action that connects the test, inferences, and decisions to the desirable consequence of learning improvement. Another focus of consequential validity is on unintended consequences. Messick (1989) gives the example of using a quantitative measure that is known to disadvantage females as an illustration of an unintended source of consequential invalidity. When judging the validity of a state assessment program, one should be alert for attributes that disadvantage subgroups within the population. As an example of consequential invalidity in statewide assessment, Haney[3] argued that the use of Texas assessment results as a requirement for high school graduation influenced minority students to drop out of school before taking the test.

Reliability. A portion of all educational achievement scores is attributable to error. The study of reliability aims to quantify the extent of such error in any score. While measurement error is associated with the test scores, it is an important consideration when *decisions* are made on the basis of scores. Particularly for high stakes decisions, it is important to consider the extent of error. When scores fall close to cut scores, interpretation of these results should be attuned to the reality that a decision based solely on the score depends on random or chance factors. Another reliability issue that is important for the interpretation of results is that the amount of error variation on a test is not constant for all scores. In the cases of decisions based on specific cut points on a test's score scale, it is desirable that the test be most reliable (yield the least error) at those points.

It is important for states to provide indicators of score stability. Of particular value, for individual scores, are indicators that relate directly to the interpretation of results. One example is confidence intervals based on standard errors of measurement that provide a clear view of the range expected for observed scores over repeated testings. Because the analysis of this report deals with groups and their means, it is useful to clarify that the standard error for a mean of a group of examinees is $\frac{1}{n}$ the standard error of an individual in the group, where *n* is the number of students in the group (Stanley, 1971). Another caveat for examining the reliability of state data relates to disaggregated scores. Because there may be sources of error that are associated with specific groups, the reliability of a test may differ across groups. Therefore, it is desirable that states provide an indicator of the error associated with scores from the various groups of interest.

[3] See W. Haney (2000), The Myth of the Texas Miracle, in *Education Policy Archives, 8*(41). Retrieved from http://epaa.asu.edu/epaa/r8n41.

Fairness/accessibility. Fairness and accessibility are related to bias in test use. Cole and Moss (1989) observed that issues of bias represent a facet of validity. Since a validity study involves the examination of evidence regarding inferences from test data to the construct specified by the domain description, bias must deal with rival hypotheses or explanations for test performance. For example, rather than indicating low mathematics achievement, a minority student's performance on a test may reflect his/her unfamiliarity with the cultural context within which problems are posed.

Usually, studies of bias focus on the differential performance on test items by subgroups of the population. Items are considered biased if the performance of examinees of a given subgroup is out of step with the examinees' overall performance on a test. The *IASA* Peer Review Guidance manual (U.S. Department of Education, 1999) outlines criteria to be applied when judging whether assessments are fair and accessible to disabled and limited English-speaking examinees. A gross indicator of accessibility is the proportion of enrolled students from various student subpopulations who take tests. Because of the potential for underrepresentation of some segment of the population, differences in such proportions across tests should raise questions about the comparability of results.

Comparability of results. The results of two tests can be considered comparable if they measure the same construct(s) and both tests yield the same score for examinees having the same ability. In practice, no two tests can meet this criterion; however, if the constructs and item characteristics measured are similar, there are statistical procedures available for equating the results from the two tests. Year-to-year comparisons require such equating procedures. It is because State NAEP assessments and tests from different states do not approximate the conditions for equating that comparisons of results from the programs are perplexing.

Administration, scoring, analysis, and reporting. Except in cases where accommodations are acceptable, test administration procedures should be the same for all examinees. Preparation for, time limits, directions, and assistance and advice during test periods should be standardized. Procedures should be in place for large-scale assessments to insure and monitor that such standardization occurs.

Selected-response type items are scored, simply as right or wrong, according to a predetermined key. However, constructed-response questions require detailed rubrics and procedures, including methods for evaluating the reliability and validity of the ratings given to student responses. Judgments about the validity of items and tests rest, in part, on the nature of scoring rubrics.

Typically, the scales used to report results require sophisticated and highly complex mathematical calculations. The details of such procedures need to be specified, as well as the quality control methods used to insure their accuracy.

All scoring, analysis, and reporting depend on highly automated systems that require quality control steps to insure accuracy. To assure the credibility of assessment systems, the output from such quality control efforts should be public.

Interpretation and use. Interpretation and use of the results of any assessment depends on the overall technical quality of the various aspects of the system. Assuming high validity and reliability, fairness and freedom from bias, comparability and high quality administration, scoring, analysis and reporting, the utility of assessment programs rests on the accuracy and user-friendliness of its reports. Reports should be as straightforward as possible. Where necessary, assessment programs should include a component for training users to interpret results. One means of assessing the merit of reports is to compare them with the purposes of the assessment system. Data on reports should align with purposes and allow them to be accomplished.

Data

The study of state assessments by external researchers depends on the availability of data from state agencies. There are two kinds of issues related to the provision of state data: One has to do with the willingness and capability of states to provide their data and the second relates to the nature of data that states retain.

In the course of the research reported herein, we encountered states that refused to provide data. In addition, the contractor of one state that agreed to participate was unable to provide data due to a shortage of staff. This study required data that had been archived for some time. In some cases, this resulted in a shortage of information about not only the data, but about the design, processes, and procedures of the assessment programs that produced the data. The availability of documentation of programs prior to the late 1990s was scant. Attempts to supplement documentation by personal communication was most often frustrated by either the unavailability or the actual absence of knowledgeable people. Both state and contractor staff members were simply too busy to deal with requests about earlier operations. Our experience with the acquisition of data from state assessments suggested that, at least in the past, states have not anticipated the needs of external researchers when planning data archiving or program documentation. This should not be surprising, since such research is typically not a purpose of the statewide assessments.

This study found variance in the nature of data available from states. One state was able to supply item-level data for individual students. Another state supplied only scale-score information on schools and districts. States also differed in the nature of demographic breakdowns of data. As states came into compliance with *IASA*, they gathered data on the demographic characteristics required by that law, so such information should be available for future researchers. This capability, together with item-level student data, would greatly facilitate the work of external researchers needing state assessment data for their studies. In particular, student data is necessary for meaningful cohort studies across grades.

Data and Program Documentation

Detailed documentation of the specifics of assessment program designs, processes, and procedures, as well as the data they yield, is essential for the effective study of the results of such assessment. Testing officers contribute greatly to the understanding of their programs by providing concise but comprehensive, self-contained written descriptions of their assessment programs. In particular, the lack of detail about technical aspects such as scaling and equating jeopardizes proper analysis and interpretation of results. In addition, it is important that specific documentation about test frameworks, test items, and procedures for equating content be available. Documentation should also clearly delineate the population tested. Guidelines for inclusion of LEP students and students with disabilities, as well as mobile students, should be spelled out.

In addition to documentation of program features, states should provide a detailed inventory of data gathered, as well as how it has been coded and stored. There should be record layouts and detailed written descriptions of all data fields. Ideally, descriptions should include the formulas, algorithms, and software used for calculations. This information will help insure proper interpretation of the results of analytic data supplied in a database.

Descriptions of Statewide Assessments for Maine, Massachusetts, and Texas

For two of the three states involved in this study, Maine and Massachusetts, the assessment programs changed substantially near the end of the period of the study. However, in Texas, the overall approach remained the same throughout.

Purposes

In the late 1980s though the mid- to late 1990s, the statewide assessments in Maine and Massachusetts were intended, primarily, to provide information for improving curriculum. Maine supplied results to parents and students to allow them to judge individual academic progress. There were no Massachusetts data available at the student level. With the implementation of the Massachusetts Comprehensive Assessment (MCAS) in 1998 and the new version of the Maine Educational Assessment System (MEA) in 1999, both undoubtedly influenced by *IASA*, there was a noticeable focus on school and district accountability. From the beginning, MCAS had sanctions for low-performing schools and districts. There were no such sanctions with the first two years of MEA.

The emphasis in Texas was different. Although TAAS supplied information useful for improving curriculum and instruction and informing parents and their youth about academic performance, the tests were the basis for the student achievement portion of the state's accountability system. Since 1994, schools have been classified as *exemplary*, *recognized*, *academically acceptable*, or *academically unacceptable* based on

the proportion of students in various demographic groups that meet the passing score.[4] In addition to these school- and district-level stakes, there are student-level incentives and repercussions. Students who answer 95% of items correctly, or master all objectives in a subject, are officially recognized, while those who fail to pass TAAS examinations must be offered remedial instruction. Moreover, the perceived importance of the TAAS system, on the part of students, is intensified by the requirement of passing exit-level tests in order to graduate from high school.[5]

In Maine and Massachusetts, through 1998 and 1996 respectively, the statewide assessments were low-stakes. However in Texas, during the period for which TAAS data were analyzed for this study, there were moderate to high stakes for districts, schools, and students. The emphasis in the use of assessment data in the two northeastern states was school- and district-level improvement and, in the case of Maine, provision of information to parents and students. Texas, in contrast, emphasized accountability, as evidenced by specific consequences associated with test results.

Design Specifications

Test source. The assessments of all three states were custom-made for the jurisdictions. The approaches of each state were different. Texas used a traditional design, administering a common set of items to nearly all students. Maine's and Massachusetts' original assessments applied item sampling. Maine mixed one form of items that was common to students, while Massachusetts had a more traditional sampling of forms across students with no common form. With the advent of MCAS, Massachusetts also utilized a set of common items.

Interpretive model. All three states reported results in terms of proficiency categories. The Texas proficiency categories were based on the Texas Learning Index (*TLI*) value corresponding to 70% of items correct in 1994. Subsequently, the *TLI* passing score has been adjusted to insure that it represents a consistent level of performance within the student population across grades. The same *TLI* score across grades indicates the same level of performance relative to the peer group at the given grade (Texas Educational Agency et al., 1999a). Thus, the *TLI* is norm-referenced to the 70% correct criterion of 1994.

It is unclear how the Maine proficiency categories were established. Initially, in 1990, Massachusetts used a normative procedure, based on an arbitrary standard deviation interval from the mean, to set category cut points. However, in 1992, the state redefined the points using a content-based procedure. Since the two approaches yielded similar results, they maintained the 1990 values (Massachusetts Department of Education, 1995a). MCAS used the content-based "Body of Work" approach (Massachusetts Department of Education, 2000) as the basis for the proficiency categories of its tests.

---

[4] http://www.tea.state.tx.us/tssas/9900method, http://www.tea.state.tx.us/perfreport/account/2002manual

[5] http://www.tea.state.tx.us/students.assessment/parents

While, in some cases, the procedures for establishing proficiency categories had normative aspects, all of the states were attempting to provide score interpretations in terms of mathematics content, rather than strictly normative calculations. State NAEP also reports data in terms of its own version of proficiency categories. Because of differing definitions and procedures of proficiency, these scales are not useful for comparing state to NAEP performance. Scale-score means and standard deviations must serve as the basis for such comparisons.

Item types. Among State NAEP and the Maine, Massachusetts, and Texas assessments, a variety of item types are represented. In 1996, roughly 40% of State NAEP items were open-ended or extended open-ended types, while the remainder were multiple-choice. The open-ended items were answered with a numbers or a brief written explanation. The extended open-ended mathematics questions call for complex problem solving and written descriptions of the solution process employed and the result.

Both Maine and Massachusetts used a combination of open-ended and multiple-choice items. While Maine made extensive use of open-ended items, available documentation did not provide information about the distribution of the various types for all forms. However, available copies of tests suggested that the nature of items and their distribution differed across the years 1990 to 1998. In 1992 through 1994, approximately 13% of released items were open-ended, while from 1995 to 1998 all of the released items from the common forms were open-ended. Between 1992 and 1996, about 25% of the Massachusetts item pool was open-ended. MCAS forms consisted of 24 multiple-choice, 6 constructed short-answer, and 7 extended-response items. Texas used multiple-choice items exclusively.

Because of the abstruse relationship between item type and cognitive demand, the differences in item types used among the various assessment systems studied for this report obfuscates comparisons among the results from the various tests. Furthermore, changes in item types within state programs across years draw into question the meaning of trends.

Item distribution. The State NAEP, as well as the Maine and Massachusetts systems, used forms of item sampling to broaden the content coverage of their assessments. Texas administered the same set of items to all students. For State NAEP, the design for sampling was intended to yield only state level results. Maine planned for school, district, and student results, employing a common set of items to obtain the latter. Prior to MCAS, Massachusetts was interested only in school and district information. With the advent of MCAS, student results were derived by means of a set of common items administered to all students on each of the test forms.

State NAEP used a sophisticated balanced incomplete block spiraling design to assure that each item was administered an appropriate number of times (Allen et al, 1997). Available documentation from Maine and Massachusetts did not present detail about the distribution of items; therefore, it is prudent to be wary about the design. Since

the sampling may have been biased, generalization from assessment results to the entire state should be made with caution.

Population. All three state assessments tested nearly all students. Those not tested were typically exempted for reasons of disability or language. The State NAEP sampling frame had a similar focus. However, specific criteria for exempting students from these groups in all likelihood differed across programs. To the extent that this is the case, comparisons are biased in favor of programs with higher levels of exemptions. Information for the programs did not allow judgments about the comparability of exemption procedures. It is expected that as states worked to become compliant with *IASA* and the *Individuals with Disabilities Education Act* (*IDEA*), their criteria for inclusion became more alike.

Reporting. All of the assessment programs studied for this report present results in terms of both IRT-derived scale scores and proficiency scores. However, because the methods for setting the proficiency categories differ across programs, comparisons are limited to the scale scores.

The primary issues in reporting are the grade-level focus and the categories for score disaggregation. Massachusetts' MEAP (1992–1996) provided consistent information about gender. The database available for this state also contained single-year (1996) data regarding the socio-economic status of schools. For the purposes of this study, this indicator was generalized to other years. MCAS, the state's successor assessment program, reported by race, disability, and LEP. The Maine database provided information about gender and race. However, because Maine is so racially homogeneous, this information was of little use for our study. Texas listed the race, gender, and free and reduced-cost lunch status of each student tested.

From 1992 through 1996, Massachusetts reported school-, district-, and state-level results. Although 25% of the items used were open ended, the results from these items were weighted 30%, compared with 70% for the multiple-choice items. MCAS reports student-level data based on the results of the state's common item set. Both Texas and Maine have consistently reported student-level results. In each case, these data were based on a set of common items administered to all students.

Score type. As indicated above, State NAEP as well as the three state programs yielded both proficiency and IRT scale scores. Due to the unavailability of documentation, details of the scaling procedures used in Maine prior to 1999 could not be determined. For MEAP, Massachusetts used Rasch procedures, with school as the unit of analysis, to create a scale with a scale mean of 1300, a standard deviation of 100. The maximum and minimum values of the scale were 1000 to 1600, respectively (Massachusetts Department of Education, 1995a). The Texas Learning Index (*TLI*) at grades 3 to 8 was designed using a scale score of 70 on the TAAS exit-level examinations in 1994 as a benchmark that represented a raw score value equivalent to 70% of the item correct (Texas Education Agency, 1999a). The standard deviations at grades 4 and 8 were 15.1 and 15.4, respectively, for the 1994 administration. A *TLI* score range of 0–100 is

reported in the *Technical Digest* (Texas Education Agency, 1999a). The Maine Department of Education (MDE) reported that the MEAP scale was originally (1985–1986) set with a mean of 250, a standard deviation of 50, and a range of 100 to 400 (Maine Department of Education, 1990). However, our calculations utilizing the database provided by MDE yielded standard deviations that were considerably larger: for example, the *SD* for 1992 at grade 4 was 150.6. In a personal communication with the contractors' representative, this discrepancy was attributed to an arbitrary truncation of the scale.

The New MEA and MCAS scales differed from their predecessors. The Maine test's scale-score range was 501 to 580, while that of the MCAS was 200-280.

Test Domains

The Maine, Massachusetts, and Texas assessments, as well as the State NAEP, are based on similar but distinct frameworks. NAEP items are classified on five "content areas and strands" and three "dimensions of general mathematical mental abilities" (College Board, National Assessment Governing Board, 1997). The content strands are: Number Sense, Properties, and Operations; Measurement; Geometry and Spatial Sense; Data Analysis, Statistics, and Probability; and, Algebra and Functions. Conceptual Understanding, Procedural Knowledge and Problem Solving comprise the mathematical mental abilities dimensions. In order to compare the domains of the four assessments, the domains of each state were compared with those of the State NAEP. In the case of Maine and Texas, experts in mathematics education did the comparisons at the item level. For Massachusetts, information from a state classification of content (Massachusetts Department of Education, 1995b) was used. In all cases, because each are based on only a single year's test, the classifications should be considered illustrative. Also, while the Texas content remained stable across the years of the study, it appears that the content of the Maine and Massachusetts assessments may have varied over time.

The state domains by State NAEP content classifications used for the 1996 assessments are presented in Tables 7.2 through 7.6. The tables show domain structures used by each state and the percentages of assessment of each category. For each state, a rough comparison of the distribution of its test's content with that of State NAEP is shown. Tables 7.2 and 7.3 suggest that at grade 4, Maine emphasized Data Analysis, Statistics, and Probability a bit more and Algebra and Functions less than did the State NAEP. At grade 8, Maine may have had slightly less emphasis on Number Sense, Properties, and Operations, and no items on Measurement. For Massachusetts, at grade 4, there was half the proportion of items covering Algebra and Functions as for the State NAEP. Otherwise, based on content classification, the measures seem comparable. At grade 8, Massachusetts seemed to place emphasis on Number Sense, Properties, and Operations and much less concentration than the State NAEP on Algebra and Functions.

Table 7.6 shows that, at grade 4, compared to the State NAEP, a greater proportion of TAAS items measured Number Sense, Properties, and Operations than did the State NAEP test, with the result that less emphasis was placed on the remaining NAEP content strands. Contrasting the subobjectives on the TAAS framework for the

two grade levels suggests that at grade 8, the content domain was more extensive than for grade 4 for the objectives of: Mathematical Relations, Functions and Other Concepts, Geometric Properties, Measurement Concepts, Probability and Statistics, and Determining Solution Strategies and Analyzing or Solving Problems.

All three states indicated that they had made changes in their testing programs in the late 1990s. Maine first administered the MEA in 1999. While it had purposes similar to its predecessor, it was based on a set of standards that were mandated by the Maine legislature in 1994.[6] Table 7.7 depicts the domain of the 1999 MEA mathematics assessment with respect to both the Maine standards and the State NAEP content strands. The Maine standards include a category called discrete mathematics that we were unable to classify with respect to the NAEP strands. At grades 4 and 8 each, only 5% of MEA items fell into this category. Based on framework category labels, the Maine MEA and

[6] http://www.state.me.us/education/lres/lres

Table 7.2
Classification of Maine and NAEP Mathematics Test Domains[1] (1993-1994, Gr.4)[2]

State Content Structure		NAEP Content Strands				
Content Strand	**Sub-strands**	**Number sense, properties & operations**	**Measurement**	**Geometry & Spatial Sense**	**Data analysis, statistics & probability**	**Algebra & functions**
Numbers and Numeration	***Numeration***	10%				
	Number Theory	4%				
	Operations-Whole Numbers	16%				
	Operations-Fractions	2%				
	Operations-Decimals	10%				
	Operations-Percent					
	Operations-Integers					
	Properties of Operations					
Variables and Relationships	***Equations/inequalities***					2%
	Functions/Coordinate Systems					2%
Geometry	***Plane and Solid Figures***			6%		
	Properties of Triangles					
	Spatial Visualization			6%		
	Perimeter. Area, and Volume			2%		
Measurement	***Using Instruments***		6%			
	Unit Equivalents		12%			
	Appropriate Units					
Problem- Solving Skills	***Estimation/ Reasonableness***		2%			
	Understanding Problem/Reasoning					2%
	Strategies	4%				
	Relevant Information					2%
Probability & Statistics					12%	

Percentage of items on MEA	46%	20%	14%	12%	8%
Percentage of items on NAEP[3]	40 – 70%	20%	15%	10%	15%

[1]No Cognitive Demand information available for MEA
[2]College Board (1996), MDE
[3]NAEP items may be classified in more than one category

Table 7.3
Classification of Maine MEA and NAEP Mathematics Test Domains[1] (1993-1994, Gr.8)[2]

State Content Structure		NAEP Content Strands				
Content Strand	**Sub-strands**	**Number sense, properties & operations**	**Measurement**	**Geometry & Spatial Sense**	**Data analysis, statistics & probability**	**Algebra & functions**
Numbers and Numeration	***Numeration***	10%				
	Number Theory	2%				
	Operations-Whole Numbers	10%				
	Operations-Fractions	4%				
	Operations-Decimals	6%				
	Operations-Percent	6%				
	Operations-Integers	4%				
	Properties of Operations					
Variables and Relationships	***Equations/inequalities***					12%
	Functions/Coordinate Systems					2%
Geometry	***Plane and Solid Figures***			6%		
	Properties of Triangles			4%		
	Spatial Visualization			4%		
	Perimeter. Area, and Volume			10%		
Measurement	***Using Instruments***					
	Unit Equivalents		4%			
	Appropriate Units					
Problem Solving Skills	***Estimation/ Reasonableness***					
	Understanding Problem/Reasoning Strategies					2%
	Relevant Information	2%				2%
Probability & Statistics					10%	
Percentage of items on MEA		44%	4%	24%	10%	18%
Percentage of items on NAEP[3]		25 – 60%	15%	20%	15%	25%

[1]No Cognitive Demand information available for MEA
[2]College Board (1996), MDE
[3]NAEP items may be classified in more than one category

Table 7.4

Percentage of Items on Massachusetts MEAP by NAEP Mathematics Test Domains[1] for Grade 4 in 1994 and 1996[2]

State Content Structure		**NAEP Content Strands**				
Content Strand	**Sub-Strand**	**Number sense, properties & operations**	**Measurement**	**Geometry & Spatial Sense**	**Data analysis, statistics & probability**	**Algebra & functions**
Numbers and Numeration	***Numeration***	8.2%				
	Number Theory	6.8%				
Operations	***Whole Numbers***	9.5%				
	Fractions & Decimals	7.2%				
	Decimals					
	Percent					
	Integers					
	Properties of Operations					
Variables and Relations						6.8%
	Algebraic Manipulations					
	Relations/Functions					
	Equations/Inequalities					
Measurement & Geometry	***Using Instruments***		6.8%			
	Units		6.8%			
	Perimeter, Area, Volume		2.3%			
	Plane/solids/figures			6.8%		
	Transformations/Spatial visualization			6.8%		
Problem Solving Skills	***Estimation and Reasonableness***	5.5%				
	Strategies	5.5%				
	Relevant Information	5.5%				
Probability & Statistics	***Probability***				6.8%	
	Statistics					
	Graphs, Tables, Charts				6.8%	

Percentage of items on MEAP	48%	16%	14%	14%	7%
Percentage of items on NAEP[3]	40 – 70%	20%	15%	10%	15%

[1]No Cognitive Demand information available for MEAP

[2]College Board (1996), MDE

[3]NAEP items may be classified in more than one category

Table 7.5

Percentage of Items on Massachusetts MEAP by NAEP Mathematics Test Domains[1] for Grade 8 in 1994 and 1996[2]

State Content Structure		NAEP Content Strands				
Content Strand	**Sub-strands**	**Number sense, properties & operations**	**Measurement**	**Geometry & Spatial Sense**	**Data analysis, statistics & probability**	**Algebra & functions**
Numbers and Numeration	***Numeration***	5.3%				
	Number Theory	5.3%				
Operations	***Whole Numbers***	6.0%				
	Fractions & Decimals (Gr. 4)	5.3%				
	Decimals	5.3%				
	Percent	5.3%				
	Integers	3.5%				
	Properties of Operations	1.4%				
Variables and Relations						7.4%
	Algebraic Manipulations					
	Relations/Functions					
	Equations/Inequalities					
Measurement & Geometry	***Using Instruments***		5.3%			
	Units		5.3%			
	Perimeter. Area, Volume		5.3%			
	Plane/solids/figures			5.3%		
	Transformations/Spatial visualization			5.3%		
Problem Solving Skills	***Estimation and Reasonableness***	5.3%				
	Strategies	5.3%				
	Relevant Information	5.3%				
Probability & Statistics	***Probability***				5.3%	
	Statistics					
	Graphs, Tables, Charts				5.3%	

Percentage of items on MEAP	53%	16%	11%	11%	7%
Percentage of items on NAEP[3]	25 – 60%	15%	20%	15%	25%

[1]No Cognitive Demand information available for MEAP

[2]NAGB (1996), MDE (1995)

[3]NAEP items may be classified in more than one category

Table 7.6
Classification of TAAS Items by TAAS and NAEP Mathematics Content Frameworks

TAAS Objectives and Instructional Targets			NAEP Content Strands				
Domain	**Objective**	**Sub-objective**	**Number sense, properties & operations**	**Measurement**	**Geometry & Spatial Sense**	**Data analysis, statistics & probability**	**Algebra & functions**
		Percentage of items on TAAS	**68%**	**12%**	**8%**	**12%**	**0%**
		Percentage of items on NAEP	**40 %**	**20%**	**15%**	**10%**	**15%**
Concepts	1.Understanding of number concepts *4 items-8%*	1.1 Translate whole numbers (name to numeral/numeral to name)	✔				
		1.2 Compare and order whole numbers					
		1.3 Use whole number place value	✔				
		1.4 Round whole numbers (to nearest ten or hundreds)	✔				
		1.5 Recognize decimal place value					
		1.6 Use odds, evens and skip counting	✔				
		1.7 Recognize and compare fractions using patterns and pictorial models					
	2.Mathematical relations, functions, & other concept *4 items-8%*	2.1 Use whole number properties and inverse operations	✔				
		2.2 Determine missing elements in patterns	✔				
		2.3 Use number line representations for whole numbers and decimals	✔ ✔				
	3.Geometric properties **4 Items 8%**	3.1 Recognize 2 and 3- dimensional figures and their properties			✔ ✔		
		3.2 Identify informal representations of congruence and symmetry			✔ ✔		
	4.Measurement concepts using metric & customary units **4 Items -8%**	4.1 Solve problems with metric and customary units and problems involving time (simple conversions; elapsed time)		✔ ✔			
		4.2 Find perimeter		✔			
		4.3 Find area (with grids)		✔			
	5. Probability and statistics **4 Items -8%**	5.1 Determine possible outcomes in a given situation				✔ ✔ ✔	
		5.2 Analyze data and interpret graphs (including line graphs)				✔	

Table 7.6 (continued)
Classification of TAAS Items by TAAS and NAEP Mathematics Content Frameworks

Domain	TAAS Objectives and Instructional Targets: Objective	Sub-objective	*NAEP Content Strands*: Number sense, properties & operations	Measurement	Geometry & Spatial Sense	Data analysis, statistics & probability	Algebra & functions
Operations	6. Use addition to solve problems (tenths and hundredths; using models) *4 Items-8%*	6.1 Add whole numbers and decimals	✔ ✔ ✔ ✔				
	7 .Use Subtraction to solve problems (tenths and hundredths; using models) *4 Items-8%*	7.1 Subtract whole numbers and decimals	✔ ✔ ✔ ✔				
	8. Use Subtraction to solve problems *4 Items-8%*	8.1 Multiply whole numbers	✔ ✔ ✔ ✔				
	9. Use Subtraction to solve problems *4 Items-8%*	9.1 Divide whole numbers (using multiplication facts)	✔ ✔ ✔ ✔				
Problem Solving	10. Estimate solutions to a problem situation *3 Items-6%*	10.1 Estimate with whole numbers	✔ ✔ ✔				
	11. Determine solution strategies and analyze or solve problems 4 Items-8%	11.1 Select strategies or solve problems using basic operations with whole numbers	✔ ✔	✔ ✔			
		11.2 Determine strategies or solve problems requiring the use of geometric concepts		✔			
	12. Solve problems using mathematical representation *4 Item -8%*	12.1 Formulate solution sentences	✔ ✔ ✔				
		12.2 Interpret charts, picture graphs, and bar graphs and use the information derived to solve problems				✔ ✔	
	13. Evaluate the reasonableness of a solution to a problem situation *3 Items-6%*	13.1 Evaluate reasonableness	✔ ✔ ✔				
Number of items			**34**	**6**	**4**	**6**	**0**

the State NAEP are quite similar in content. In comparison to its predecessors, the MEA has better balance across the NAEP categories. For example, at grade 4 in 1996, the MEA had no coverage of Algebra and Functions; in 1999, 21% of the items fell into this category.

Table 7.7
Percentage of Items on 1999 Maine MEA by NAEP Mathematics Test Domain

	NAEP Content Strands									
Maine Content Standards[1]	**Number Sense, Properties, & Operations**		**Measurement**		**Geometry & Spatial Sense**		**Data Analysis, Statistics & Probability**		**Algebra & Functions**	
	Grades		***Grades***		***Grades***		***Grades***		***Grades***	
	4	***8***	***4***	***8***	***4***	***8***	***4***	***8***	***4***	***8***
Number and Number Sense	15%	14%								
Computation	15%	11%								
Data Analysis & Statistics							12%	11%		
Probability							8%	11%		
Geometry					12%	13%				
Measurement			12%	10%						
Patterns, Relations, Functions									12%	15%
Algebra Concepts									9%	15%
Discrete Mathematics										
Percentage MEAP	30%	25%	12%	10%	12%	13%	20%	22%	21%	30%
Percentage NAEP	25-60%	25-60%	15%	15%	20%	20%	15%	15%	25%	25%

[1]Maine Department of Education, 1998-1999 Maine Educational Assessment Technical Manual

MCAS was first administered to Massachusetts' students in 1998. Tables 7.8 and 7.9 show the content of this test in terms of both the Massachusetts and State NAEP frameworks. MCAS appears to have a distribution of items among State NAEP strands that is similar to NAEP itself. In comparison to the earlier Massachusetts MEAP, there is less emphasis on Number Sense, Properties, and Operations and a larger proportion of items covering Algebra and Functions. The MCAS also categorized their items according to the mathematical thinking skill categories of Conceptual Understanding, Procedural Knowledge, and Problem Solving. There were twice as many Problem Solving items on the grade 8 MCAS as at grade 4. The distribution of thinking skills on grade 8 MCAS and State NAEP were similar. At grade 4, there were far fewer problem-solving items and a considerably larger number of items measuring Procedural Knowledge.

Beginning in the 1998–1999 school year, the TAAS was intended to reflect the revised statewide curriculum called the *Texas Essential Knowledge and Skills (TEKS).*[7] However, the following statement from the Technical Digest (Texas Education Agency, 1999a) suggests that the test domain itself did not actually change.

[7] http://www.tea.state.tx.us/teks/

Table 7.8
Percentage of Items on Massachusetts MCAS by NAEP Mathematics Test Domains[1] for Grade 4 in 1999[2]

State Content Structure		NAEP Content Strands				
Content Strand	**Sub-strand**	**Number sense, properties & operations**	**Measurement**	**Geometry & Spatial Sense**	**Data analysis, statistics & probability**	**Algebra & functions**
Number Sense	***Number and Number Relations***	5.1%				
	Concepts of Whole Number Operations	7.7%				
	Fractions & Decimals	7.7%				
	Estimation	5.1%				
	Whole Number Computation	12.8%				
Patterns, Relations, & Functions	***Patterns& Relations***					12.8%
	Algebra & Mathematical Structures					7.7%
Geometry & Measurement	***Geometry & Spatial Sense***			10.3%		
	Measurement		10.3%			
Statistics & Probability	***Statistics & Probability***				20.5%	

Percentage of items on MCAS	38%	10%	10%	21%	21
Percentage of items on NAEP[1]	25 – 60%	15%	20%	15%	25%

Distribution of MCAS and NAEP items by mathematical thinking skills[3,4]

Thinking Skill	**MCAS**	**NAEP**
Conceptual Understanding	40%	37%
Procedural Knowledge	40%	22%
Problem Solving	20%	41%

[1]No Cognitive Demand information available for MCAS
[2]NAGB (1996), MDE, 1999 Released MCAS Grade 4 Mathematics items, MCAS Technical Report
[3]NAEP items may be classified in more than one category
[4]MCAS Technical Report
[5]Allen et al., 1997

Table 7.9
Percentage of Items on Massachusetts MCAS by NAEP Mathematics Test Domains[1] for Grade 8 in 1999[2]

State Content Structure		NAEP Content Strands				
Content Strand	**Sub-strand**	**Number sense, properties & operations**	**Measurement**	**Geometry & Spatial Sense**	**Data analysis, statistics & probability**	**Algebra & functions**
Number Sense	***Number and Number Relations***	7.7%				
	Number Systems & Number Theory	7.7%				
	Computation & Estimation	7.7%				
	Ratio, Proportion, Percent	5.1%				
Patterns, Relations, & Functions	***Patterns& functions***					12.8%
	Algebra					12.8%
Geometry & Measurement	***Geometry***			12.8%		
	Measurement					
	Geometric Measurement		12.8%			
Statistics & Probability	***Statistics***				10.3%	
	Probability				10.3%	
Percentage of items on MCAS		28%	13%	13%	21%	26
Percentage of items on NAEP[3]		25 – 60%	15%	20%	15%	25%

Distribution of MCAS and State NAEP items by mathematical thinking skills[4,5]

Thinking Skill	MCAS	NAEP
Conceptual Understanding	30%	38%
Procedural Knowledge	25%	25%
Problem Solving	45%	37%

[1]No Cognitive Demand information available for MCAS
[2]NAGB (1996), MDE, 1999 Released MCAS Grade 8 Mathematics items, MCAS Technical Report
[3]NAEP items may be classified in more than one category
[4]MCAS Technical Report
[5]Allen et al., 1997

> " . . . test items were matched to TEKS, and TAAS assessed only those areas common to the TEKS and the Essential Elements." (Texas Education Agency, 1999a, p. 2)

This statement implies that while items aligned to the earlier curriculum framework (called the Essential Elements[8]) but not aligned to the TEKS would be eliminated from the test domain and that no items that aligned with TEKS but not with the Essential Elements would be added to the TAAS domain. Thus, it seems fair to conclude that the test domain of TAAS remained essentially unchanged from 1994 through 2000.

In summary, the domains of the earlier versions of Massachusetts and Maine tests are difficult to pin down. It seems likely that Maine's domain changed somewhat from year to year. With the creation of MCAS and the MEA, the domains became much more clearly defined and the adopted test development procedures lend credibility to claims for comparability across years. TAAS domains and assessment content were stable across all the years of the study.

Technical Characteristics

There were considerable differences in description of technical characteristics among the various assessments. Little technical information was available about the MEAP in Maine. Massachusetts supplied some description of validity, reliability, and bias review procedures. The MCAS, TAAS, and the MEA in Maine had more extensive documentation on these topics, as well as equating information.

Massachusetts reported reliability coefficients of .97 and .98 on its grade 4 and grade 8 MEAP instruments, respectively. The coefficients were obtained based on the average of alpha coefficients of the various forms. Massachusetts' study of MEAP validity was limited to procedures for aligning items to the perceived state curriculum. MEAP documentation stated that the committees transformed judgments about "what would be appropriate to teach at or prior to the grade level tested" (Massachusetts Department of Education, 1995a) into the set of mathematics items the comprised the item domain. Beyond this item selection criterion, there were no estimates of content, or any other type, of validity offered. Massachusetts checked for bias by asking item developers to avoid the inclusion of items that, it their opinion, would advantage specific genders or races. No discussion of reliability, validity, or bias review for the original Maine MEA could be located.

The TEA reported Kuder Richardson–20 reliability coefficients for the total group of .89 and .92 for grades 4 and 8, respectively (Texas Education Agency, 1999a). TEA cites committee processes of aligning TEKS with TAAS frameworks as evidence of content validity (Texas Education Agency, 1999a, pp. 46-47). However, because, the domain of TAAS was limited in recent years to content common to TEKS and the earlier

[8] Texas Administrative Code, Chapter 75 curriculum, subchapter B-D

Essential Elements frameworks, the Texas mathematics test must be considered to have reduced content validity since the introduction of TEKS. Data were presented for one small criterion-related validity study for TAAS. It yielded a correlation of .32 between TAAS mathematics scale scores and end-of-year grades in 1992 and 1993.

Both Maine and Massachusetts made significant advances in their documentation of procedures that made judgments possible about the technical quality of the MEA and MCAS. For MCAS, Cronbach's αs for grade 4 and grade 8 1998 mathematics tests were .87 and .91, respectively. Several procedures were used to estimate the statistical accuracy and consistency of proficiency scale classification (Massachusetts Department of Education, 1998b; 1999a). In 1998, estimates of the consistency of classification into proficiency categories based on two parallel forms of the MCAS mathematics tests were .68 and .71 at grade 4 and grade 8, respectively. The rigorous implementation of a systematic test-development process was offered as evidence for the alignment of MCAS instruments with the Massachusetts standards. A concurrent validity coefficient between the grade 4 MCAS and the Stanford Achievement Test mathematics score at that grade for a single district was .69. Results of studies indicating the similarity between gender and ethnic group performance differences on MCAS and the two commercial achievement tests are further evidence of concurrent validity. Differential Item Functioning (DIF) procedures were used to identify potential male-female and Black-White bias among test items. In 1998 for gender, DIF values were negligible for 93% and 88% of the items at grades 4 and 8, respectively. The corresponding value for race was 80% for each grade.

Maine's 1998–1999 *Technical Manual* (Maine Department of Education, 2001) offers Cronbach's αs for grade 4 and grade 8 1998 MEA mathematics of .80 and .85, respectively. Indicators of the consistency and accuracy of level classification on the proficiency were reported. The accuracy statistic indicated that students were properly classified in 75% of the cases in grade 4 and 79% in grade 8. Estimates of the consistency of classification from parallel forms were .67 and .71 for grades 4 and 8, respectively. Evidence of content validity is based on test development processes for aligning MEA and Maine's Curriculum Framework. DIF analysis, by gender, of the MEA mathematics tests at grades 4 and 8 yielded negligible values for 83% and 73% of the items at grades 4 and 8, respectively. There was no report on the disposition of those items that functioned differentially.

The TEA developed two sets of equating procedures (Texas Education Agency, 1999a, p. 49-53) for TAAS. The first aimed to maintain comparable content from year to year, while the second set of strategies was designed to keep the level of difficulty of tests constant. The first criterion is achieved by aligning items based on content specifications, while the second, difficulty equivalence, is accomplished through a two-step pre- and post-equating strategy. These equating procedures use 1994 results as the base year. Pre-equating is done by selecting items for new tests that, with respect to their content specifications and Rasch difficulty values, are quite similar. Post-equating uses the results from the actual administration to, if deemed necessary, recalibrate difficulties to the 1994 scale in order to allow consistent comparison across years. The Massachusetts

Department of Education described its equating procedures as an "anchor-test-nonequivalent groups design." The application of this design is detailed in the Department's *Technical Report* (Massachusetts Department of Education, 1999c, pp. 68-71). Probably because the latest available documentation of Maine's assessment pertains to 1998–1999, the initial year of MEA, the state has yet to report equating procedures used for the MEA.

Data

The NAEP offers extensive data and information to qualified researchers. Demographic information about schools, communities, students, and their families is available. Nearly all of the data associated with sampling and data gathering are furnished (Allen et al., 1997). Data and information available for the MEAP (Masssachesetts) and MEA (Maine) were limited. Because of item sampling and the non-release of most items, as well as lack of clarity about actual sampling procedures, it was not possible to understand the content structure represented by data. These issues limited evaluation of comparability across years. While Massachusetts did provide information about the characteristics of communities, it was not in a form that proved particularly useful for comparative analysis of NAEP data. Both the MCAS (Massachusetts) and the most recent MEA (Maine) are designed to overcome many of these data deficiencies. While Texas offered student-level data with indicators for gender, race, and economic disadvantage, as well as substantial information about scaling and equating, the state was not able to provide linking parameters that would have permitted the evaluation of test comparability from year to year at each grade level. From our research perspective, it would have been desirable for TAAS to emulate NAEP and establish a common scale that spanned grades 3 through 8. This would have allowed for more powerful cohort analysis.

This study of the impact of three SSIs illustrates the use of data from large-scale educational assessments for a purpose other than that for which they were originally intended. Not surprisingly, such an effort quickly identifies desirable but unavailable data elements. As research on student achievement and on state assessment programs becomes more crucial, existing state and national databases are being sought as sources of data. Issues raised regarding the utility of these sources and the means of making data useful for such unintended but important purposes are being considered (Turnbull, Grissmer, & Ross, in Grissmer & Ross, 2000). Research to extend the understanding of the correlates of achievement would be aided if states, as well as the federal government, were to consider the requirements of such research as they design their assessment systems. Collaboration in this regard would be helpful in providing necessary comparative research data.

It would be well to consider the nature of the data that are required for cohort and cross-sectional studies. Standardization of definitions would enhance the quality of comparative studies. Some states provide extensive data via the Internet. However, for research purposes, gaps remain. The cooperative introduction of common indicators, as well as data structures, would facilitate the use of state assessment data for research

purposes. Further, states also differ in terms of their willingness to allow access to their data. It would be helpful for states to institute common criteria and procedures for permitting the use of data for research purposes. A good first step toward achieving standardized state practices could be promoted by a joint CCSSO/AERA/NCME (National Council on Measurement in Education) taskforce charged with making recommendations regarding appropriate procedures.

Documentation

While none of the states offered researchers written information about policies or procedures for data access and acquisition, staff members from each of the three state education agencies were helpful in providing data and other information, as well as answering questions to clarify voids, or vagueness, in available documentation regarding their respective assessment systems. The states did differ, however, in the amount of detail that was available about their programs, their processes, and procedures. Also, for Maine and Massachusetts, there were substantial improvements in the breadth and depth of the documentation provided for MEA and MCAS, as compared to the earlier assessment systems in those states. In particular, the congruence of the contents of the MCAS Technical Reports (Massachusetts Department of Education, 1999a) with the *Standards for Education and Psychological Testing* (American Educational Research Association, 1999) and the *Code of Fair Testing Practices* (American Psychological Association, 1999) provides an excellent example for other states.

For the original MEAP (Massachusetts Department of Education, 1995a, 1995b, 1996), there was considerable information about scoring and scaling; however, little attention was paid to item sampling, inclusion/exclusion, and equation. For the original MEA (Maine Department of Education, 1989), there was little information concerning the details of design or technical characteristics. The Texas Educational Agency offers extensive information and data about assessment and accountability availability through its Internet site. As referenced earlier in this report, the *Technical Digest* (1999a) provided substantial information about factors affecting the technical quality of the data, such as test development, reliability, validity, equating and scaling methods, and procedures. However, there are a few areas where information is lacking. For example, with regard to equating, while *p*-value information was available for the released items, no item data relating to Rasch scaling could be obtained. So, in the case of linking items, no parameters are reported. In fact, the TEA claims that these values are proprietary.

Analysis and Findings[9]

State Assessment data were the focus of two types of study. The first was a comparison of the trends exhibited in the state results with those evident in the State NAEP. The second was a comparison of state assessment results for students or schools

[9] More detailed information about the state assessment studies may be found in Kaufman (2002).

connected with SSI interventions within a state with those not involved in SSIs. Data from all three states were subjected to comparison of state performance trends with those of the State NAEP. Since, in Texas, it was not possible to determine the level of SSI intervention at the local level, only for Maine and Massachusetts were SSI versus non-SSI comparisons conducted.

Comparison of State Assessment and State NAEP Trends

The primary procedure used to compare state assessment and State NAEP trends was comparison of the effect size (Cohen, 1969; Klein et al., 2000) of changes across programs. Effect sizes were calculated as:

$ES = \frac{\mu_1 - \mu_2}{s_{pooled}}$, where μ_1 and μ_2 are group means and the pooled varience is obtained by :

$s_{pooled} = \sqrt{\frac{(n_1 - 1)(s_1) + (n_2 - 1)(s_2)}{n_1 + n_2 - 2}}$, with n_1 and n_2 being sample sizes and s_1 and s_2 standard deviations.

Cohen (1969) suggested that, in general, effect-sizes of .2 were small, of .5 were medium, and of .8 were large.

For Maine and Massachusetts, effect sizes were calculated for scale score mean changes in mathematics results between 1992 and 1996. In the case of Texas, the data allowed study of changes between 1994 to1996, as well as from 1996 to 2000. Table 7.10 shows the results of the comparisons for the three states. These results, viewed altogether, show that in some cases state and NAEP trends agreed and in other cases they did not. For example, NAEP and state assessment results for Massachusetts and Texas (1994–1996) indicated similar trends, while Maine's data for the same period did not. The small-to-medium improvements suggested on MEA between 1992 and 1996 are not replicated by State NAEP data. Likewise, for Texas over the 1996–2000 period, the small-to-medium improvement shown by TAAS did not show up in the NAEP data for the same period. TAAS data allowed for study results by race and SES. Figures 7.2 and 7.3 depict an achievement gap on the basis of race. Table 7.11 shows a gap by economic disadvantaged category. These figures, examined together with Table 7.11, suggest that a very slight gap closing for both race and economic disadvantage may have occurred at 4th grade between 1994 and 1996, as well as between 1996 and 2000. These same figures and tables suggest a similar reduction in the gap at grade 8 between 1992 and 1996 and a larger impact between 1996 and 2000. However, the comparison of the effect sizes for Texas changes for race as compared with those found on State NAEP, which are shown in Table 7.12, suggest no gap-closing during the 1992–1996 period and only a slight advantage for grade 4 minorities at grade 4 over the 1996–2000 period. One possible explanation for these differences between State NAEP and TAAS results may be attributable to the better alignment of TAAS with the content areas that were emphasized

Figure 7.2. Trends in TAAS grade 4 TLI means by ethnic category.

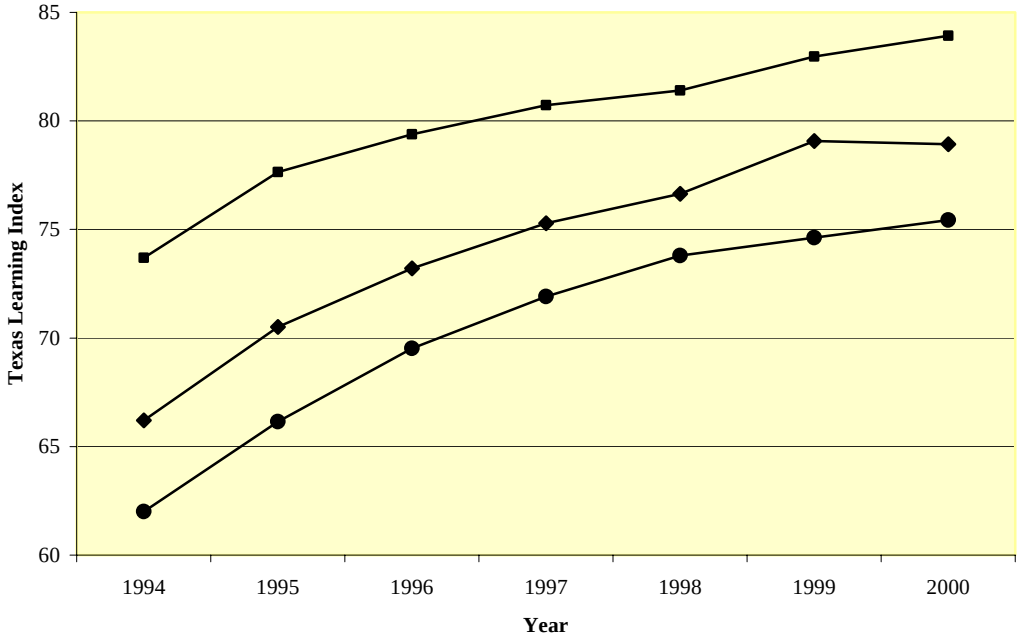
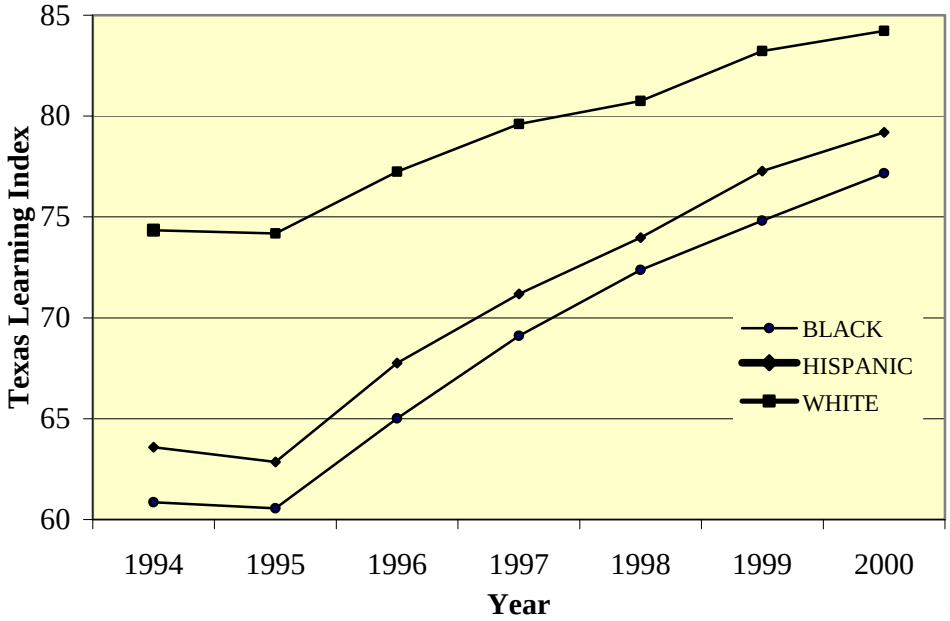


Figure 7.3. Trends in TAAS grade 8 TLI means by ethnic category.



in the instruction of student minorities. Another explanation might arise from the greater incentives to perform that are associated with TAAS.

Table 7.10
Comparison of Effect Sizes for State Assessments and State NAEP Across Massachusetts, Maine, and Texas

	Massachusetts		***Maine***		***Texas***			
	1992 -1996		***1992 -1996***		***1994 -1996***		***1996 -2000***	
	MEAP	***NAEP***	***MEA***	***NAEP***	***TAAS***	***NAEP***	***TAAS***	***NAEP***
Grade 4	-.02	.1	.4	.02	.4	.4	.3	.1
Grade 8	-.1	.1	.3	.2	.2	.2	.6	.1

An apparent explanation for Maine's 1992–1996 and Texas'1996–2000 results, in which MEA and TAAS showed larger gains than did State NAEP, is that the state assessments were better aligned with curriculum that guided instruction in the respective states. Content differences between the assessment systems of these states and State NAEP were discussed earlier. The stakes associated with TAAS suggest another reason that Texas students might show greater improvement on TAAS than on the State NAEP. While NAEP offers little incentive for students to perform well, the school and student consequences associated with TAAS are likely to motivate most students to do their best on the examinations.

SSI versus Non-SSI Comparisons

Both Maine and Massachusetts were able to identify schools that were the focus of SSI efforts. Massachusetts identified districts involved in SSI activities by their initial year of participation (D. Perda, personal communication, 2000). There were four cohorts, labeled Cohort 1, Cohort 2, Cohort 3, and Cohort 4, which corresponded to initial participation in the 1992–1993, 1993–1994, 1994–1995, and 1995–1996 school years, respectively. Table 7.14 provides summary statistics for these cohorts, as well as the effect sizes for 1992–1996 and 1998–1999 scale score changes within each cohort. It is instructive to know something of the characteristics of the students in the various cohorts. Figures 7.4 through 7.7 show that schools in the first cohort scored substantially lower than any of the other groupings of schools. Furthermore, non-SSI schools seem to score consistently higher than any of the SSI cohorts. With regard to socio-economic status (SES), Figures 7.8 and 7.9 suggest that the first cohort had a substantially larger proportion of students receiving free or reduced-cost lunch than did the other cohorts. In fact, if this index is accepted as a sound indicator of SES, one can conclude that most of the lower-SES schools were located in the first two SSI cohorts. Such a focus on lower-SES students would be consistent with the National Science Foundation's sixth driver (Shields et al., 1998) for the SSIs that called for a reduction in the achievement gap between traditionally underserved students and their peers.

Table 7.11

Scale Score Means and Standard Deviations for TAAS in 1994 through 2000, and Effect Sizes for 1994-1996 and 1996-2000

	1994		1995		1996		1997		1998		1999		2000		Effect Size [1]	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	94-96	96-00
Grade 4																
Total	69.9	15.2	74.0	14.0	76.2	13.7	77.9	12.7	78.9	11.9	80.6	10.6	81.0	10.9	.4	.3
Gender																
Male	69.5	15.4	74.0	14.2	76.2	13.9	77.9	12.9	78.7	12.1	80.5	10.4	81.2	10.9	.4	.3
Female	70.2	15.0	74.0	13.9	76.2	13.5	77.8	12.5	79.0	11.6	80.6	10.1	80.8	10.9	.4	.3
Race																
Black	62.0	16.2	66.1	15.4	69.5	15.7	72.0	14.8	73.8	13.6	74.6	12.8	75.4	13.3	.5	.4
Hispanic	66.2	15.6	70.5	14.6	73.2	14.6	75.3	13.7	76.6	12.7	79.1	10.7	78.9	11.8	.5	.4
White	73.7	13.3	77.6	11.9	79.4	11.5	80.8	10.5	81.4	10.0	83.0	8.2	83.9	8.2	.4	.3
Economic Disadvantaged																
Yes	64.7	15.6	69.0	15.1	71.8	15.1	74.0	14.2	75.5	13.2	77.7	11.6	78.0	12.3	.5	.5
No	74.1	13.3	78.0	11.7	80.0	11.1	81.3	10.1	82.0	9.6	83.2	8.1	84.1	8.3	.4	.3
Grade 8																
Total	69.3	15.4	68.8	15.0	72.6	14.9	75.5	14.0	77.4	12.7	80.1	11.0	81.5	9.2	.2	.6
Gender																
Males	69.0	16.0	68.9	14.5	72.2	15.3	75.3	14.4	76.9	13.4	79.7	11.6	81.5	9.9	.2	.6
Female	69.5	14.8	68.7	15.3	72.9	14.5	75.7	13.6	77.9	11.8	80.5	10.4	81.5	9.1	.2	.6
Race																
Black	60.9	15.3	60.6	14.4	65.0	15.5	69.1	15.0	72.4	13.7	74.9	12.5	77.2	10.8	.3	.8
Hispanic	63.6	15.4	62.9	14.6	67.8	15.4	71.2	14.7	74.0	13.5	77.3	11.9	79.2	10.2	.3	.8
White	74.3	13.3	74.2	12.8	77.2	12.6	79.6	11.8	80.8	10.6	83.2	8.9	84.2	7.7	.2	.5
Economic Disadvantaged																
Yes	62.5	15.5	62.1	14.7	66.8	15.5	70.3	15.0	73.3	13.8	76.5	12.1	78.5	10.5	.3	.8
No	72.7	14.2	72.5	13.7	76.2	13.3	78.8	12.3	80.2	11.0	82.5	9.6	83.6	8.2	.2	.5

[1]SD=15. See text for explanation

Table 7.12
Comparison of Effect Sizes of Change by Race and Economic Disadvantage Categories for State NAEP and TAAS

		State NAEP		TAAS	
		92-92	***96-00***	***94-96***	***96-00***
Grade 4					
Race					
	Black	.5	.3	.5	.4
	Hispanic	.3	.3	.5	.4
	White	.5	.1	.4	.3
Economic Disadvantaged					
	Yes		.2	.5	.5
	No		.1	.4	.3
Grade 8					
Race					
	Black	.2	.1	.3	.8
	Hispanic	.2	.2.	.3	.8
	White	.2	.1	.2	.5
Economic Disadvantaged					
	Yes		.2	.3	.8
	No		.1	.2	.5

Table 7.13

Scale Score Means and Standard Deviations for Massachusetts SSI Cohorts in 1990 Through 2000, and Effect Sizes for 1992-1996 and 1996-2000

		MEAP						MCAS						Effect Size	
		1990		1992		1994		1996		1998		1999			
	Cohort	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	92-96	98-99
Grade 4															
	92-93	1226.3	106.2	1237.1	108.5	1248.4	118.9	1261.2	121.5	227.8	9.1	229.9	8.6	.2	.2
	93-94	1283.3	92.8	1298.7	93.7	1300.4	92.7	1302.6	91.5	233.7	8.9	235.0	9.0	.04	.1
	94-95	1346.7	95.12	1361.6	93.78	1358.9	97.7	1347.3	83.5	237.6	8.4	239.5	8.4	-.2	.2
	95-96	1322.0	100.8	1343.3	94.3	1335.6	93.3	1322.5	85.2	237.2	9.0	237.9	8.1	-.2	.1
	Non-SSI	1342.6	107.3	1363.0	104.2	1363.1	95.8	1355.1	87.4	238.1	9.2	239.2	8.8	-.1	.1
Grade 8															
	92-93	1202.2	106.9	1216.0	102.9	1200.7	103.4	1221.6	94.7	218.1	11.2	219.9	10.9	.1	.2
	93-94	1275.7	93.2	1300.8	91.1	1308.9	70.7	1301.6	80.8	226.1	11.3	225.4	10.7	.01	-.1
	94-95	1335.0	91.1	1354.9	99.4	1354.4	84.1	1345.4	80.2	232.2	10.4	231.4	9.7	-.1	-.1
	95-96	1328.1	118.5	1358.1	108.2	1339.7	103.4	1336.8	87.4	232.8	10.0	231.4	10.1	-.2	-.1
	Non-SSI	1346.0	102.9	1367.6	106.2	1352.3	87.1	1351.2	84.2	233.9	10.7	232.5	10.7	-.2	-.1

Figure 7.4. Grade 4 MEAP cohort trends.

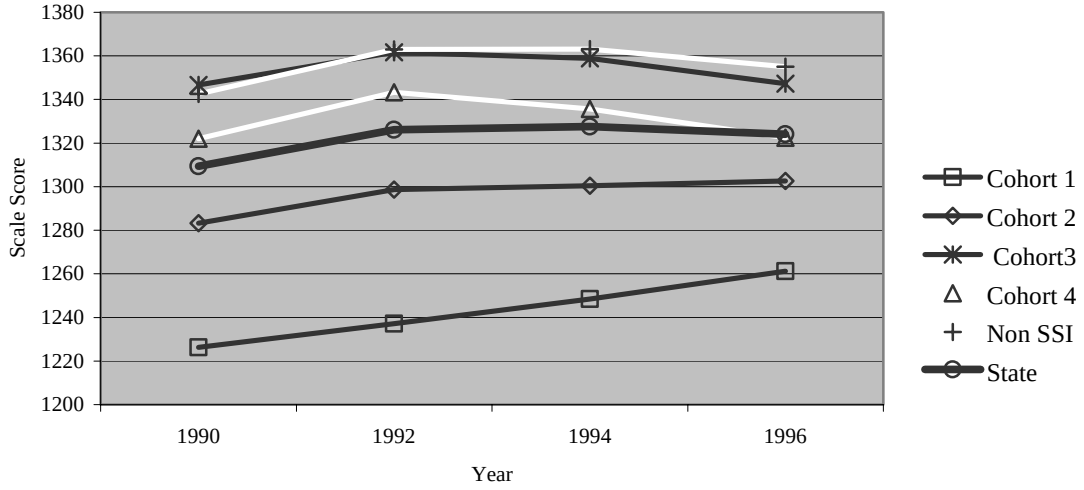


Figure 7.5. Grade 8 MEAP cohort trends.

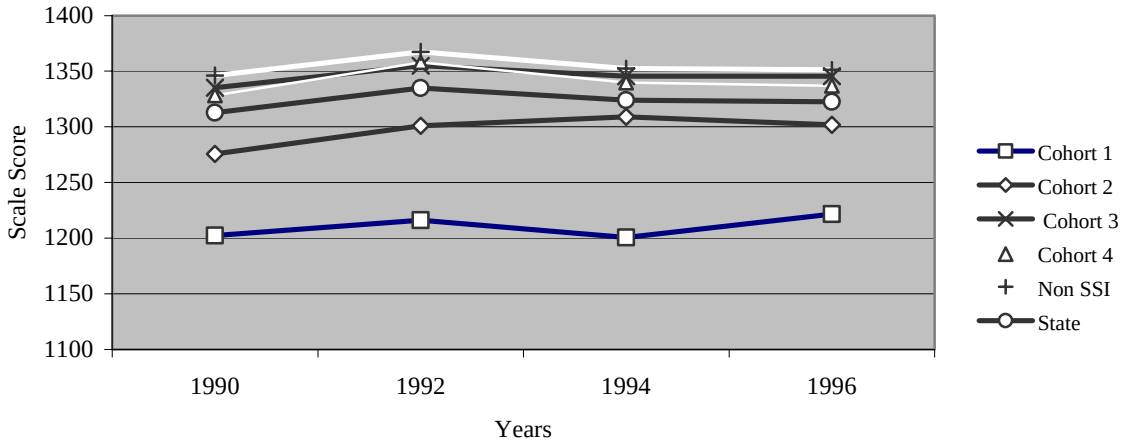


Figure 7.6. Grade 4 MCAS SSI cohort trends.

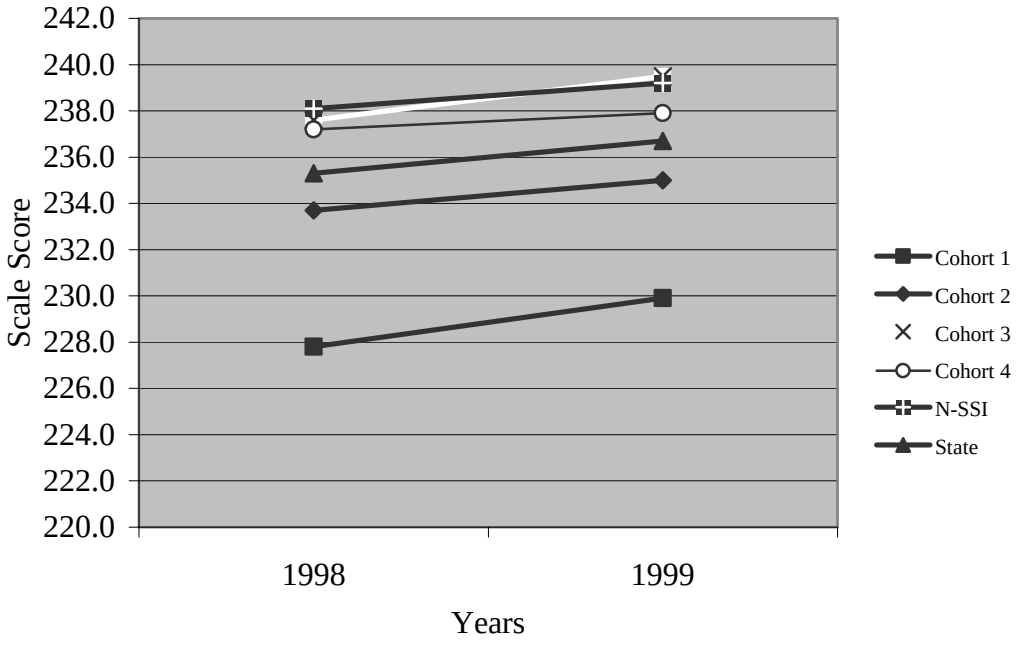


Figure 7.7. Grade 8 MCAS SSI cohort trends.

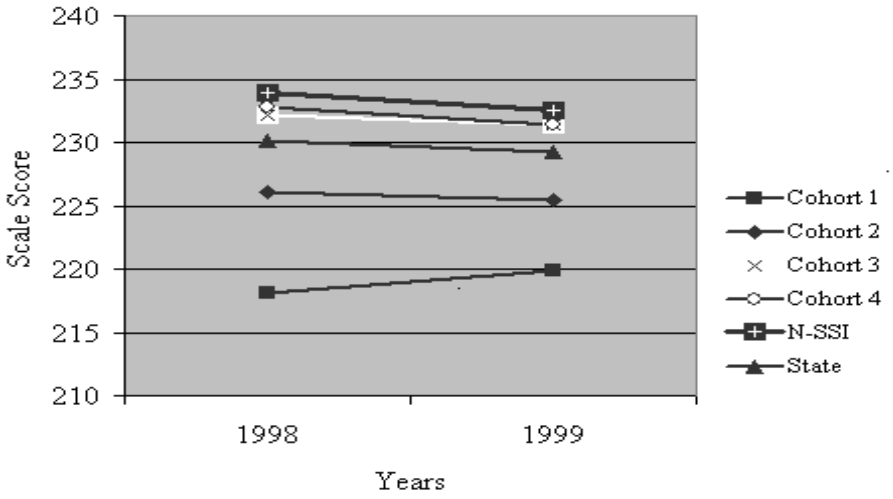


Figure 7.8. Proportion of grade 4 free and reduced-cost lunch students in four Massachusetts SSI cohorts and the non-SSI category.

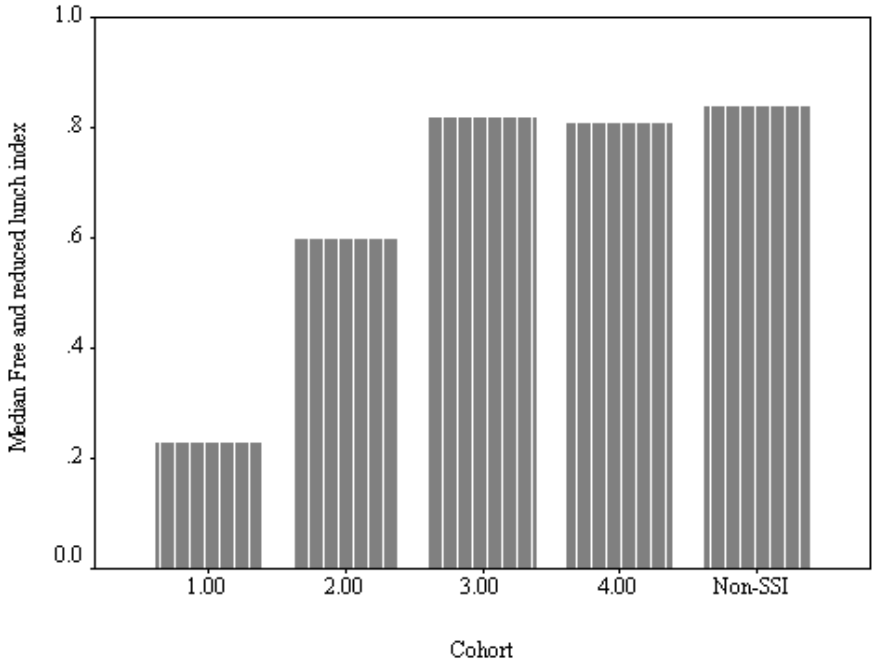
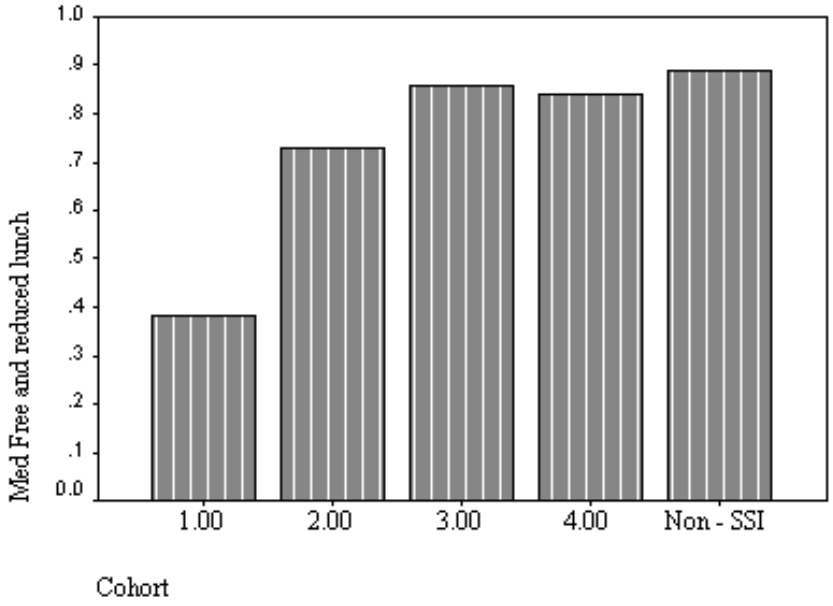


Figure 7.9. Grade 8 SSI free and reduced-cost lunch students in four Massachusetts SSI cohorts and the non-SSI category.



Although, at best, the effect sizes for each of the SSI cohorts are small, the differences among the change indicators are of some interest. Figures 7.10 and 7.11, which show changes in MEAP scale scores between 1992 and 1996, allow for concentration on the contrasts in achievement patterns between Cohort 1, the other SSI cohorts, non-SSI schools, and the state as a whole. Cohort 1 shows small improvement at both grade 4 and grade 8, while the other groups show negligible change or, for Cohorts 3 and 4, declines in achievement level. These patterns suggest that for the initial cohorts, which represents the greatest proportion of underserved youth, the extent of the gaps shown in comparison with their more advantaged peers were decreased.

Figure 7.10. MEAP grade 4 SSI cohort changes between 1992 and 1996.

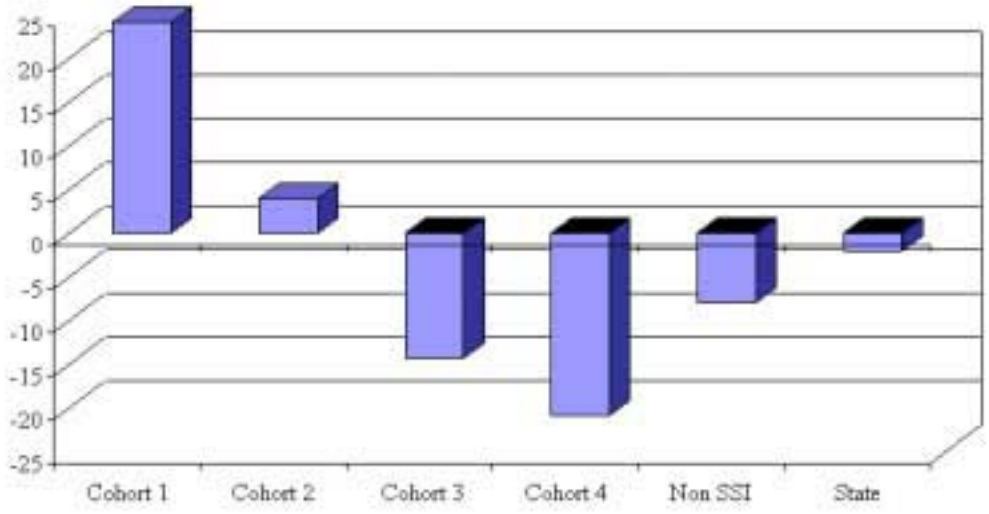
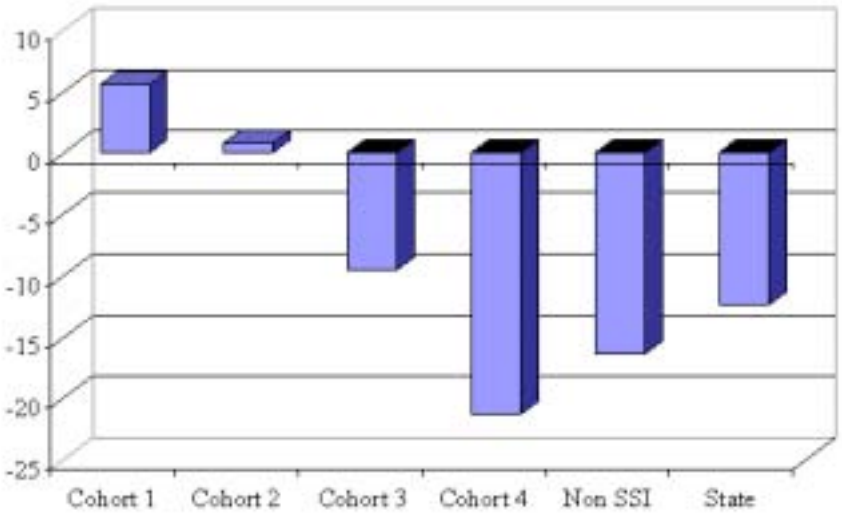


Figure 7.11. MEAP grade 8 SSI cohort changes between 1992 and 1996.



In Maine, a group of "Beacon Schools" (Schultz, 1994) was the focus of SSI efforts. These target schools were located in areas with a high proportion of underserved students. The initiative had five components: professional preparation and development; higher education cooperation and collaboration; community involvement, curriculum, instruction, and development; and, systemic planning and evaluation (Schultz, 1994). Figure 7.12 indicates that the SSI Beacon Schools had a higher proportion of minority students than their non-SSI cohorts.[10] At grade 8, Table 7.13 shows the effect sizes for the SSI and non-SSI populations over the 1992–1996 periods. Using Cohen's labels (Cohen, 1969), changes in both groups were small to medium, the grade 4 SSI group being slightly higher and the grade 8 cohort slightly lower. The magnitude of either difference suggests a meaningful advantage for either group.

Figure 7.12. Percentages of minorities participating in the 1996 MEA by SSI status.

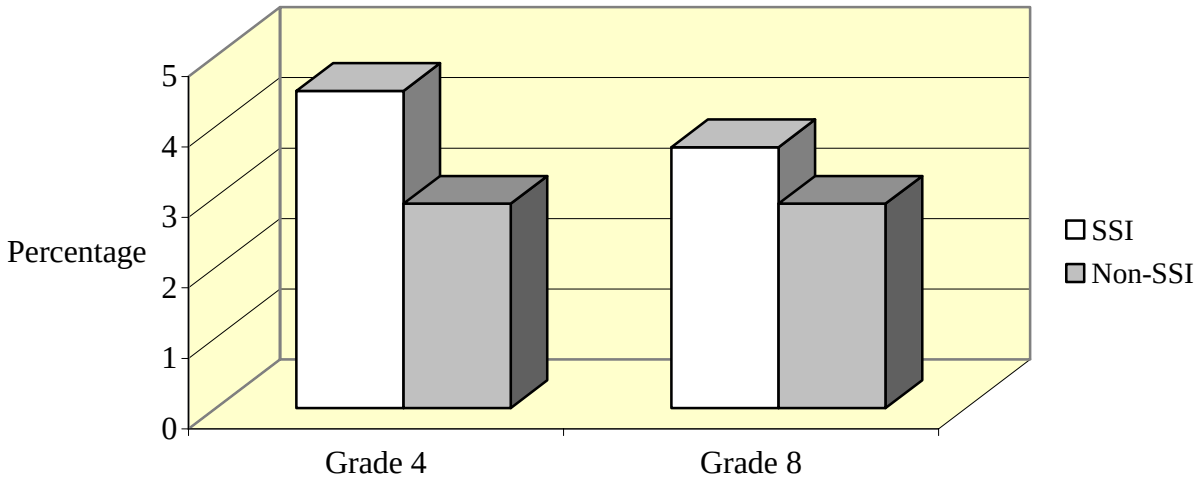


Table 7.14
Maine and State NAEP Effect Size for 1992 – 1996

	MEA					***State NAEP***		
	Total	***SSI***	***Non-SSI***	***Male***	***Female***	***Total***	***Male***	***Female***
Grade 4	.4	.5	.4	.4	.4	.02	.1	-.02
Grade 8	.3	.2	.3	.3	.3	.2	.2	.2

[10] X^2 =9.09, df=1, p=. 003 at grade 4. X^2=2.07, df=1, p=. 157 at grade 8

Summary

The three state assessments used for this study of SSI impact illustrate the diversity among state programs. For our purposes, TAAS provided the most useful information. Its technical quality, comprehensive documentation, and consistency over a number of years allowed for a more meaningful use of its data when making a comparison with State NAEP results. Until recently, the Massachusetts and Maine programs lacked year-to-year data and documentation that would allow trustworthy longitudinal studies. The primary reasons for this were the lack of evidence for either content or statistical comparability across years and the absence of student-level item data that could be used to calculate scale scores. The item-sampling designs, as well as the low-stakes purposes of these programs, no doubt contributed to these shortcomings. In fairness, it needs to be mentioned that, in the design of their initial assessments, neither the Maine nor Massachusetts departments of education set a priority on producing the type of data that were required for our SSI Impact Study analysis. Both states have greatly improved the utility of their data with MCAS (Massachusetts) and the "new" MEAP in Maine. When it comes to assessment program documentation, the MCAS model, which follows the schema of the AERA/APA/NCME *Standards for Educational and Psychological Testing* (1999), is exemplary. This sort of comprehensive design and technical information, together with the transparent availability of student-level item information, should make state data much more useful for research purposes.

A primary objective for including the study of state assessment results in this SSI Impact Study was to identify multiple indications of mathematics achievement in a few states. It was of interest to learn whether NAEP and state results were congruent, as well as to identify possible reasons for divergent results. Two anticipated explanations for differences in NAEP and state assessment trends were differences in student test-taking motivation and test content. Particularly with respect to content, Mehrens (2000) has observed that the validity of state tests is best judged by their alignment with state curriculum frameworks or standards. Insofar as motivation is concerned, TAAS, because of its stakes, should be assumed to provide greater motivation for students than the State NAEP or either the Massachusetts or Maine assessments of the early to mid-90s.

For the 1992–1996 periods, Texas and Massachusetts State NAEP and state assessment results were comparable, while Maine MEAP showed achievement improvements that were not found in NAEP results. Only the TAAS data allowed comparison with NAEP over the 1996–2000 period. TAAS results indicated substantially more improvement than did NAEP. This raises the question of the comparability of the assessments. Even though the TAAS is more focused on the basics of Numbers and Operations, it is well aligned to the Texas curriculum (Mehrens, 2000). This suggests that it is a sound indicator of the intended curriculum of the Texas public schools. Based on this reasoning, one would conclude that the TAAS results are a better indicator of the impact of policy and instructional interventions that took place as a result of SSI than State NAEP. Also, because of the stakes associated with TAAS, one may conclude that TAAS is a more accurate reflection of true achievement than the State NAEP. Because of the introduction of graduation and promotion requirements, it is to be expected that

perceived student stakes increased in the latter half of the previous decade. Looking back to TAAS/NAEP comparisons earlier in the decade, it should be recalled that NAEP gains over four years were compared to TAAS improvement over half of that interval. Had it been possible to examine TAAS data going back to 1992, comparisons of NAEP and TAAS might have been similar to the 1996–2000 period. A reasonable conclusion, then, would be that Texas students have shown considerable improvement in mathematics achievement during and after SSI implementation in their state. It is also fitting to point out that the greater levels of improvement on TAAS compared to NAEP during most of the decade would suggest that the enacted mathematics curriculum in Texas is narrower than the national consensual content reflected in the State NAEP.

Similarly, in the case of Maine, one might argue that the MEA results best reflect the enacted curriculum in Maine and, thus reflect the state's SSI interventions. This argument is weakened somewhat by the fact that, during the period from 1992 to 1996, there was no clear-cut state mathematics curriculum on which to base the content of the MEA assessments. The within-state study of the impact of SSI in Maine yielded scant, ambiguous results. In grade 4, between 1992 and 1994, the SSI cohort improved slightly more than cohorts in the non-SSI group. However, at grade 8, the converse was true. While neither NAEP nor MEAP showed improvement in Massachusetts mathematics performance, it should be noted that the cohort most impacted by SSI was small. In fact, within-state data indicate a positive differential in performance between this group and the state's non-SSI cohort at both grade levels.

Part of the rationale for using state assessment information to study SSI impact was its potential for extending 1996 NAEP results. With the extension of the SSI Impact Study project and the availability of 2000 State NAEP data, it would have been desirable to compare NAEP and state assessment gains between 1996 and 2000 for all three states. Unfortunately, the lack of equating studies between earlier and current assessment systems in both Maine and Massachusetts limited the value of extension studies and made 1995–2000 comparisons impossible. Perhaps the most important insight gleaned from this extensive study of state assessment information and its comparison with NAEP data relates to the merit of equating within-state assessment systems. At a minimum, high-quality horizontal studies are essential if the data they yield are to be of value for longitudinal studies. Sound vertical studies are necessary for cohort research across grade levels. Another nearly equally critical ingredient for state assessments is a comprehensive description of content domain that is stated in terms of knowledge and cognitive skills, as well as academic and real-world tasks to be performed. All three of the states studied are making strides in both equating and domain definition. This should have the effect of making test results more meaningful for longitudinal accountability schemes such as that required by *The No Child Left Behind Act of 2001* as well as increasing their usefulness to researchers. This should have the effect of making test results more meaningful for longitudinal accountability schemes such as that required by the *No Child Left Behind Act,* as well as increasing their usefulness to researchers.

CHAPTER 8

SSI AND NON-SSI ACHIEVEMENT USING STATE NAEP DATA: EMPIRICAL BAYES AND BAYESIAN ANALYSES

Introduction

State NAEP reported mathematics scale scores of each state that participated in the assessment years for grades 4 and 8 in 1992, 1996, and 2000. One of the advantages of NAEP scale scores is that they are comparable across data collection years and across grades (Allen et al., 1997). This enables us to compare mathematics score means of states in each assessment year and to assess the trends of states participating in consecutive assessment years. The main purpose of this chapter is to examine the differences between SSI and non-SSI states in mathematics scale scores using longitudinal and cross-sectional analytic approaches. For these analyses, only the 27 states that participated all three years in NAEP are used. This sample consists of 14 SSI states and 13 non-SSI states.

New methods are required for the use of State NAEP data for the longitudinal and cross-sectional analyses because of the unique nature of the State NAEP: sparse data at the state level (around 40 states); states' *voluntary* participation in tests, violating the assumption of randomness; and, the heterogeneous variance structure of each state, instead of the homogeneity required of the hierarchal linear model (Raudenbush, 2000; Raudenbush, Fotiu, & Cheong, 1999).

To address these characteristics of the State NAEP data, we developed an analytic method that will enable us to study the change in achievement scores at each grade and in cohort growth across states with the advantage of "comparable metrics" of the State NAEP scores across data collection years and across grades (Webb et al., 2001). Our longitudinal analyses employed two methods: empirical Bayes and fully Bayesian methods (Raudenbush, 2000; Raudenbush et al., 1999). Empirical Bayes analyses using the HLM program via restricted maximum likelihood (Raudenbush, Bryk, Cheong, & Congdon, 2000) and fully Bayesian analyses using the WinBUGS program via Gibbs sampling (Spiegelhalter, Thomas, & Best, 2000) are similar to a meta-analysis described in Bryk and Raudenbush (1992, Chapter 7). These methods incorporate the estimated standard errors with estimated state means for grades 4 and 8 in each test year. Thus, we employed empirical Bayes and Bayesian analyses to synthesize each state mean to estimate an overall mean across SSI states and non-SSI states. In a longitudinal perspective, these methods estimated the average state mean in 1992 and the average state growth rate per year.

The basic, or unconditional Bayesian model we used, follows:

$$\mathbf{b}_{ts}|\beta_{ts} \sim N(\beta_{ts}, V_{ts})$$
$$\beta_{ts}|\gamma \sim N(\gamma_0 + \gamma_1(\text{Time}), T)$$

where $\mathbf{b}_{ts}$ is a vector of estimated means from each state s at time t, β_{ts} is the corresponding vector of parameters, and $\mathbf{V}_{ts}$is the known variance-covariance matrix of the estimates $\mathbf{b}_{ts}$. Time is an annual growth variable coded as 0 in 1992, 4 in 1996, and 8 in 2000 and T is the between-state residual covariance matrix. We assume a multivariate normal prior distribution for γ and an inverse Wishart prior distribution for T (for details, see Browne & Draper, 2000a, 2000b; Raudenbush et al., 1999; Spiegelhalter et al., 2000).

The input data of state means and jackknife standard errors are listed in Tables 8A.1 and 8A.2 in the Appendix. These estimates are based on the results after taking into account the NAEP sampling design (Allen et al., 1997).

This chapter begins with longitudinal analyses of grade 8 data, grade 4 data, and two cohort data in order to compare the overall trends of SSI states and non-SSI states over the assessment years. The next section discusses the results of cross-state analyses using the data of grades 4 and 8 in 1992, 1996, and 2000. The latter analyses allow us to detect the differences between the two groups, the SSI and non-SSI states, in each assessment year. In each section, the statistical models used in the analyses are discussed.

Longitudinal Analysis: Empirical Bayes and Fully Bayesian Methods

The tables in this section display the parameter estimates from a longitudinal analysis of grade 8 students in mathematics scale scores from 1992 to 2000. In Tables 8.1 to 8.4, column one describes two different growth models (e.g., unconditional and conditional models) used in our analyses. Columns two and three summarize means, standard deviation, and credibility intervals obtained using the empirical Bayes method and the fully Bayesian method. The upper part of the tables presents the fixed effects or coefficients, and the lower part lists the random effects or variance components. The fully Bayesian method included both fixed and random effects, while the empirical Bayesian method had fixed effects only.

Overall, the estimates from the two methods appear to be fairly similar. As noted by many researchers, however, the fully Bayesian method has many properties that for this type of analysis are superior to the empirical Bayes method. In particular, the fully Bayesian method takes into account uncertainty regarding the parameters of interest. Thus, even though the results of both methods for the comparison are presented in the next section, we will focus mainly on the results from the fully Bayesian estimates of each parameter.

For all data sets of grade 8, grade 4, and the cohorts, the unconditional fully Bayesian model is based on samples of 20,000 iterations with 5,000 burn-in iterations. For the conditional models, the fully Bayesian results were run for 30,000 iterations after a burn-in of 10,000 to approximate the marginal posteriors of the parameters. For each parameter, the posterior distribution indicates the relative likelihood of potential values for that parameter and is affected both by the observed data and prior beliefs about that parameter as specified in the selected distribution.

Linear Growth Models of Grade 8

Unconditional model. Table 8.1 displays the results from linear growth models of grade 8 data from 1992, 1996, and 2000 using empirical Bayes and fully Bayesian methods. First, we begin with an unconditional model to estimate the average state mean in 1992 and average state growth rate across 27 states. The fully Bayesian estimate of average state mean in 1992 is 265.70. Average state growth rate per year from 1992 to 2000 is 0.81. This means that grade 8 students across the states are gaining an average 0.81 points per year. Both posterior estimates are statistically significant.

Regarding the variance components, the fully Bayesian posterior means for the variance of average state mean in 1992 and growth rate are 98.54 and 0.18 (Table 8.1). The 95% credibility intervals for these parameters range from 56.52 to 169.50 for average state mean in 1992 and from 0.08 to 0.34 for average growth rate. Since Figure 8.1 displays the posterior distribution of the variances of each estimate, there is evidence of between-state heterogeneity in 1992 in both average state mean and growth rate.

Conditional model. The next conditional model presents the differences in state mean in 1992 and the growth rate between SSI states and non-SSI states (Table 8.1). On average, SSI states started behind non-SSI states by 1.24 points in 1992 (average SSI differential effect in state mean in 1992 is –1.24). But, the growth rate of grade 8 students in SSI states is 0.10 points per year faster than that of their counterparts in non-SSI states (average SSI differential effect in state growth rate is 0.10). As a result, the learning gap between SSI states and non-SSI states was modestly reduced in 2000. However, the SSI differential effects are not statistically significant.

Table 8.1 and Figure 8.2 show the posterior distribution of the variance of all estimates in state mean in 1992 and growth rate. All four variance estimates for SSI states and non-SSI states are positive. The results indicated a substantial variability in 1992 in state mean and growth rate across both SSI states and non-SSI states.

Figures 8.3 and 8.4 show the Bayesian posterior estimates with the 95% credibility intervals for each state. Considering the state mean in 1992, each state does vary considerably in its performance (Figure 8.3). Half of each of non-SSI states and SSI states score below or above the average state mean in 1992. The low-performing states were Mississippi, Louisiana, and Alabama, and the high-performing states were Maine, Minnesota, and North Dakota. Two of three states that scored highest are non-SSI states. More interestingly, both the highest performing and lowest performing states are non-SSI states.

However, when we look at the posterior distribution of annual growth rates of each state, the pattern of state mean in 1992 is reversed (Figure 8.4). This also confirms a relative advantage for the SSI states over non-SSI states in growth rate. Despite the low scores in 1992, grade 8 students in SSI states were more likely to show a gain than their counterparts in non-SSI states. Indiana, West Virginia, Maryland, Michigan, Massachusetts, and Texas were fast-gaining states and half of these states are SSI states. But, on average, the group differences between SSI states and non-SSI states are not statistically different from zero, even though the Bayesian posterior

variance estimates provide clear evidence of heterogeneity in growth rate both among SSI states and non-SSI states.

Table 8.1
Longitudinal Analysis of Grade 8 Data over 1992, 1996, and 2000: Empirical Bayes and Fully Bayesian Estimates After Considering Jackknife Standard Errors

Model	**Empirical Bayes**		**Fully Bayesian**			
Fixed Effect	*Coefficient*	*SD*	*Coefficient*	*SD*	*Credibility Interval*	
					2.5%	*97.5%*
Linear Growth Model –Time						
Average state mean in 1992	265.799***	1.659	265.700	1.928	261.800	269.500
Average state growth rate (per year)	0.806**	0.322	0.811	0.092	0.629	0.994
Linear Growth Model –Time, SSI, and Time x SSI						
Non-SSI State						
Average Non-SSI state mean in 1992	266.303**	2.417	266.200	2.867	260.000	271.500
Average Non-SSI state growth rate (per year)	0.758~	0.469	0.757	0.165	0.436	1.084
SSI State						
Average SSI state mean in 1992	265.328		264.960			
Average SSI state growth rate (per year)	0.851		0.858			
SSI Effect						
Average SSI differential effect in state mean	-0.975	3.360	-1.240	3.701	-7.452	7.328
Average SSI differential effect in state growth rate	0.093	0.651	0.101	0.238	-0.371	0.569
Random Effect						
Linear Growth Model –Time						
Variance (Mean)			98.540	29.220	56.520	169.500
Variance (Time)			0.175	0.066	0.081	0.336
Linear Growth Model –Time, SSI, and Time x SSI						
Variance (Mean)			113.800	52.770	52.670	252.600
Variance (Time)			0.314	0.149	0.135	0.687
Variance (SSI)			11.180	18.220	0.221	64.280
Variance (Time x SSI)			0.405	0.253	0.119	1.071

~ $p \leq .1$, * $p \leq .05$, ** $p \leq .01$, *** $p \leq .001$

Figure 8.1. Posterior distribution of the variance: Unconditional model.

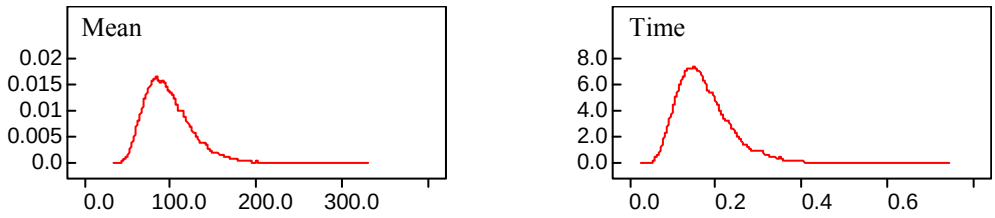


Figure 8.2. Posterior distribution of the variance: Conditional model.

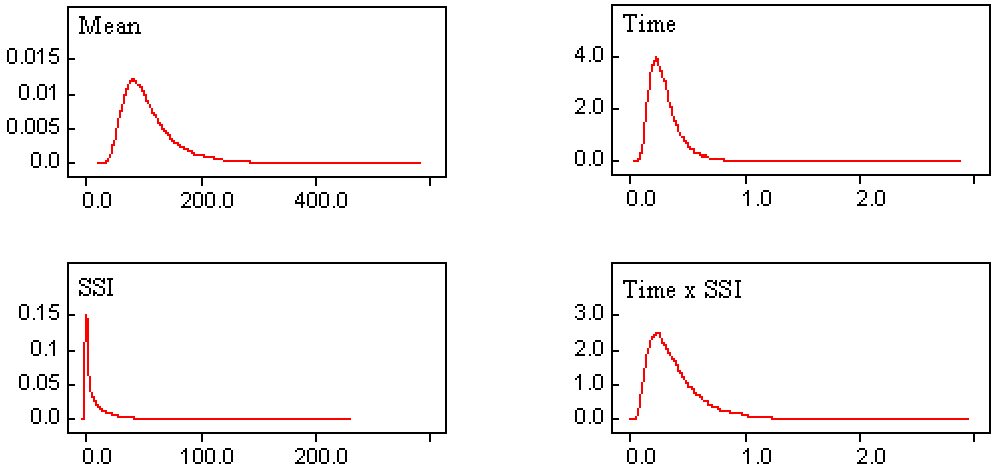


Figure 8.3. Posterior distributions of average scale scores of grade 8 in 1992.

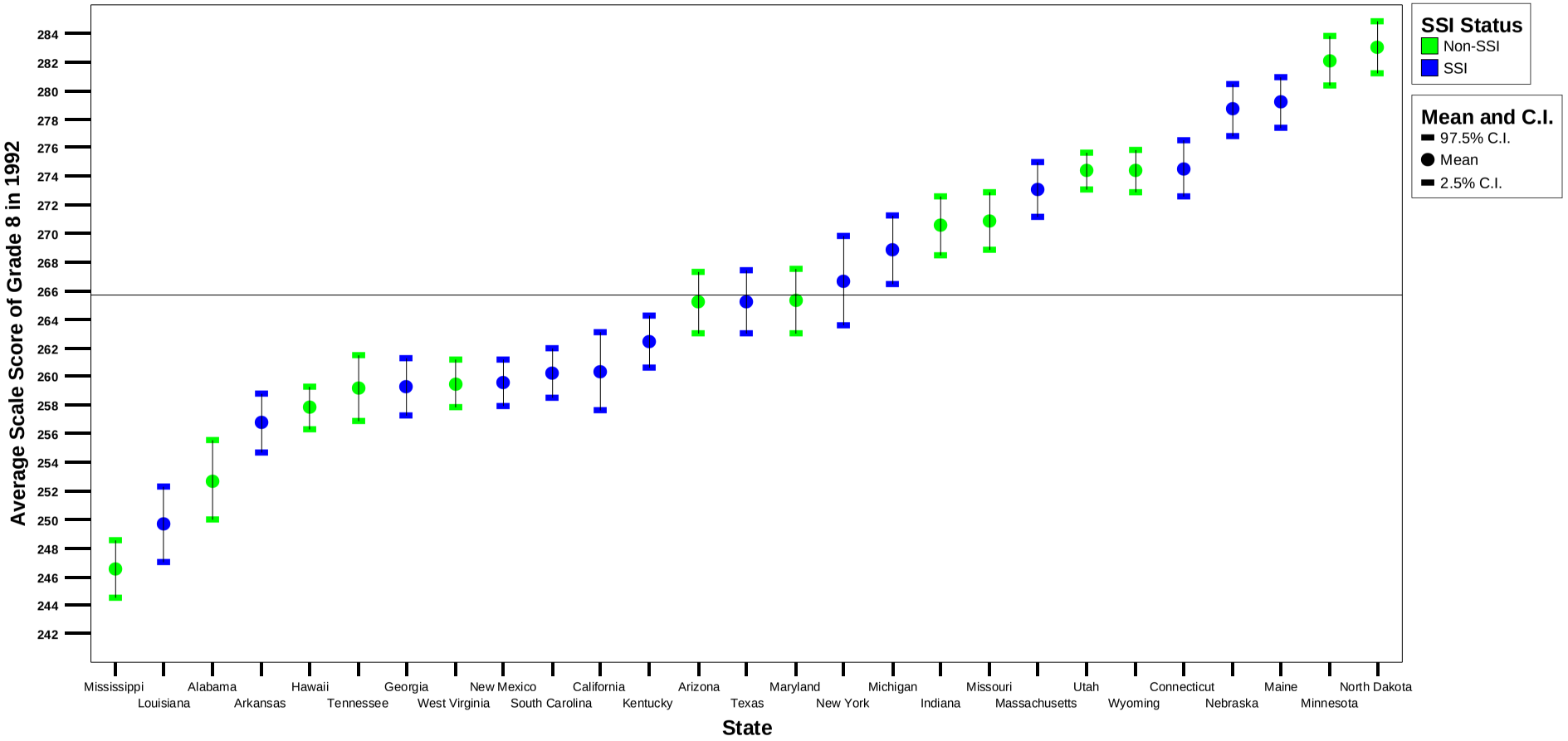
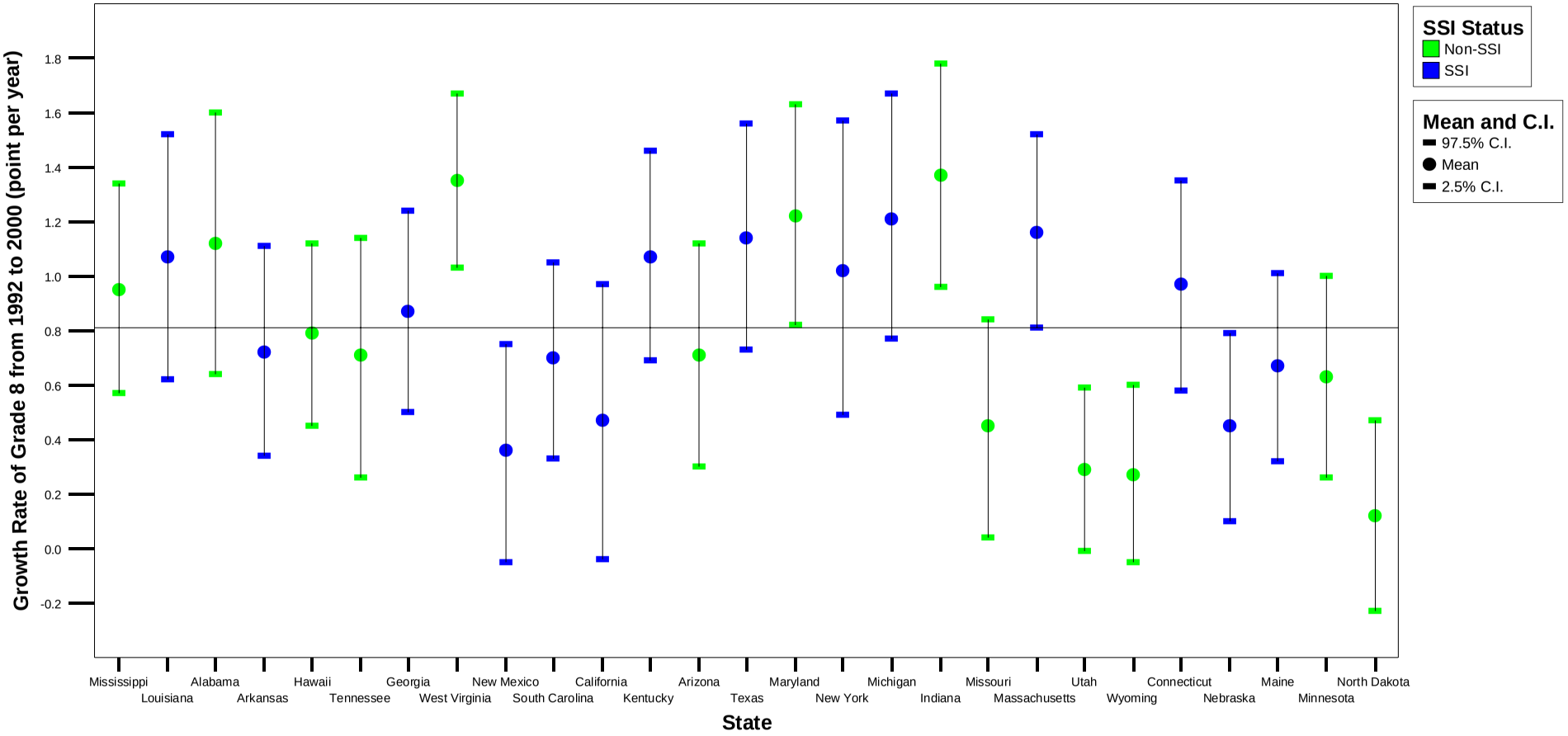


Figure 8.4. Posterior distributions of growth rates of grade 8 from 1992 to 2000.



Linear Growth Models of Grade 4

Unconditional model. The results from the unconditional linear growth model of grade 4 data are presented in Table 8.2. Both empirical Bayes and fully Bayesian estimates of average state mean in 1992 and average growth rate are identical. The fully Bayesian estimated average state mean in 1992 and average state growth rate for grade 4 mathematics scale scores were 217.80 and 0.82, respectively. This indicates that the average grade 4 mathematics scores in 1992 across SSI and non-SSI states was 217.80 points and grade 4 students were increasing at the rate of 0.82 points per year from 1992 to 2000.

Table 8.2 also shows that there was a true variation in average state mean in 1992 and average growth rate among states. The posterior means of the variance are 63.840 for average state mean in 1992 and 0.22 for average state growth rate. However, the values of the variance of average state mean in 1992 can be as small as 36.63 and as large as 110.40. The posterior distribution of the variance of average growth rate ranges from 0.11 to 0.41. The results indicate that a quite substantial variability in these two estimates exists between states.

Conditional model. Table 8.2 shows the results of a conditional model after adjusting for SSI status. For non-SSI states, the average mathematics score mean in 1992 was 218.10 and average growth rate per year was 0.79. For SSI states, grade 4 students began with 217.63 points (218.10 – 0.47), but they gained more, at 0.84 points (0.79 + 0.04), per year. Thus, students in SSI states were more likely to score lower initially but tended to learn faster than their counterparts in non-SSI states. However, the two SSI differential effects are not statistically significant.

Figure 8.6 displays the posterior distributions of the variance of four estimates in average state mean in 1992 and average growth rate. Each of the groups of SSI and non-SSI states had a considerable heterogeneity in average state mean in 1992 and average growth rate.

This is also confirmed by the line charts in Figures 8.7 and 8.8. The average state mean in 1992 differed significantly from state to state (Figure 8.7). Each state mean can be as low as 203 points and as high as 231 points. The top three states were Minnesota, North Dakota, and Maine, and the lowest three states were Mississippi, Louisiana, and California. Two of the three higher-performing states were non-SSI states. Figure 8.8 displays an interesting picture regarding state growth rate. While there is some between-state variation in growth rate, both the greatest gaining and least gaining states are SSI states. For example, Texas is a high-gaining state and Maine a low-gaining state. As indicated in Table 8.2, there were no overall differences between SSI and non-SSI states in state mean in 1992 and state gain rate. But, the patterns of Figures 8.7 and 8.8 are consistent with the results of Table 8.2, which show considerable variance across both SSI states and non-SSI states in the two estimates.

Table 8.2
Longitudinal Analysis of Grade 4 Data in 1992, 1996, and 2000: Empirical Bayes and Fully Bayesian Estimates After Considering Jackknife Standard Errors

Model	**Empirical Bayes**		**Fully Bayesian**			
Fixed Effect	*Coefficient*	*SD*	*Coefficient*	*SD*	*Credibility Interval*	
					2.5%	*97.5%*
Linear Growth Model –Time						
Average state mean in 1992	217.894***	1.329	217.800	1.558	214.700	220.900
Average state growth rate (per year)	0.818**	0.258	0.820	0.099	0.623	1.016
Linear Growth Model –Time, SSI, and Time x SSI						
Non-SSI State						
Average Non-SSI state mean in 1992	218.164***	1.937	218.100	1.719	214.500	221.500
Average Non-SSI state growth rate (per year)	0.798*	0.375	0.793	0.148	0.498	1.088
SSI State						
Average SSI state mean in 1992	217.642		217.628			
Average SSI state growth rate (per year)	0.836		0.837			
SSI Effect						
Average SSI differential effect in state mean	-0.522	2.694	-0.472	2.340	-4.912	4.142
Average SSI differential effect in state growth rate	0.038	0.522	0.044	0.240	-0.419	0.511
Random Effect						
Linear Growth Model –Time						
Variance (Mean)			63.840	19.020	36.630	110.400
Variance (Time)			0.220	0.078	0.109	0.410
Linear Growth Model –Time, SSI, and Time x SSI						
Variance (Mean)			50.220	21.980	22.680	104.700
Variance (Time)			0.253	0.116	0.110	0.540
Variance (SSI)			28.650	43.570	0.265	153.900
Variance (Time x SSI)			0.376	0.250	0.107	1.023

~ $p \leq .1$, * $p \leq .05$, ** $p \leq .01$, *** $p \leq .001$

Figure 8.5. Posterior distribution of the variance: Unconditional model.

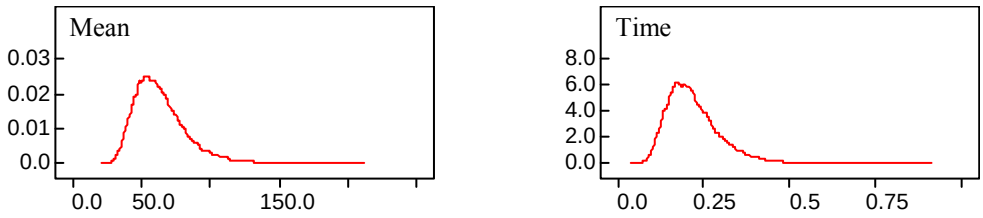
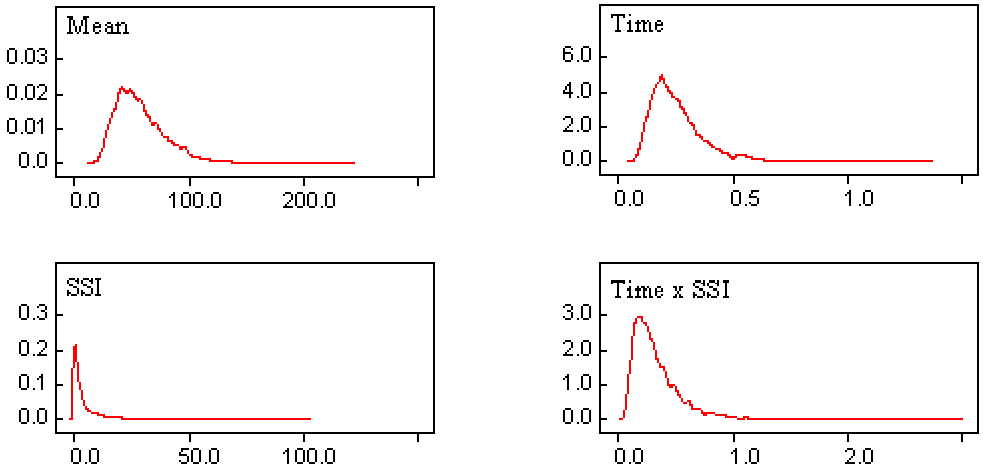


Figure 8.6. Posterior distribution of the variance: Conditional model.



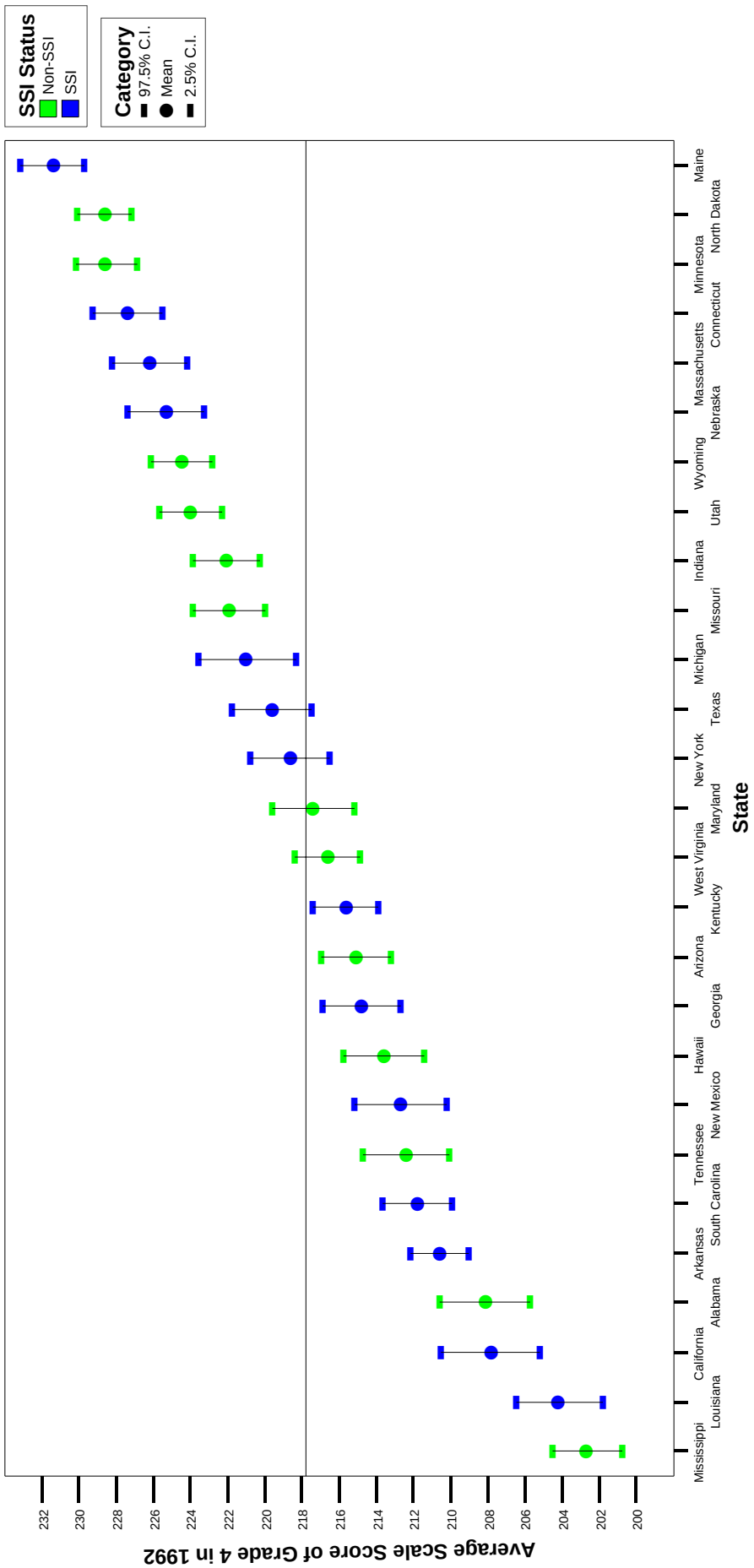


Figure 8.7. Posterior distributions of average scale scores of grade 4 in 1992.

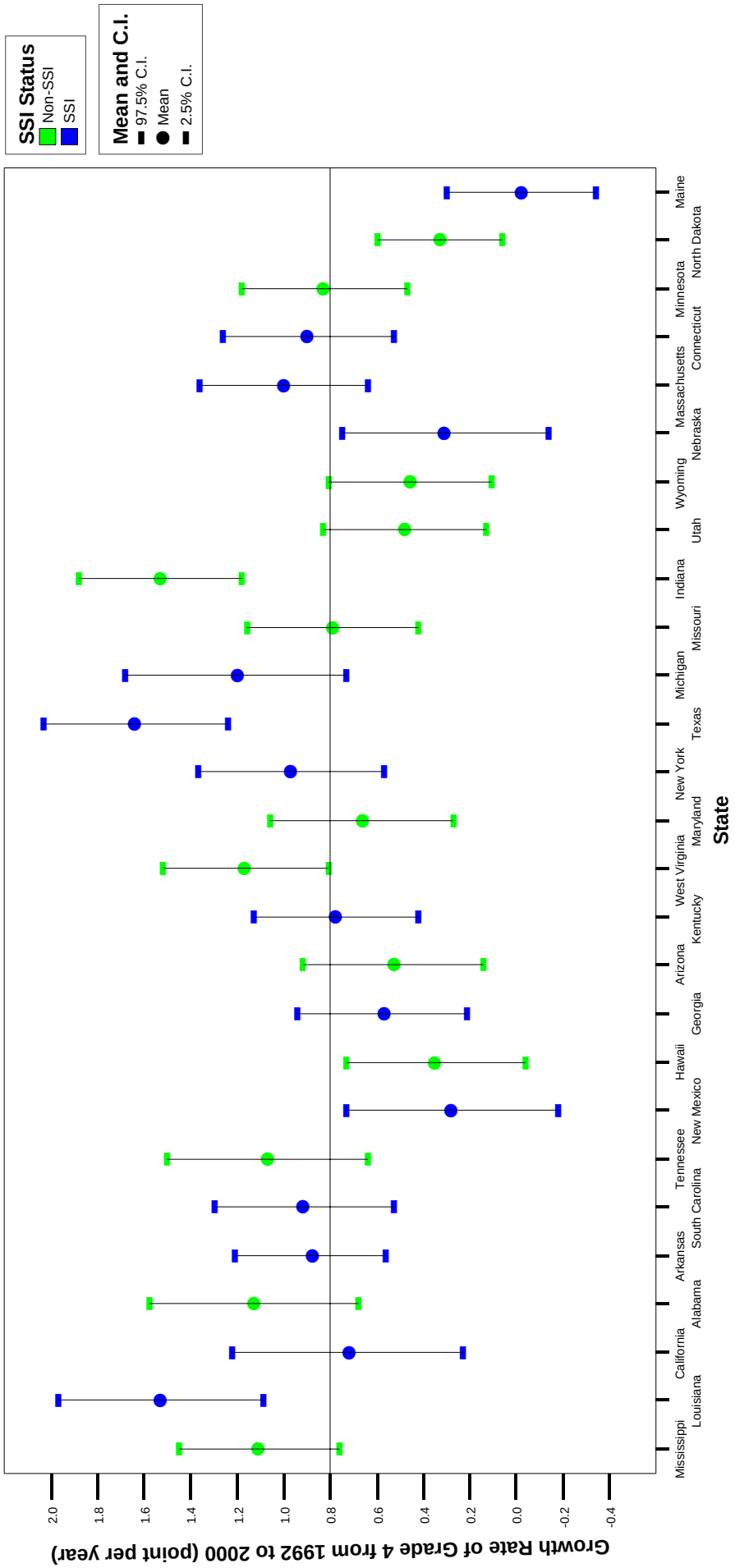


Figure 8.8. Posterior distributions of growth rates of grade 4 from 1992 to 2000.

Linear Growth Models of Cohort 1—Grade 4 in 1992 and Grade 8 in 1996

Unconditional model. Table 8.3 displays the estimates of empirical Bayes and Bayesian methods of cohort data analysis for grade 4 in 1992 and grade 8 in 1996 in the unconditional linear growth model. The estimated posterior mean of average state mean in 1992 is 217.70 points. This is nearly identical to the estimate in the previous result of grade 4 data in Table 8.2 because the base data point for Cohort 1 is grade 4 in 1992. The overall state growth rate of the cohort was 12.88 points per year. That is, the grade 4 Cohort 1 among all states was likely to gain 51.52 (12.88×4) points in mathematics scale scores from 1992 to 1996. Each estimate of the cohort growth model is statistically significant.

Posterior variations in average state mean in 1992 and average state growth rate are also listed in Table 8.3. The estimated posterior variance mean of average state mean in 1992 is 61.94 and that of average state growth rate is 0.30. As Figure 8.9 shows, the posterior distributions of these estimates indicate that both estimates vary significantly between states. The 95% credibility interval for average state mean in 1992 ranges from 35.56 to 107.40. Average state growth rate in 1992 ranges from 0.12 to 0.63.

Conditional model. Table 8.3 displays the results from the conditional models. In general, students in non-SSI states scored higher in 1992, but gained less than their counterparts in SSI states. On average, the non-SSI state mean in 1992 was 217.80 points and the growth rate 12.85 points per year. But, the state mean in 1992 and growth rate for SSI states were 217.52 (217.80 - 0.28) points and 12.92 (12.85 + 0.07) points, respectively. The posterior estimates of SSI differential effects are not statistically significant.

Next, the posterior estimates of the variance of average state mean in 1992 and average state growth rate for each SSI state and non-SSI state are presented in Table 8.3. All four posterior means of the variance are positive and significant, indicating that each group, SSI states and non-SSI states, is heterogeneous in state mean in 1992 and state growth rate. The posterior distributions of these four fully Bayesian estimates in Figure 8.10 clearly indicate the variability of these estimates.

Figures 8.11 and 8.12 display the line charts describing the posterior distribution of average state mean in 1992 and average state cohort growth rate for each state. As previously noted, the chart of average state mean in 1992 is nearly identical to those shown in grade 4 data. Thus, only Figure 8.12 will be discussed. Each state varies in its posterior mean of state cohort growth rate. Especially, the low-gaining states (e.g., Mississippi, Louisiana, and Alabama) appear different from the high-gaining states (Nebraska, North Dakota, and Minnesota) in the estimated state growth rate. The distributions of SSI states and non-SSI states display a similar pattern. The estimated means among SSI states and among non-SSI states were both quite variable, as indicated by the 95% credibility intervals of state growth rate. The distribution of the estimated posterior means illustrates the results presented in Table 8.3, indicating no significant differences between SSI and non-SSI states in average state cohort growth rate.

Table 8.3
Longitudinal Analysis of Cohort 1 Data for Grade 4 in 1992 and Grade 8 in 1996: Empirical Bayes and Fully Bayesian Estimates After Considering Jackknife Standard Errors

Model	**Empirical Bayes**		**Fully Bayesian**			
Fixed Effect	*Coefficient*	*SD*	*Coefficient*	*SD*	*Credibility Interval*	
					2.5%	*97.5%*
Linear Growth Model –Time						
Average state mean in 1992	217.752***	1.686	217.700	1.539	214.700	220.800
Average state growth rate (per year)	12.913***	0.597	12.880	0.135	12.610	13.140
Linear Growth Model –Time, SSI, and Time x SSI						
Non-SSI State						
Average Non-SSI state mean in 1992	217.941***	2.475	217.800	1.799	214.400	221.500
Average Non-SSI state growth rate (per year)	12.892***	0.877	12.850	0.214	12.430	13.260
SSI State						
Average SSI state mean in 1992	217.574		217.518			
Average SSI state growth rate (per year)	12.932		12.917			
SSI Effect						
Average SSI differential effect in state mean	-0.367	3.440	-0.282	2.238	-4.471	4.005
Average SSI differential effect in state growth rate	0.040	1.219	0.067	0.320	-0.563	0.700
Random Effect						
Linear Growth Model –Time						
Variance (Mean)			61.940	18.600	35.560	107.400
Variance (Time)			0.300	0.133	0.117	0.631
Linear Growth Model –Time, SSI, and Time x SSI						
Variance (Mean)			57.180	25.250	26.290	120.400
Variance (Time)			0.485	0.248	0.184	1.121
Variance (SSI)			9.982	12.150	0.250	41.920
Variance (Time x SSI)			0.441	0.312	0.116	1.234

~ $p \leq .1$, * $p \leq .05$, ** $p \leq .01$, *** $p \leq .001$

Figure 8.9. Posterior distribution of the variance: Unconditional model.

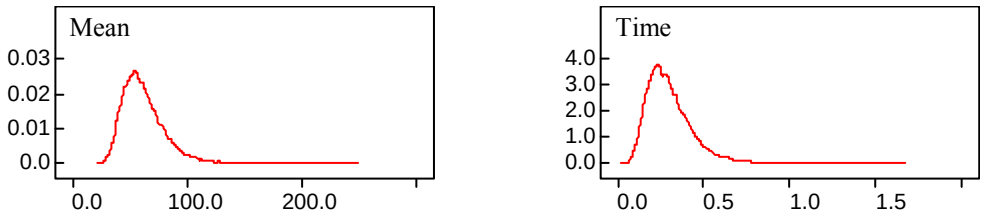


Figure 8.10. Posterior distribution of the variance: Conditional model.

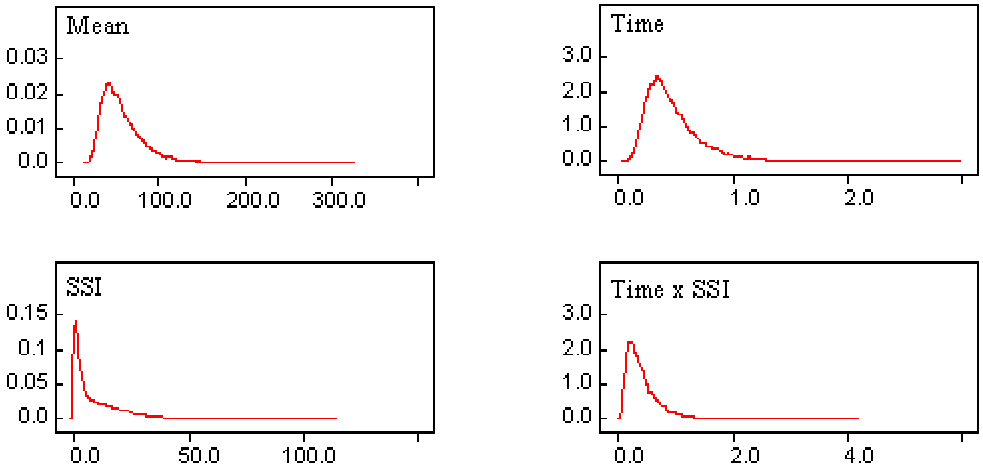


Figure 8.11. Posterior distributions of average scale scores of Cohort 1—grade 4 in 1992 and grade 8 in 1996.

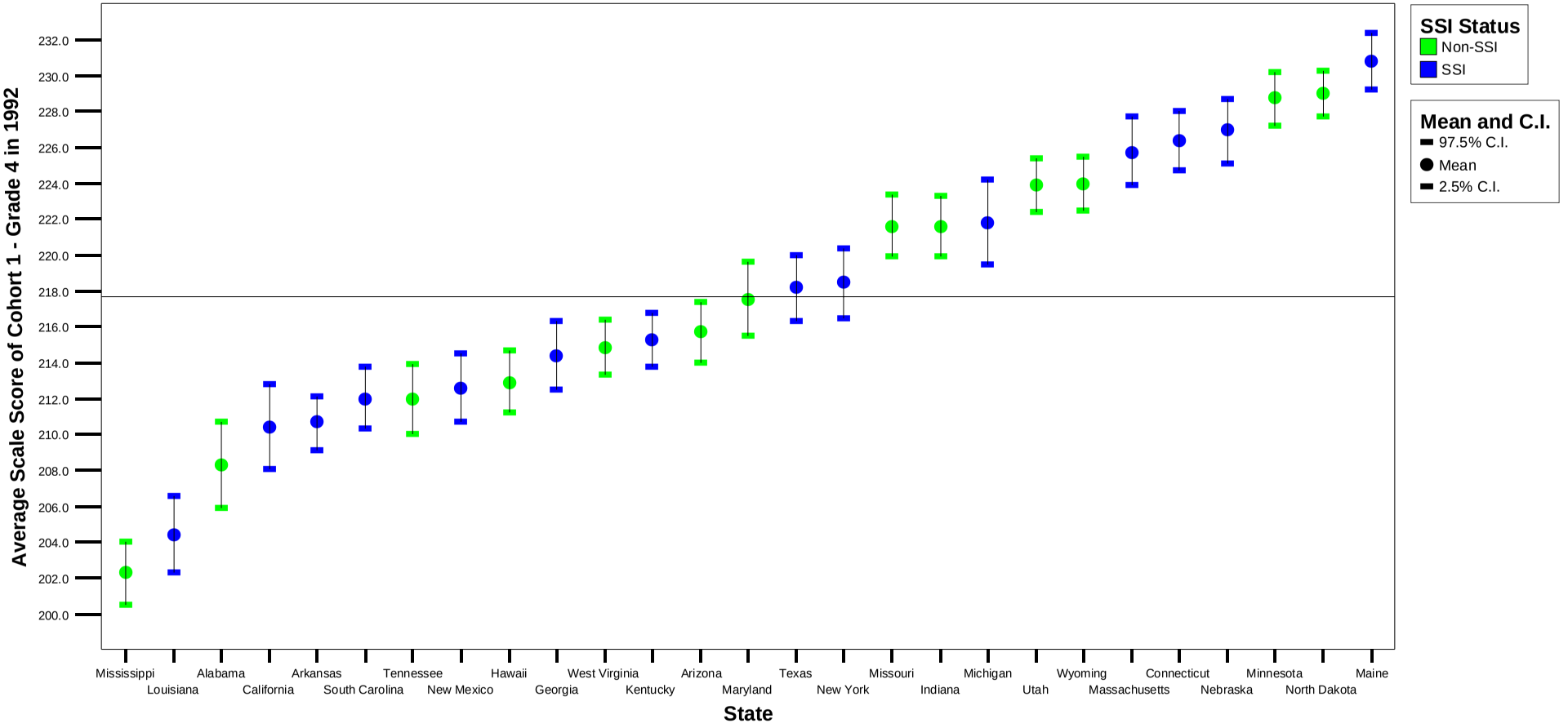
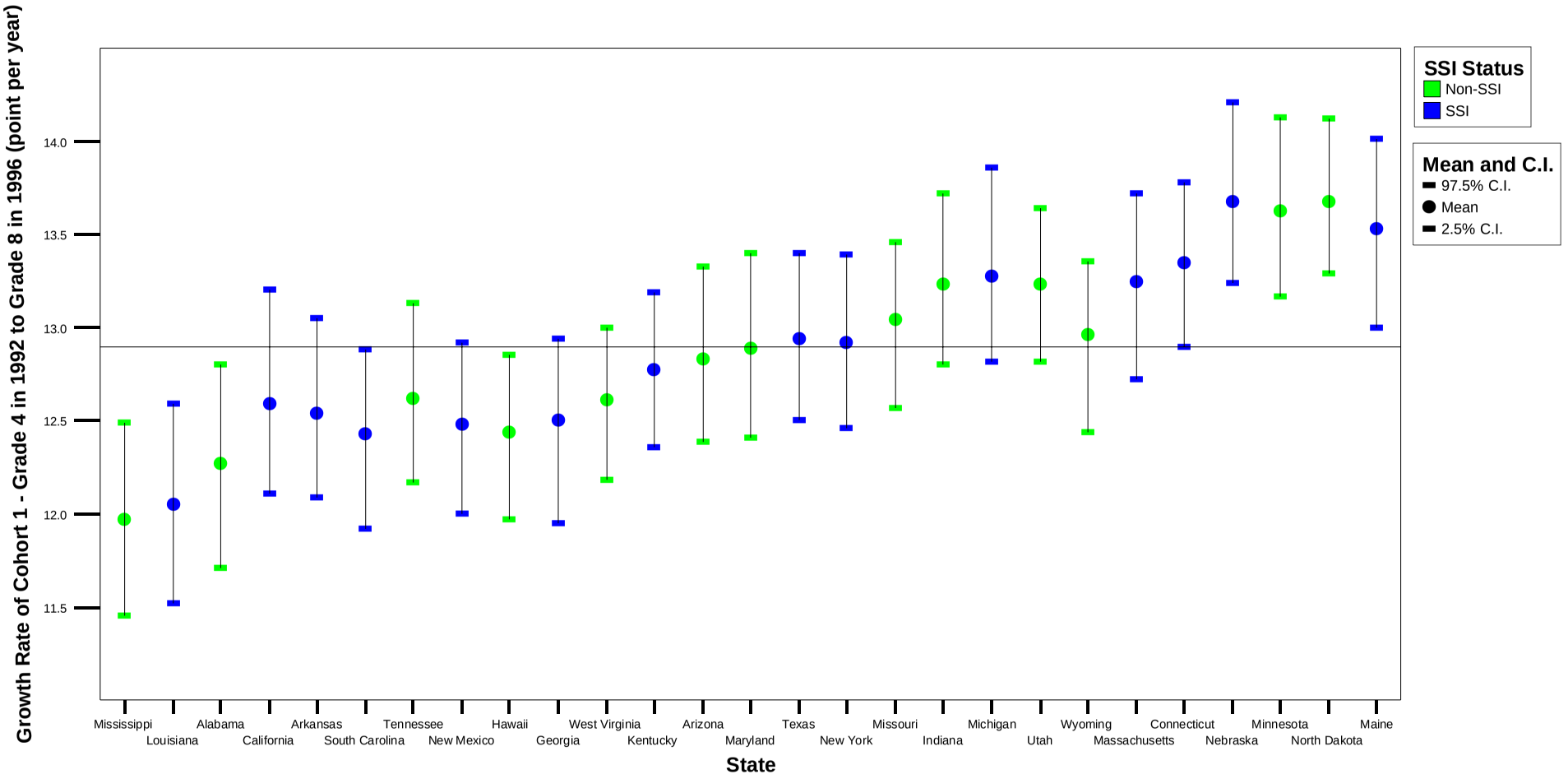


Figure 8.12. Posterior distributions of growth rates of Cohort 1—grade 4 in 1992 and grade 8 in 1996.



Linear Growth Models of Cohort 2—Grade 4 in 1996 and Grade 8 in 2000

Unconditional model. The results of empirical Bayes and Bayesian methods of Cohort 2 data analysis for grade 4 in 1996 and grade 8 in 2000 are shown in Table 8.4. In this unconditional model, grade 4 data in 1996 is the baseline for Cohort 2. The estimated average state mean in 1996 and average state cohort growth rate per year from 1996 to 2000 are 221.40 and 12.62, respectively. Thus, the grade 4 students in 1996 were gaining 12.62 points per year in average mathematics scale scores. Both posterior means of the Cohort 2 growth model are statistically significant.

Table 8.4 and Figure 8.13 show that each state varied significantly in both posterior variations in average state mean in 1996 and average state growth rate. The posterior means of the variance are 59.25 for average state mean in 1996 and 0.32 for average growth rate. The posterior distributions for these parameters range from 33.87 to 103.30 for average state mean in 1996 and from 0.10 to 0.73 for average state growth rate of Cohort 2.

Conditional model. Table 8.4 shows the results of the conditional models using SSI status as a factor. For non-SSI states, the average state mean in 1996 was 221.80 points and average growth rate per year is 12.57 points. For SSI states, grade 4 students in 1996 started with 221.17 (221.80 -0.63) points, but the growth rate of these cohort students is 12.69 (12.57 + 0.12) points per year, faster than that of their counterparts in non-SSI states. Neither of the estimated SSI differential effects is statistically significant.

Table 8.4 and Figure 8.14 also display the posterior distribution of the variance of all estimates in average state mean in 1996 and average state growth rate. The results indicate there is a substantial variability in state mean in 1996 and state cohort growth rate across SSI states and non-SSI states.

Figures 8.15 and 8.16 display the Bayesian posterior estimates for each state as the line charts of average state mean in 1996 and average state growth rate. In general, compared to SSI states, most non-SSI states scored above average state mean in 1996, while there were no overall differences in distribution of SSI states and non-SSI states in average growth rate. But these two figures show an evident pattern for non-SSI states. That is, the lowest scored Mississippi and the highest scored Minnesota were also the least gaining and highest gaining states. Moreover, both are non-SSI states.

Table 8.4

Longitudinal Analysis of Cohort 2 Data for Grade 4 in 1996 and Grade 8 in 2000: Empirical Bayes and Fully Bayesian Estimates After Considering Jackknife Standard Errors

Model	**Empirical Bayes**		**Fully Bayesian**			
Fixed Effect	*Coefficient*	*SD*	*Coefficient*	*SD*	*Credibility Interval*	
					2.5%	*97.5%*
Linear Growth Model –Time						
Average state mean in 2000	221.426***	1.633	221.400	1.508	218.500	224.400
Average state growth rate (per year)	12.659***	0.578	12.620	0.142	12.340	12.900
Linear Growth Model –Time, SSI, and Time x SSI						
Non-SSI State						
Average Non-SSI state mean in 2000	221.774***	2.397	221.800	1.627	218.600	225.200
Average Non-SSI state growth rate (per year)	12.627***	0.848	12.570	0.252	12.060	13.070
SSI State						
Average SSI state mean in 2000	221.099		221.167			
Average SSI state growth rate (per year)	12.688		12.690			
SSI Effect						
Average SSI differential effect in state mean	-0.675	3.331	-0.633	2.340	-5.420	3.652
Average SSI differential effect in state growth rate	0.061	1.179	0.120	0.351	-0.564	0.816
Random Effect						
Linear Growth Model –Time						
Variance (Mean)			59.250	17.910	33.870	103.300
Variance (Time)			0.322	0.165	0.100	0.727
Linear Growth Model –Time, SSI, and Time x SSI						
Variance (Mean)			43.670	18.100	21.000	89.550
Variance (Time)			0.656	0.327	0.249	1.485
Variance (SSI)			15.710	15.510	0.365	53.640
Variance (Time x SSI)			0.616	0.454	0.136	1.794

~ $p \leq .1$, * $p \leq .05$, ** $p \leq .01$, *** $p \leq .001$

Figure 8.13. Posterior distribution of the variance: Unconditional model.

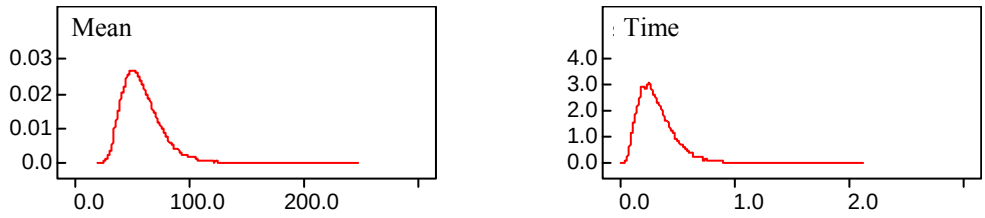


Figure 8.14. Posterior distribution of the variance: Conditional model.

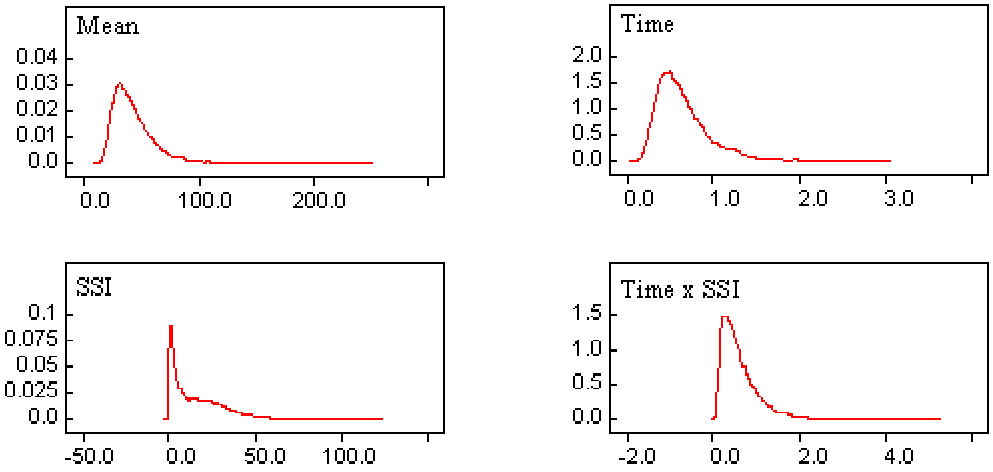


Figure 8.15. Posterior distributions of average scale scores of Cohort 2—grade 4 in 1996 and grade 8 in 2000.

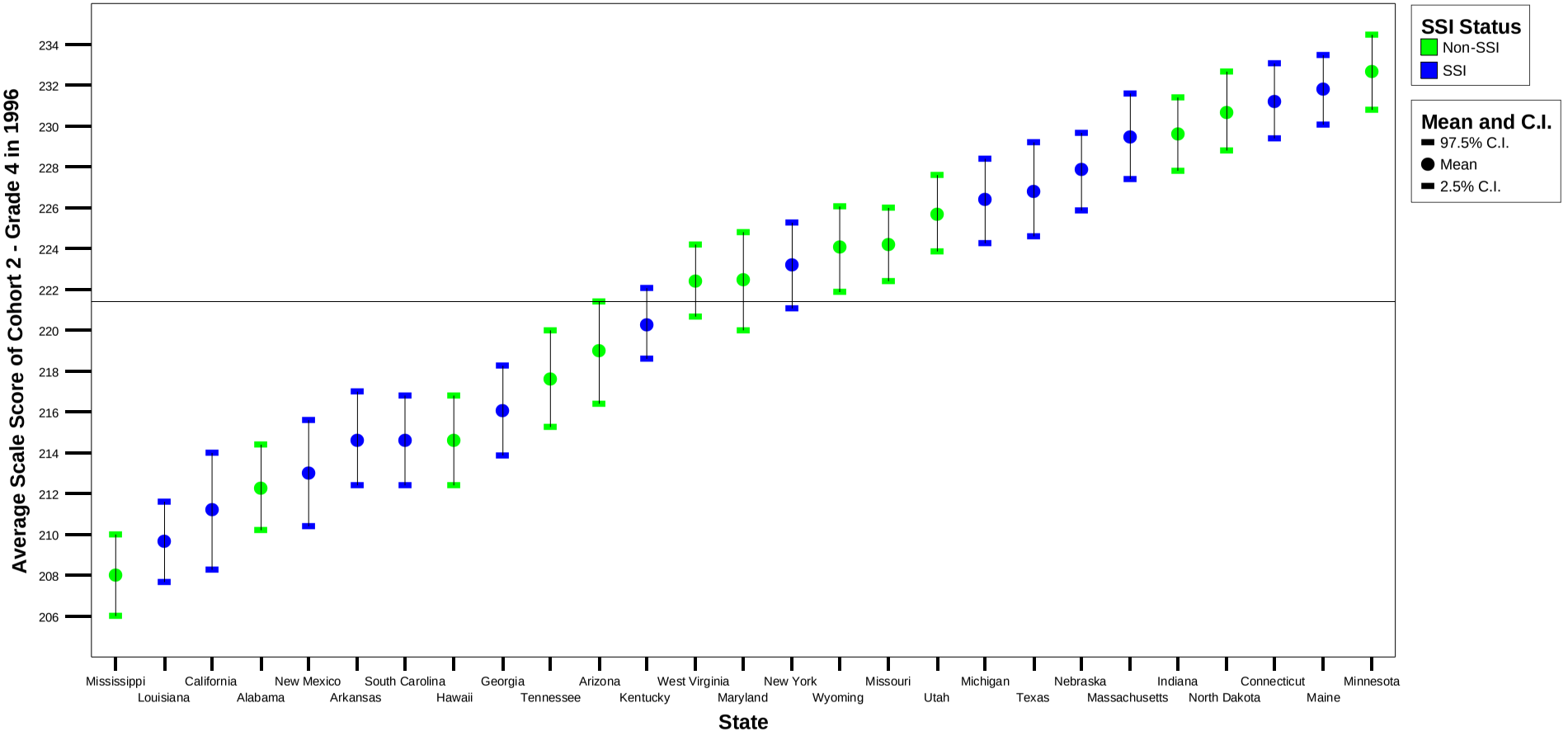
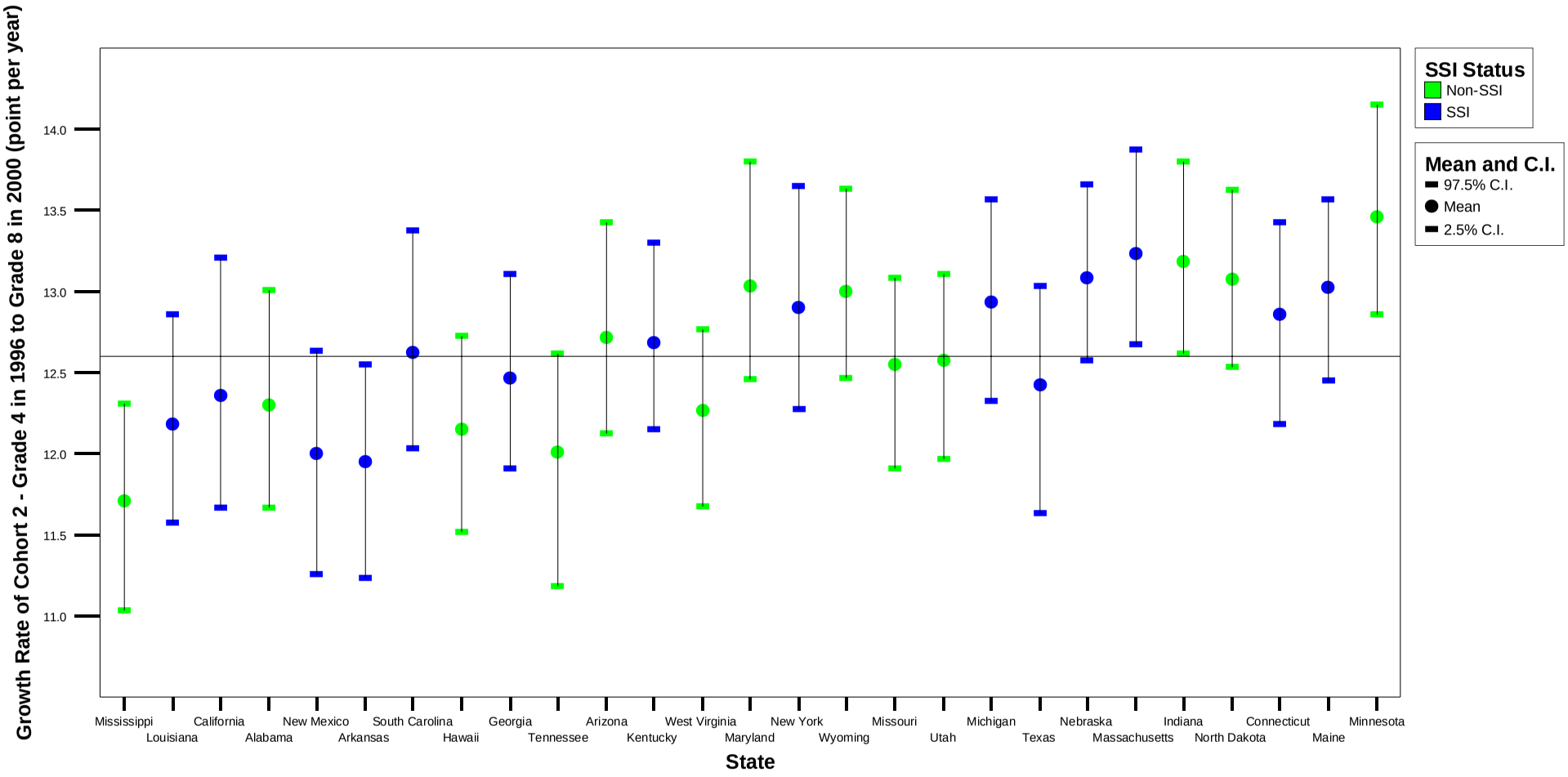


Figure 8.16. Posterior distributions of growth rates of Cohort 2—Grade 4 in 1996 and grade 8 in 2000.



Cross-Sectional Analysis: Empirical Bayes Method

Grade 8

Unconditional model. Table 8.5 presents the results of the unconditional models, using the empirical Bayesian method. Over three assessment years in 1992, 1996, and 2000, average state means increased from 265.61 points in 1992 to 269.40 points in 1996 to 272.06 points in 2000. This was a gain of close to 4 points from 1992 to 1996 and almost 3 points from 1996 to 2000. Thus, there was evidence that grade 8 students across the states were likely to gain in mathematics scale scores from 1992 to 2000. The estimated variance of average state mean is also displayed in Table 8.5. Overall, all three estimates of the variance indicate that significant variability between states existed in each of the average state means. The estimated variance ranges from 83.09 to 91.36, which is substantial.

Conditional model. In the conditional models, we tried to detect any differences between SSI and non-SSI states in average state mean over the period of 1992-1996-2000. The results show that differential effects of SSI states are not significant and negative, as large as -1.17 points and as small as -0.20 points. Overall, grade 8 students in non-SSI states outperformed counterparts in SSI states over all three assessment years. However, the gaps between SSI and non-SSI states narrowed by .97 points from 1992 to 1996 and widened by .23 points from 1996 to 2000. As the variance components in Table 8.5 indicate, SSI differential effects cannot account for between-state variance in average state means. Much of the variability between states still remains to be explained.

Table 8.5
Cross-Sectional Analysis of Grade 8 Data: Empirical Bayes Estimates After Considering Jackknife Standard Errors

Model	**1992**		**1996**		**2000**	
Fixed Effect	*Coefficient*	*SE*	*Coefficient*	*SE*	*Coefficient*	*SE*
Unconditional Model						
Average state mean	265.610***	1.855	269.399***	1.859	272.057***	1.776
Conditional Model – SSI						
Non-SSI State						
Average Non-SSI state mean	266.217***	2.718	269.504***	2.730	272.281***	2.607
SSI State						
Average SSI state mean	265.043		269.300		271.847	
SSI Effect						
Average SSI differential effect	-1.174	3.778	-0.204	3.794	-0.434	3.624
Random Effect	*Variance*	*SD*	*Variance*	*SD*	*Variance*	*SD*
Unconditional Model	91.356***	9.558	91.257***	9.553	83.093***	9.116
Conditional Model	94.688***	9.731	94.964***	9.7445	86.451***	9.298

~ $p \leq .1$, * $p \leq .05$, ** $p \leq .01$, *** $p \leq .001$

Grade 4

Unconditional model. The results of grade 4 average state means in 1992, 1996, and 2000 are displayed in Table 8.6. As presented earlier in grade 8 results, grade 4 students also show continued increases in mathematics scores from 1992 and 2000. The empirical Bayes estimates of average state means were 217.76 points across states in 1992 data, 221.43 points across states in 1996 data, and 224.30 points across states in 2000 data. The overall gain of grade 4 students was about 6.5 points over the two assessment periods. As the bottom part of Table 8.6 indicates, both estimated average state means varied significantly between states. The estimates for the variance of average state means in 1992, 1996, and 2000 were 59.56, 57.55, and 52.25, respectively. This suggests that state means in 1992 were more heterogeneous than those in 1996 and 2000.

Conditional model. Table 8.6 presents the empirical Bayes estimated results of the conditional models, including SSI status, as a covariate. In general, the mathematics scores for grade 4 students in both SSI and non-SSI states increased substantially over the three assessment years of 1992, 1996, and 2000. But, non-SSI states scored slightly higher than SSI states. In 1992, average state means were 217.98 points for non-SSI states and 217.58 (217.95 - 0.37) points for SSI states. In 1996, the average non-SSI state mean was 221.78 points and average SSI state mean was 221.11 (221.78 - 0.67) points. In 2000, the estimated average means were 224.33 and 224.27 (224.33 - 0.06) points for non-SSI states and SSI states. Thus, the mathematics score gap between SSI and non-SSI states was increased by 0.30 points from 1992 to 1996 and then decreased by 0.61 points from 1996 to 2000. Table 8.6 also shows how much variation the SSI status indicator explains in average state means. The SSI differential effects explained no between-state variance in 1992, 1996, and 2000.

Table 8.6
Cross-Sectional Analysis of Grade 4 Data: Empirical Bayes Estimates After Considering Jackknife Standard Errors

Model	**1992**		**1996**		**2000**	
Fixed Effect	*Coefficient*	*SE*	*Coefficient*	*SE*	*Coefficient*	*SE*
Unconditional Model						
Average state mean	217.758***	1.503	221.433***	1.482	224.303***	1.413
Conditional Model – SSI						
Non-SSI State						
Average Non-SSI state mean	217.949***	2.205	221.779***	2.174	224.334***	2.074
SSI State						
Average SSI state mean	217.578		221.108		224.271	
SSI Effect						
Average SSI differential effect	-0.371	3.066	-0.671	3.021	-0.063	2.884
Random Effect	*Variance*	*SD*	*Variance*	*SD*	*Variance*	*SD*
Unconditional Model	59.559***	7.717	57.550***	7.586	52.254***	7.229
Conditional Model	61.952***	7.871	59.804***	7.799	54.407***	7.376

~ $p \leq .1$, * $p \leq .05$, ** $p \leq .01$, *** $p \leq .001$

Summary and Conclusions

In this chapter, we report longitudinal and cross-sectional analyses of the State NAEP data conducted on mathematics achievement for grades 4 and 8 students in 1992, 1996, and 2000, and for two cohorts of students, grade 4 in 1992 and grade 8 in 1996, and grade 4 in 1996 and grade 8 in 2000. Using the empirical Bayes and fully Bayesian methods, we found that both SSI and non-SSI states showed an overall gain in mathematics scores across the assessment years. The results also revealed that a substantial variation in average state gain existed among states. Comparing the achievement growth between SSI and non-SSI states in average mathematics scale scores, the following summarizes the major findings in this chapter:

- In 1992, grade 8 students in SSI states scored 1.24 points lower than those in non-SSI states, but showed 0.10 points in faster annual growth from 1992 to 2000 than those in non-SSI states. Among the 27 states, half of the six highest gaining states—Michigan, Massachusetts, and Texas—were SSI states.
- Grade 4 students in SSI states started 0.47 points behind those in non-SSI states at points in 1992, but gained more, 0.04 points per year, than their counterparts in non-SSI states. Louisiana and Texas made the highest gains.
- For the cohort students in grade 4 in 1992 and grade 8 in 1996, SSI states scored lower than non-SSI states in 1992 by 0.28 points and gained more at 0.07 points per year. Nebraska, North Dakota, Minnesota, and Maine had the highest annual gains.
- Grade 4 Cohort 2 students in SSI states performed lower in 1996 by 0.63 points than those in non-SSI states, but learned faster by 0.12 points per year than their counterparts in non-SSI states. Minnesota and Massachusetts were the highest gaining states.
- There was clear evidence of the variance in average growth rate across SSI states and non-SSI states.
- In each year of grade 8 data in 1992, 1996, and 2000, the mathematics scores in SSI states were lower than those in non-SSI states. But, the gaps between SSI and non-SSI states were reduced from 1992 to 1996, and then widened slightly from 1996 to 2000. But for grade 4 data in 1992, 1996, and 2000, the pattern was reversed. Much of the between-state variance in average state means remains to be further explained.

Our findings regarding the effectiveness of SSI states in improving mathematics achievement over non-SSI states need to be interpreted with care in terms of data limitations. For longitudinal analyses of the two cohorts, our results were based on only two time points and therefore may not provide adequate data on the overall trends of growth for cohort students. State means used in this study as the input data for empirical Bayes and fully Bayesian analyses were also not adjusted for student socioeconomic and demographic backgrounds, school composition, and other variables reported to be associated with student achievement scores.

In the longitudinal and cross-sectional analyses, we did not attempt to determine which SSI-related factors contributed to achievement growth of the SSI states. In other chapters of this study, we will extend current approaches to assessing the effects of state policies and practices related to the SSI program on student mathematics achievement.

Appendices

Appendix A
Table 8A.1
Summary of State Means and Jackknife Estimated Standard Errors for Grade 8 Data

Table 8A.2
Summary of State Means and Jackknife Estimated Standard Errors for Grade 4 Data

Appendix A

Table 8A.1
Summary of State Means and Jackknife Estimated Standard Errors for Grade 8 Data

State ID	State Name	SSI Status	1992			1996			2000		
			Stu N	Mean	SE	Stu N	Mean	SE	Stu N	Mean	SE
1	Alabama		2522	252.19	1.66	2261	256.59	2.15	2327	262.16	1.77
4	Arizona		2617	265.37	1.26	2136	267.87	1.56	1786	270.72	1.53
5	Arkansas	✓	2556	256.31	1.19	1845	261.65	1.52	2170	261.36	1.37
6	California	✓	2516	260.89	1.66	2290	262.77	1.85	1628	262.17	2.04
9	Connecticut	✓	2613	273.74	1.14	2485	279.59	1.12	2454	281.90	1.37
13	Georgia	✓	2589	259.36	1.16	2364	262.47	1.65	2513	266.33	1.25
15	Hawaii		2454	257.41	0.86	2189	262.13	0.97	2277	262.77	1.34
18	Indiana		2659	270.10	1.14	2347	275.53	1.44	1855	283.05	1.45
21	Kentucky	✓	2756	262.24	1.11	2461	266.59	1.07	2294	271.56	1.40
22	Louisiana	✓	2582	249.98	1.66	2599	252.38	1.57	2359	258.98	1.50
23	Maine	✓	2464	278.64	0.98	2258	284.06	1.29	2102	283.64	1.19
24	Maryland		2399	264.83	1.28	2137	269.68	2.13	2401	276.01	1.43
25	Massachusetts	✓	2456	272.78	1.05	2280	277.57	1.74	2303	283.12	1.25
26	Michigan	✓	2616	267.35	1.39	2155	276.87	1.79	1975	278.45	1.60
27	Minnesota		2471	282.39	0.96	2425	284.05	1.34	1525	287.65	1.44
28	Mississippi		2498	246.46	1.18	2487	250.22	1.19	2394	254.03	1.30
29	Missouri		2666	271.13	1.19	2386	273.28	1.39	2329	273.58	1.46
31	Nebraska	✓	2285	277.65	1.11	2610	282.77	1.02	1916	280.62	1.12
35	New Mexico	✓	2561	259.61	0.90	2371	261.97	1.22	1919	259.84	1.74
36	New York	✓	2158	266.42	2.08	1962	270.23	1.66	1633	276.26	2.09
38	North Dakota		2314	283.21	1.14	2602	284.22	0.91	2227	283.07	1.07
45	South Carolina	✓	2625	260.77	0.97	2143	260.78	1.54	2306	266.35	1.39
47	Tennessee		2485	258.83	1.39	2300	263.12	1.40	2232	263.44	1.72
48	Texas	✓	2614	264.59	1.30	2245	270.20	1.43	2317	274.85	1.47
49	Utah		2726	274.34	0.73	2697	276.77	1.03	2472	275.44	1.16
54	West Virginia		2690	259.09	1.01	2578	264.87	1.02	2463	270.78	1.00
56	Wyoming		2444	275.08	0.86	2696	274.78	0.91	2634	276.69	1.18

Table 8A.2

Summary of State Means and Jackknife Estimated Standard Errors for Grade 4 Data

State ID	State Name	SSI Status	1992			1996			2000		
			Stu N	Mean	SE	Stu N	Mean	SE	Stu N	Mean	SE
1	Alabama		2605	208.33	1.56	2541	211.65	1.24	2438	217.94	1.41
4	Arizona		2741	215.25	1.07	2113	217.58	1.73	2082	218.77	1.42
5	Arkansas	✓	2621	210.21	0.89	2047	215.85	1.46	2262	217.06	1.13
6	California	✓	2412	208.40	1.56	2063	209.13	1.84	1656	213.57	1.84
9	Connecticut	✓	2600	226.80	1.13	2565	232.03	1.10	2499	234.24	1.16
13	Georgia	✓	2766	215.59	1.23	2542	215.46	1.49	2681	219.56	1.06
15	Hawaii		2625	214.06	1.31	2578	214.97	1.45	2439	215.85	1.15
18	Indiana		2593	221.04	1.04	2470	229.39	1.05	1864	234.42	1.08
21	Kentucky	✓	2703	215.05	1.01	2579	219.99	1.07	2275	220.99	1.17
22	Louisiana	✓	2792	204.14	1.46	2671	209.02	1.11	2513	217.96	1.40
23	Maine	✓	1898	231.64	1.00	2115	232.21	1.02	2132	230.57	0.92
24	Maryland		2844	217.32	1.28	2465	220.69	1.56	2645	222.31	1.27
25	Massachusetts	✓	2549	226.60	1.17	2497	228.97	1.35	2292	234.96	1.12
26	Michigan	✓	2412	219.88	1.71	2382	226.26	1.27	1903	230.89	1.43
27	Minnesota		2640	228.49	0.90	2425	232.19	1.08	1822	235.27	1.32
28	Mississippi		2712	201.83	1.08	2716	208.43	1.22	2831	210.97	1.07
29	Missouri		2509	222.22	1.19	2643	224.73	1.07	2330	228.55	1.19
31	Nebraska	✓	2327	225.33	1.23	2678	227.54	1.18	1396	225.95	1.72
35	New Mexico	✓	2342	213.30	1.44	2389	213.84	1.75	1933	213.87	1.48
36	New York	✓	2284	218.45	1.25	2248	222.63	1.24	1753	226.56	1.33
38	North Dakota		2193	228.66	0.77	2666	230.90	1.23	2456	230.89	0.86
45	South Carolina	✓	2771	212.50	1.08	2364	213.19	1.30	2501	220.42	1.39
47	Tennessee		2708	210.95	1.35	2473	219.18	1.40	2488	219.84	1.49
48	Texas	✓	2623	217.92	1.21	2413	228.71	1.36	2171	232.67	1.21
49	Utah		2799	224.04	0.97	2625	226.52	1.15	2639	227.29	1.22
54	West Virginia		2786	215.27	1.05	2530	223.35	1.01	2431	224.85	1.20
56	Wyoming		2605	225.38	0.93	2758	223.20	1.38	1739	229.25	1.30

CHAPTER 9

A COMPARISON OF SSI AND NON-SSI PERFORMANCE ON STATE NAEP MATHEMATICS ASSESSMENT ITEMS

A comparison of the mean mathematics composite scores of SSI states with non-SSI states on the state NAEP over multiple years indicates relatively small differences between the two groups. We detected some variations between the groups, however, when we analyzed their performance on the five content strands. Non-SSI states performed slightly higher than SSI states on the grade 8 Number and Operations content strand and on the Measurement content strand. But even though two groups may perform similarly on the average, this does not mean that students' knowledge is uniform on all aspects of mathematics. It is possible for two groups to get the same mean score on a test, but to perform differently on individual items.

In order to look more deeply into the performance of students in the two groups for the 1992, 1996, and 2000 administration of the State NAEP, we used a differential item functioning (DIF) analysis technique. DIF analysis is typically associated with detecting item bias between two groups, such as between White students and Black students or between male and female students. The technique uses statistical analyses to indicate whether a group of students performed significantly differently (lower or higher) on an item than expected, based on the total score of the full group. One assumption is that DIF is a nuisance dimension that intrudes on the ability testers intended to be measured (Ackerman, 1992; Roussos & Stout, 1996). In the present study, the focus was on whether the construct being measured in the SSI and non-SSI samples was the same. The presence of DIF items represents a difference in the construct between what is being measured and the items identified. If the set of DIF items favoring one or the other of two groups represents a meaningful and recognizable mathematical construct, it is implied that students in the favored group may have had some experience that caused them to perform better on these items than could be expected compared to students in the total group.

An underlying vision of NSF for the SSIs was that all students could achieve the ambitious mathematics outcomes described by the National Council of Teachers of Mathematics' *Curriculum and Evaluation Standards for School Mathematics* (1989) (Zucker, Shields, Adelman, Corcoran, & Goertz, 1998). The NCTM *Standards* emphasize the importance of reasoning, communication, and problem solving for the in-depth study of mathematics, along with learning to compute with numbers, analyze data, use geometry, and apply the principles of algebra. Mathematics reform strategies in many SSI states, and non-SSI states, included emphasis on these factors. Because of the strong emphasis by NSF on this content, it is reasonable to investigate whether SSI states showed greater improvement on that content given increased attention in the NCTM *Standards*. It was hypothesized that NAEP items reflecting outcomes in such areas might function differently in SSI and non-SSI states. In particular, examinees in states emphasizing reform would be expected to perform qualitatively differently on such items than examinees in states that did not.

Item Response Theory (IRT) (Hambleton & Swaminathan, 1985; Lord, 1980) models for NAEP mathematics items (Allen, Jenkins, Kulick, & Zelenak, 1997) were used to analyze items administered as part of the State NAEP assessment. Likelihood ratio tests for differential item functioning (DIF) were used to compare the performance of students in SSI and non-SSI states. The focus of this study was to determine whether patterns existed in these data that might indicate the influence of SSI participation.

Methods

Data

Data for this study were taken from the State NAEP Assessment results for grades 4 and 8 for 1992, 1996, and 2000. The sample sizes and ethnic group composition of these samples are given in Table 9.1. DIF comparisons between SSI and non-SSI samples were conducted on these data. It is evident that the sample sizes are quite large. The states included in the analyses are those states for which data were available for all three years and is the longitudinal sample used in the other analyses performed by this project. There were 14 states in the SSI sample and 13 states in the non-SSI sample (see Table 9.2).

The items in the State NAEP assessment were administered in blocks. Each student received only a portion of the total block of items available for that year and grade level. In most cases, students received three blocks of between 10 and 15 items each. Some blocks consisted of multiple-choice items, some consisted of constructed response items and others contained both multiple-choice and constructed-response items. Some of the constructed-response items were considered short enough to fit with a 2-parameter IRT model (2PL), while other constructed-response items were fit with a graded-response model. The IRT models used in this study were those described for each item in Allen et al. (1997). The numbers of items and blocks by grade and year are given in Table 9.3.

It can be seen in Table 9.3 that the number of blocks was generally consistent across years. Thirteen blocks were provided in each of the other grade-by-year administrations. The number of items changed, however, somewhat markedly. The largest percentage of items for any one year was allocated to measuring content area 1 (Number and Operations), although, from 1992 to 2000, the proportion of grade 8 items in the 13 blocks allocated to measuring Number and Operations declined from 32% in 1992 to 26% in 2000. Over this period, the proportion of grade 8 items in those same blocks allocated to measuring Algebra and Functions increased slightly, from 16% in 1992 to 24% in 2000. The proportion of items allocated to the other three content areas in the seven blocks remained fairly constant from 1992 to 2000. Over time, process areas 1 (conceptual understanding) and 3 (problem solving) received more emphasis than procedural knowledge and in 2000 had the highest proportion of items across both grades. The most frequent type of item was clearly the multiple-choice item.

Table 9.1
Sample Sizes of NAEP State Assessments

		Caucasian	African-American	Hispanic	Asian and Pacific Islander	Native American	Unclassified	Total
1992 Grade 4	SSI	67,338	17,841	15,504	2,484	2,010	123	105,300
	Non-SSI	71,097	14,196	8,718	5,982	2,871	216	103,080
1996 Grade 4	SSI	64,775	16,653	14,262	2,235	2,568	162	100,655
	Non-SSI	67,113	14,220	9,372	5,403	2,664	237	99,009
2000 Grade 4	SSI	54,888	16,509	13,797	2,100	2,247	192	89,733
	Non-SSI	59,418	14,037	8,334	5,991	2,586	153	90,519
1992 Grade 8	SSI	71,454	16,572	14,045	2,442	1,392	267	106,172
	Non-SSI	71,725	12,138	6,762	5,976	2,064	168	98,833
1996 Grade 8	SSI	63,269	15,195	12,911	2,678	1,735	195	95,983
	Non-SSI	67,315	12,000	6,723	5,473	1,837	200	93,548
2000 Grade 8	SSI	57,193	15,849	12,349	2,571	1,500	126	89,588
	Non-SSI	60,837	11,451	6,240	6,555	1,572	87	86,742

Table 9.2
SSI and Non-SSI States in the Sample

SSI States	Non-SSI States
Arkansas	Alabama
California	Arizona
Connecticut	Hawaii
Georgia	Indiana
Kentucky	Maryland
Louisiana	Minnesota
Maine	Mississippi
Massachusetts	Missouri
Michigan	North Dakota
Nebraska	Tennessee
New Mexico	Utah
New York	West Virginia
South Carolina	Wyoming
Texas	

Table 9.3
Number of Items by Item Block, Content, Process, and Type

Year		1992		1996		2000	
Grade		4th	8th	4th	8th	4th	8th
Total number of items (Number of blocks)		156 (13)	183 (13)	144 (13)	162 (13)	145 (13)	160 (13)
Content	1 Number & Operations	63	58	59	46	58	42
	2 Measurement	29	32	25	27	27	24
	3 Geometry	27	36	25	31	25	31
	4 Data Analysis, Statistics, & Probability	20	28	17	25	14	24
	5 Algebra & Functions	17	29	18	33	21	39
Process	1 Conceptual Understanding	64	67	59	59	61	59
	2 procedural knowledge	31	45	30	44	36	48
	3 Problem Solving	56	65	48	51	48	53
	4 Problem Solving Extended Open Question	5	6	7	8	0	0
Item Type	1 Open Ended	54	59	57	61	60	62
	2 Multiple Choice	97	118	80	93	85	98
	3 Extended Open Ended	5	6	7	8	0	0

NAEP Item Categories

The item categories for the State NAEP assessments are given in Tables 9.4 and 9.5 by grade and by year. NAEP categorizes items as belonging to one of five content areas for grade 4[t] and 8 mathematics: (1) Number and Operations, (2) Measurement, (3) Geometry, (4) Data Analysis, Statistics, and Probability, and (5) Algebra and Functions. Four process categories are identified for these same items: (1) conceptual understanding, (2) procedural knowledge, (3) problem solving (these are multiple-choice items), and (4) problem solving with extended response (typically, these ask students to show their work). The content and process breakdowns by item format are given in Table 9.4 for grade 4 and in Table 9.5 for grade 8 for each year. When possible, we have used this designation to identify patterns of DIF in the SSI versus non-SSI comparisons.

Table 9.4
Grade 4 Item Type Designations by Year, Strand, Process, and Item Type

Table 9.4a
Grade 4, 1992 (content by item type)

Item type / Content	1 (Open ended)	2 (Multiple choice)	3 (Ext. open ended)	Total
1 (Number & Operations)	20	41	2	63
2 (Measurement)	7	22	0	29
3 (Geometry)	12	14	1	27
4 (Data Analysis, Stat. & Prob.)	9	10	1	20
5 (Algebra & Functions)	6	10	1	17
Total	54	97	5	156

Table 9.4b
Grade 4, 1992 (process by item type)

Item type / Process	1 (Open ended)	2 (Multiple choice)	3 (Ext. open ended)	Total
1 (Conceptual Understanding)	17	47	0	64
2 (Procedural Knowledge)	14	17	0	31
3 (Problem Solving)	23	33	0	56
4 (Prob. Solving with Ext.)	0	0	5	5
Total	54	97	5	156

Table 9.4c
Grade 4, 1996 (content by item type)

Item type / Content	1 (Open ended)	2 (Multiple choice)	3 (Ext. open ended)	Total
1 (Number & Operations)	24	34	1	59
2 (Measurement)	6	19	0	25
3 (Geometry)	13	10	2	25
4 (Data Analysis, Stat. & Prob.)	6	9	2	17
5 (Algebra & Functions)	8	8	2	18
Total	57	80	7	144

Table 9.4d
Grade 4, 1996 (process by item type)

Item type / Process	1 (Open ended)	2 (Multiple choice)	3 (Ext. open ended)	Total
1 (Conceptual Understanding)	18	41	0	59
2 (Procedural Knowledge)	13	17	0	30
3 (Problem Solving)	26	22	0	48
4 (Prob. Solving with Ext.)	0	0	7	7
Total	57	80	7	144

Table 9.4e
Grade 4, 2000 (content by item type)

Item type / Content	1 (Open ended)	2 (Multiple choice)	3 (Ext. open ended)	Total
1 (Number & Operations)	24	34	0	58
2 (Measurement)	5	22	0	27
3 (Geometry)	13	12	0	25
4 (Data Analysis, Stat. & Prob.)	7	7	0	14
5 (Algebra & Functions)	11	10	0	21
Total	60	85	0	145

Table 9.4f
Grade 4, 2000 (process by item type)

Item type / Process	1 (Open ended)	2 (Multiple choice)	3 (Ext. open ended)	Total
1 (Conceptual Understanding)	19	42	0	61
2 (Procedural Knowledge)	16	20	0	36
3 (Problem Solving)	25	23	0	48
4 (Prob. Solving with Ext.)	0	0	0	0
Total	60	85	0	145

Table 9.5
Grade 8 Item Type Designations

Table 9.5a
Grade 8, 1992 (content by item type)

Item type / Content	1 (Open ended)	2 (Multiple choice)	3 (Ext. open ended)	Total
1 (Number & Operations)	15	41	2	58
2 (Measurement)	12	19	1	32
3 (Geometry)	15	20	1	36
4 (Data Analysis, Stat. & Prob.)	10	17	1	28
5 (Algebra & Functions)	7	21	1	29
Total	59	118	6	183

Table 9.5b
Grade 8, 1992 (process by item type)

Item type / Process	1 (Open ended)	2 (Multiple choice)	3 (Ext. open ended)	Total
1 (Conceptual Understanding)	15	52	0	67
2 (Procedural Knowledge)	14	31	0	45
3 (Problem Solving)	30	35	0	65
4 (Prob. Solving with Ext.)	0	0	6	6
Total	59	118	6	183

Table 9.5c
Grade 8, 1996 (content by item type)

Item type / Content	1 (Open ended)	2 (Multiple choice)	3 (Ext. open ended)	Total
1 (Number & Operations)	14	30	2	46
2 (Measurement)	10	15	2	27
3 (Geometry)	15	15	1	31
4 (Data Analysis, Stat. & Prob.)	10	12	3	25
5 (Algebra & Functions)	12	21	0	33
Total	61	93	8	162

Table 9.5d
Grade 8, 1996 (process by item type)

Item type / Process	1 (Open ended)	2 (Multiple choice)	3 (Ext. open ended)	Total
1 (Conceptual Understanding)	18	41	0	59
2 (Procedural Knowledge)	11	33	0	44
3 (Problem Solving)	32	19	0	51
4 (Prob. Solving with Ext.)	0	0	8	8
Total	61	93	8	162

Table 5, continued:

Table 9.5e
Grade 8, 2000 (content by item type)

Item type / Content	1 (Open ended)	2 (Multiple choice)	3 (Ext. open ended)	Total
1 (Number & Operations)	12	30	0	42
2 (Measurement)	8	16	0	24
3 (Geometry)	14	17	0	31
4 (Data Analysis, Stat. & Prob.)	14	10	0	24
5 (Algebra & Functions)	14	25	0	39
Total	62	98	0	160

Table 9.5f
Grade 8, 2000 (process by item type)

Item type / Process	1 (Open ended)	2 (Multiple choice)	3 (Ext. open ended)	Total
1 (Conceptual Understanding)	21	38	0	59
2 (Procedural Knowledge)	9	39	0	48
3 (Problem Solving)	32	21	0	53
4 (Prob. Solving with ext.)	0	0	0	0
Total	62	98	0	160

Detection of DIF

Detection of DIF was done in the context of IRT (Hambleton & Swaminathan, 1985; Lord, 1980) using the likelihood ratio test for DIF (Thissen, Steinberg, & Wainer, 1988, 1993). The IRT models used in this study were the 2- and 3-parameter logistic models and the graded- response models as described by Allen et al. (1997) for each of the items. Previous research has shown that the likelihood ratio test for DIF controls Type I error at the item level for the 2- and 3-parameter IRT dichotomous models (Cohen, Kim, & Wollack, 1996) and for the graded- response models (Kim & Cohen, 1998). The likelihood ratio test for DIF was conducted using output from the computer program MULTILOG (Thissen, 1991). All DIF results were obtained within blocks for both dichotomous and constructed-response items.

Results

The presence of DIF items in this study indicates a difference in the underlying construct being measured in the SSI and non-SSI groups. Our focus is on the use of DIF items as an indication of the impact of the SSI initiative on the mathematics construct in grades 4 and 8. In this section, we discuss some of the possible patterns that are present in the DIF results. Grade 4 DIF items for 1992, 1996, and 2000 are identified in Table 9.6 and grade 8 items for the three years are identified in Table 9.7.

Item types that differentiated between the SSI and non-SSI groups changed over the three administrations of the State NAEP. A greater proportion of the DIF items at both grade 4 and grade 8 consisted of open-ended items (type 1) in 2000 than in the previous two years (Tables 9.6 and 9.7). The increase in open-ended DIF items was larger than the increase in the proportion of these items included on the assessment (38% to 43%) (Tables 9.4 and 9.5). At grade 4, in 1992 five of the 14 DIF items (36%) were open-ended items. In 1996, 10 of the 23 DIF items (43%) were open-ended items (type 1 or 3). In 2000 at grade 4, the number of open-ended DIF items continued to increase to 17 of 35 (48%). A similar trend occurred for grade 8, but was even more dramatic. Twelve of the 26 DIF items (46%) that distinguished between SSI and non-SSI states were open-ended items (type 1). In 1996, 25 of the 33 grade 8 DIF items were open-ended (76%). This trend continued in 2000, when 24 of the 39 DIF items (62%) were open-ended items. The trend toward a greater number of DIF items being open-ended at both grade 4 and grade 8 provides some evidence that students in the two groups of states, SSI and non-SSI, performed more similarly from 1992 to 2000 on multiple-choice items and increasingly differently on the open-ended items. This trend was more apparent at grade 8 than at grade 4.

In addition to having an increasing proportion of DIF items over time as open-ended items, there was one other way the two item types, open-ended and multiple-choice, distinguished between SSI states and non-SSI states. At grade 4 in 2000, a noticeably higher proportion of DIF items favored the SSI states that were multiple-choice items (13 DIF multiple-choice items favoring SSI states and 5 DIF multiple-choice items favoring non-SSI states). At grade 8, there was a similar trend, but not as

striking (9 DIF multiple-choice items favoring SSI states and 6 DIF multiple-choice items favoring non-SSI states). However, the distribution of multiple-choice items by group was not statistically significant using a χ^2 statistic. Other than there being a rising proportion of open-ended items identified with DIF for both groups of states and a higher proportion of DIF items favoring the SSI states that were multiple-choice items, the items by type were nearly evenly distributed between the two groups of states.

Table 9.6
DIF Items by Content, Process, and Item Type

Table 9.6a
Items Showing DIF Between SSI and Non-SSI (Grade 4, 1992)

Block	Item	Content	Process	Item-type	SSI favored	Non-SSI favored
M3	6. INTERPRET SPEEDOMETER READING	2	2	2		X
M4	9. SOLVE MULTI-STEP STORY PROBLEM	1	3	2		X
	11. SOLVE STORY PROBLEM (DIVISION)	1	1	2		X
M5	4. DRAW SQUARE W/2 CORNERS AT GIVEN POINTS	3	3	1		X
	6. TOTAL # NEWSPAPERS LEE DELIVERS IN 5 DAYS: 5 X BOX	5	1	2	X	
	9. ESTIMATE: WHEN CHEN DECIDED IF HE HAD ENOUGH MONEY	1	1	2	X	
	16. MARBLE TAKEN FROM BAG - MOST LIKELY TO BE RED	4	1	2		X
M6	2. DRAW AN OBTUSE ANGLE	3	3	1	X	
	6. COMPLETE A BAR GRAPH	4	2	1	X	
	7. READ A RULER	2	3	1	X	
M8	15. VISUALIZE WRITTEN STATEMENT	3	3	2		X
M13	4. MULTIPLY A NUMBER BY ZERO	1	1	2	X	
	11. OUTLINE FOUR-SIDED SHAPE	3	1	1		X
M15	5. SELECT REASONABLE UNIT OF MEASURE	2	1	2		X

Table 9.6b
Items Showing DIF Between SSI and Non-SSI (Grade 4, 1996)

Block	Item	Content	Process	Item-type	SSI favored	Non-SSI favored
M3	4. READ THERMOMETER	2	2	1		X
M4	2. COMPARE WEIGHTS	2	1	2		X
	7. APPLY PROPERTIES OF A CUBE	3	3	2		X
	14. SOLVE AN INEQUALITY	5	1	2	X	
M5	1. RECOGNIZE BEST MEASUREMENT UNIT	2	1	2		X
	4. LIST DIFFERENT POSSIBILITIES (COINS)	5	3	1		X
	6. COMPLETE PATTERN AND WRITE RULE	5	3	1	X	
	10. FIND ALL POSSIBLE COMBINATIONS	4	4	3	X	
M6	2. DRAW AN OBTUSE ANGLE	3	3	1		X
	3. VISUALIZE A GEOMETRIC FIGURE	3	3	1		X
	7. READ A RULER	2	3	1	X	
M8	8. INTERPRET READING ON A GAUGE	2	2	2		X
	15. VISUALIZE WRITTEN STATEMENT	3	3	2		X
M9	6. COUNT CUBES IN SOLID	2	1	2		X
M10	4. ASSEMBLE PIECES TO FORM SHAPE	3	3	1	X	
M11	3. APPLY PROPERTY OF A CUBE	3	1	2	X	
	11. FIND SIMPLE PROBABILITY	4	1	2	X	
M12	2. IDENTIFY MULTIPLE OF 5	1	1	2	X	
	9. USE PROBABILISTIC REASONING	4	4	3	X	
M13	4. MULTIPLY A NUMBER BY ZERO	1	1	2	X	
M14	6. IDENTIFY MOST LIKELY OUTCOME	4	3	2	X	
M15	3. COMPLETE GEOMETRIC PATTERN	5	3	2		X
	4. USE RULER TO DRAW TRIANGLE	3	2	1	X	

Table 9.6c
Items Showing DIF Between SSI and Non-SSI (Grade 4, 2000)

Block	Item	Content	Process	Item-type	SSI favore	Non-SSI
M3	4. READ THERMOMETER	2	2	1		X
	7. FIND AREA OF SQUARE GIVEN AREA INSCRIBED TRIANGLE	2	1	2	X	
	8. IDENTIFY NUMBER IN A	5	3	2	X	
	13.REASON USING THE CONCEPT OF FRACTIONS	1	1	1		X
M4	10.APPLY CONCEPT OF EQUALITY	5	1	2		X
	14.SOLVE AN INEQUALITY	5	1	2	X	
M5	1. RECOGNIZE BEST MEASUREMENT UNIT	2	1	2	X	
	3. IDENTIFY CORRECT ORDER	5	3	2	X	
	10.FIND ALL POSSIBLE COMBINATIONS	4	3	1	X	
M6	2. DRAW AN OBTUSE ANGLE	3	3	1	X	
	6. COMPLETE A BAR GRAPH	4	2	1	X	
	7. READ A RULER	2	3	1		X
	8. COMPLETE A LETTER PATTERN	5	3	1		X
M7	2. COMPLETE GEOMETRIC PATTERN	5	1	2		X
	4. COMPLETE RATIONAL NUMBER LINE	5	1	1		X
M8	10.FIND DIFFERENCE IN TIMES	2	3	2		X
M9	7. PLOT TWO POINTS ON A GRID	5	2	1		X
M10	2. FIND THE FIGURE THAT HAS AN ANGLE GREATER THAN 90	2	2	2	X	
M11	3. APPLY PROPERTY OF A CUBE	3	1	2	X	
	4. SUBTRACT AND ESTIMATE SOLUTION OF WORD PROBLEM	1	2	2		X
	5. APPLY PROPERTIES OF RECTANGLES	3	3	2	X	
	10. IDENTIFY POINT ON NUMBER LINE	1	1	1		X
	14.DISPLAY DATA IN PICTOGRAPH	4	2	1	X	
	15.IDENTIFY PROPERTY OF TRIANGLES	3	1	2		X
M12	4. FIND THE NUMBER OF STUDENTS ON EACH TEAM	1	2	2	X	
	7. WORK WITH LARGE NUMBERS	1	2	1		X
M13	6. DETERMINE PROBABILITY THAT ARROW STOPS ON A SPACE	4	1	2	X	
	7. COMPLETE BAR GRAPH	4	2	1	X	
	8. USE CONCEPT OF REMAINDER TO SOLVE WORD PROB.	1	3	1	X	

Block	Item	Content	Process	Item-type	SSI favore	Non-SSI
M14	4. REASON WITH NUMBER MULTIPLES	5	2	1	X	
	6. IDENTIFY MOST LIKELY OUTCOME	4	3	2	X	
	8. JUSTIFY TWO DIFFERENT INTERPRETATIONS	1	1	1		X
	10.DETERMINE RELATIONSHIP BETWEEN TWO VARIABLES	5	3	1	X	
M15	3.COMPLETE GEOMETRIC PATTERN	5	3	2	X	
	5.USE PLACE VALUE	1	1	2	X	

Table 9.6d
DIF and Non-DIF Between SSI and Non-SSI (Grade 4, 1992 and 1996)

1992	1996	# of items
SSI favored DIF	SSI favored DIF	2
	Non-SSI favored DIF	1
	Non-DIF	1
Non-SSI favored DIF	SSI favored DIF	0
	Non-SSI favored DIF	1
	Non-DIF	4
Non-DIF	SSI favored DIF	4
	Non-SSI favored DIF	6
	Non-DIF	78
Total		97

Table 9.6e
DIF and Non-DIF Between SSI and Non-SSI (Grade 4, 1996 and 2000)

1996	2000	# of items
SSI favored DIF	SSI favored DIF	4
	Non-SSI favored DIF	1
	Non-DIF	4
Non-SSI favored DIF	SSI favored DIF	3
	Non-SSI favored DIF	1
	Non-DIF	6
Non-DIF	SSI favored DIF	12
	Non-SSI favored DIF	10
	Non-DIF	78
Total		119

Table 9.6f
DIF and Non-DIF Between SSI and Non-SSI (Grade 4, 1992, 1996, and 2000)

1992	1996	2000	# of items
SSI favored DIF	SSI favored DIF	SSI favored DIF	0
		Non-SSI favored DIF	1
		Non-DIF	1
	Non-SSI favored DIF	SSI favored DIF	1
		Non-SSI favored DIF	0
		Non-DIF	0
	Non-DIF	SSI favored DIF	1
		Non-SSI favored DIF	0
		Non-DIF	0
Non-SSI favored DIF	SSI favored DIF	SSI favored DIF	0
		Non-SSI favored DIF	0
		Non-DIF	0
	Non-SSI favored DIF	SSI favored DIF	0
		Non-SSI favored DIF	0
		Non-DIF	1
	Non-DIF	SSI favored DIF	0
		Non-SSI favored DIF	0
		Non-DIF	4
Non-DIF	SSI favored DIF	SSI favored DIF	2
		Non-SSI favored DIF	0
		Non-DIF	1
	Non-SSI favored DIF	SSI favored DIF	0
		Non-SSI favored DIF	1
		Non-DIF	4
	Non-DIF	SSI favored DIF	7
		Non-SSI favored DIF	7
		Non-DIF	50
Total			81

Table 9.7
Grade 8 DIF Items by Content, Process, and Item Type

Table 9.7a
Items Showing DIF Between SSI and Non-SSI (Grade 8, 1992)

Block	Item	Content	Process	Item-type	SSI favored	Non-SSI favored
M3	3. MULTIPLY TWO NEGATIVE INTEGERS	1	2	2		X
	5. FIND AMOUNT OF RESTAURANT TIP	1	1	2	X	
	7. READ DIALS ON A METER	2	2	2		X
	10. FIND NUMBER DIAGONALS-POLYGON	5	3	1	X	
M4	3. APPLY TRANSFORMATIONAL GEOMETRY	3	1	2	X	
	7. APPLY PROPERTIES OF A CUBE	3	3	2		X
	12. SOLVE STORY PROBLEM (FRACTIONS)	1	2	2		X
	18. INTERPRET MEASUREMENT TOLERANCE	2	1	2	X	
M5	16. MARBLE TAKEN FROM BAG - MOST LIKELY TO BE RED	4	1	2	X	
	18. 500 BATTERIES, 2 OUT OF 25 DEAD - 40 DEAD TOTAL	4	3	2	X	
M6	2. DRAW AN OBTUSE ANGLE	3	3	1	X	
	5. APPLY PART-WHOLE RELATIONSHIP	1	1	1		X
	7. READ A RULER	2	3	1		X
	11. DRAW A LINE OF SYMMETRY	3	1	1	X	
M8	3. FIND CHECKBOOK BALANCE	1	3	2		X
	9. FIND AN AVERAGE	4	3	1	X	
M9	2. SOLVE AND EXPLAIN NUMBER PROBL	1	3	1	X	
	5. USE DATA FROM A LINE GRAPH	4	2	2		X
M10	2. ASSEMBLE PIECES TO FORM SHAPE	3	3	1	X	
	5. COMPARE AREAS OF TWO SHAPES	2	3	1		X
	6. COMPARE PERIMETERS OF SHAPES	2	3	1	X	
	7. USE DATA FROM A CHART	4	3	1		X
M11	12. CONVERT HOURS TO ACTUAL TIME	2	3	2		X
	17. FIND SIMPLE PROBABILITY	4	1	2		X
M13	10. LOCATE OBJECT ON A GRID	5	3	1		X
M14	1. IDENTIFY THE COMMON MULTIPLE OF TWO NUMBERS	1	1	2	X	

Table 9.7b

Items Showing DIF Between SSI and Non-SSI (Grade 8, 1996)

Block	Item	Content	Process	Item-type	SSI favored	Non-SSI favored
M3	2. WRITE FRACTION THAT REPRESENTS SHADED REGION	1	1	2		X
	12. REASON ABOUT BETWEENNESS	3	3	1		X
	13. REASON TO MAXIMIZE DIFFERENCE	1	4	3	X	
M4	5. INTERPRET PIE CHART DATA	4	3	2	X	
	11. SOLVE STORY PROBLEM (DIVISION)	1	1	2	X	
M5	6. COMPLETE PATTERN AND WRITE RULE	5	3	1	X	
	11. INTERPRET TRIP GRAPH	4	4	3	X	
M6	1. SOLVE A NUMBER SENTENCE	5	1	1	X	
	4. APPLY PLACE VALUE	1	1	1	X	
	7. READ A RULER	2	3	1		X
	9. USE A NUMBER LINE GRAPH	1	1	1		X
	11. DRAW A LINE OF SYMMETRY	3	1	1	X	
	13. EXPLAIN SAMPLING BIAS	4	1	1	X	
	14. GRAPH AN INEQUALITY	5	1	1		X
M7	4. COMPLETE RATIONAL NUMBER LINE	5	1	1		X
	6. APPLY LINE SYMMETRY	3	3	1	X	
M8	3. FIND CHECKBOOK BALANCE	1	3	2	X	
	9. FIND AN AVERAGE	4	3	1	X	
	17. ORDER FRACTIONS	1	1	2		X
	18. CONVERT TEMPERATURES	5	2	2		X
M9	9. FIND PROBABILITY AND EXPLAIN	4	4	3	X	
M10	1. COMPARE GEOMETRIC SHAPES	3	1	1		X
	2. ASSEMBLE PIECES TO FORM SHAPE	3	3	1	X	
	7. USE DATA FROM A CHART	4	3	1		X
M11	3. APPLY PROPERTY OF A CUBE	3	1	2	X	
	10. IDENTIFY POINT ON NUMBER LINE	1	1	1		X
M12	5. USE PERCENT INCREASE	1	3	1	X	
	9. RECOGNIZE MISLEADING GRAPH	4	4	3		X
M14	2. IDENTIFY BETTER SURVEY	4	3	1	X	
	9. ANALYZE ROAD DETOUR	1	4	3		X
M15	3. DRAW A VECTOR	2	2	1	X	
	5. DRAW TWO RECTANGLES USING RULER	2	2	1		X
	6. INTERPRET MEANING OF 2X	5	1	2		X

Table 9.7c
Items Showing DIF Between SSI and Non-SSI (Grade 8, 2000)

Block	Item	Content	Process	Item-type	SSI favored	Non-SSI favored
M3	7. FIND THE AREA OF A PENTAGON	2	1	1	X	
	9. READ A COMPLEX BAR GRAPH	4	3	1		X
M4	16. FIND A MEDIAN	4	2	2	X	
M5	3. IDENTIFY CORRECT ORDER	5	3	2	X	
	6. COMPLETE PATTERN AND WRITE RULE	5	3	1	X	
	7. DRAW PATH ON GRID	5	1	1		X
	11. INTERPRET TRIP GRAPH	4	3	1		X
M6	4. APPLY PLACE VALUE	1	1	1	X	
	7. READ A RULER	2	3	1		X
	8. COMPLETE A LETTER PATTERN	5	3	1		X
	9. USE A NUMBER LINE GRAPH	1	1	1		X
	10. LIST SAMPLE SPACE	4	1	1		X
	11. DRAW A LINE OF SYMMETRY	3	1	1	X	
	13. EXPLAIN SAMPLING BIAS	4	1	1	X	
M7	10. VISUALIZE FOLDED BOXES	3	3	1	X	
M8	3. FIND CHECKBOOK BALANCE	1	3	2		X
	6. IDENTIFY TRIANGLE TYPE	3	1	2	X	
	9. FIND AN AVERAGE	4	3	1	X	
	10. FIND A PROBABILITY	4	3	1	X	
M9	3. SELECT GRAPH FOR INEQUALITY	5	2	2		X
	6. USE MEASUREMENT OF CENTRAL TENDENCY	4	1	2		X
	9. FIND PROBABILITY AND EXPLAIN	4	3	1	X	
M11	3. APPLY PROPERTY OF A CUBE	3	1	2	X	
	12. CONVERT HOURS TO ACTUAL TIME	2	3	2		X
	17. FIND SIMPLE PROBABILITY	4	3	2	X	
M12	3. FIND A PERCENT	1	2	2		X
	4. GRAPH POINTS, CONNECT, AND FIND PERIMETER	3	3	1	X	
	5. SOLVE FOR X	5	2	2	X	
	9. SOLVE PROBLEM INVOLVING POSTAGE RATES	5	3	1		X
M13	4. DETERMINE WHICH OPERATION RESULTS IN ODD INTEGERS	1	1	2		X
	6. IF SQUARE ROOT OF N = 6, THEN N =	5	1	2	X	
	7. FIND ANGLE MEASURE IN A CIRCLE USING ARC MEASURE	3	1	2	X	
	10. LOCATE THE POSITION OF AN OBJECT ON A GRID	5	3	1		X
M14	4. USE NUMERICAL REASONING	1	1	2	X	
	8. COMPARE MEAN AND MEDIAN	4	1	1	X	
	9. ANALYZE ROAD DETOUR	1	3	1		X
M15	2. USE PROPORTIONAL REASONING	5	1	1	X	
	5. DRAW TWO RECTANGLES USING RULER	2	2	1		X
	9. USE SCALE TO SOLVE TILING PROBLEM	2	3	1		X

Table 9.7d
DIF and Non-DIF Between SSI and Non-SSI (Grade 8, 1992 and 1996)

1992	1996	# of items
SSI favored DIF	SSI favored DIF	3
	Non-SSI favored DIF	0
	Non-DIF	7
Non-SSI favored DIF	SSI favored DIF	1
	Non-SSI favored DIF	2
	Non-DIF	10
Non-DIF	SSI favored DIF	8
	Non-SSI favored DIF	8
	Non-DIF	75
Total		114

Table 9.7e
DIF and Non-DIF Between SSI and Non-SSI (Grade 8, 1996 and 2000)

1996	2000	# of items
SSI favored DIF	SSI favored DIF	7
	Non-SSI favored DIF	2
	Non-DIF	6
Non-SSI favored DIF	SSI favored DIF	0
	Non-SSI favored DIF	4
	Non-DIF	6
Non-DIF	SSI favored DIF	11
	Non-SSI favored DIF	9
	Non-DIF	88
Total		133

Table 9.7f
DIF and Non-DIF Items Between SSI and Non-SSI (Grade 8, 1992, 1996, and 2000)

1992	1996	2000	# of items
SSI favored DIF	SSI favored DIF	SSI favored DIF	2
		Non-SSI favored DIF	0
		Non-DIF	0
	Non-SSI favored DIF	SSI favored DIF	0
		Non-SSI favored DIF	0
		Non-DIF	0
	Non-DIF	SSI favored DIF	0
		Non-SSI favored DIF	0
		Non-DIF	4
Non-SSI favored DIF	SSI favored DIF	SSI favored DIF	0
		Non-SSI favored DIF	1
		Non-DIF	0
	Non-SSI favored DIF	SSI favored DIF	0
		Non-SSI favored DIF	1
		Non-DIF	0
	Non-DIF	SSI favored DIF	1
		Non-SSI favored DIF	2
		Non-DIF	4
Non-DIF	SSI favored DIF	SSI favored DIF	4
		Non-SSI favored DIF	0
		Non-DIF	3
	Non-SSI favored DIF	SSI favored DIF	0
		Non-SSI favored DIF	1
		Non-DIF	4
	Non-DIF	SSI favored DIF	5
		Non-SSI favored DIF	5
		Non-DIF	57
Total			94

Grade 4 Items

At grade 4 in 1992, the performance of students in the SSI states was slightly different from those in the non-SSI states on both conceptual understanding items and procedural knowledge items (Tables 9.8 and 9.9). (See also Figure 9.4 at the end of the chapter.) There were a total of 13 DIF items for conceptual understanding with eight favoring SSI states and five DIF items for procedural knowledge with four favoring SSI states. This difference was removed in 1996 but continued in 2000 with the number of DIF items for each of these abilities being about the same (Tables 9.10 to 9.13 and Figures 9.5 and 9.6). Notably, in 2000 students from SSI states compared to students of equal ability from non-SSI states performed better on a greater number of problem-

solving items. Of the 12 problem-solving DIF items, nine favored the SSI states. One interpretation for this trend is that the students in SSI states, in 1992, prior to the states' full implementation of the systemic initiative program, demonstrated slight advantage on conceptual understanding and procedural knowledge items. However, this advantage disappeared after four years of the SSI program, with both students from SSI states and those from non-SSI states performing about the same on all three categories of mathematical abilities. Eight years after the initiation of the SSI programs, more of the DIF items categorized as problem solving favored students from the SSI states.

By mathematics topic, in 1992, grade 4 students from the SSI states had a greater number of DIF items favoring them in the area of Number and Operations, whereas the students from non-SSI states had a greater number of DIF items favoring them in the area of Measurement (Tables 9.8 and 9.9 and Figure 9.1). In 1996, students from the non-SSI states continued their advantage in Measurement, but grade 4 students from the SSI states had a higher number of DIF items favoring them in the area of Data Analysis, Statistics, and Probability (Tables 9.10 and 9.11 and Figure 9.2). The advantage students from the SSI states had in data analysis continued and even increased by two items in 2000, while students from the non-SSI states did not have a noticeable (three or more items) advantage in any of the five mathematics topics (Tables 9.12 and 9.13 and Figure 9.3). Eight years after the initiation of the SSI program, whatever advantage students from non-SSI states had in measurement disappeared, while students from SSI states appeared to gain an advantage in the area of Data Analysis, Statistics, and Probability.

Tracking specific DIF items across the three testing times provides some insight into the differences in the underlying constructs of mathematics that students in each group of states had. Tables 9.6d, 9.6e, and 9.6f lists the number of grade 4 items by DIF status for different combinations of years—1992 and 1996, 1996 and 2000, and for all three years. Not all of the items were given all three years, so it is not possible to track all of the DIF items for 1992, 1996, and 2000. For example, there were 97 common items administered in 1992 and 1996, 119 items in 1996 and 2000, and 81 items in all three years. As evident in Table 9.6f, there were very few cells with multiple items, indicating some clustering of items. In order to learn more about how what students in SSI states knew about mathematics changed over time, it is particularly instructive to look at DIF items that first favored non-SSI students in 1992, early in the SSI program, and then either favored SSI students in 1996 and 2000, during and after the SSI program, or were non-DIF items. When a DIF item in one year favoring non-SSI students is a non-DIF item in the next administration, this is a positive finding for the SSI states because it signifies that the students in the SSI states were performing similarly to students in the non-SSI states. It also is informative to consider non-DIF items in 1992 or 1996 that then favored students from SSI states in either 1996 or 2000, respectively. There were two items, Block M4 Item 14 and Block M11 Item 3, on which the performance of the students from SSI states and students from non-SSI states was the same in 1992, but were DIF items that favored students from SSI states in both 1996 and 2000. This indicates that the performance on these items favoring the SSI states was sustained over four years. Both of these items were classified as procedural knowledge and required students to recall factual information. Item M4-14 (Algebra and Functions) asked students to solve

an inequality and Item M11-3 (Geometry) required students to apply properties of a cube. There were no DIF items that favored the non-SSI states in 1992 and then favored the SSI states in 1996 (Table 9.6d). However, there were four DIF items that favored the non-SSI states in 1992 that were not DIF items in 1996. These items can be interpreted as items that indicated some change in the performance of students from SSI states because their performance became more similar to the performance of students in the non-SSI states. These four items required students to recall a fact or concept requiring some interpretation:

Item M3-6 (Measurement) interpret a speed odometer
Item M4-9 (Number and Operations) solve a multi-step story problem
Item M4-11 (Number and Operations) solve a story problem requiring computation
Item M13-11 (Geometry) outline a four-sided shape

These four DIF items indicated some shift in the nature of performance from 1992 by students from SSI states compared to students of similar abilities from non-SSI states. Most of these items measured a range of basic skills. The skills required to work the items successfully were not restricted to one content area, but included concepts and procedures from four of the five content areas.

The collection of eight items that represented some sustainability in performance by students in the SSI states from 1996 to 2000 included three Data Analysis, Statistics, and Probability items and two Algebra and Functions items. Both of these content areas were given more emphasis in reform mathematics. Four DIF items favored SSI states in 1996 and in 2000 (Table 9.6e):

Item M4-14 (Algebra and Functions) solve an inequality
Item M5-10 (Statistics and Probability) find all possible combinations
Item M11-3 (Geometry) apply properties of a cube
Item M14-6 (Statistics and Probability) identify most likely outcome

Four other DIF items favored SSI states in 1996, but were not DIF items in 2000 (Table 9.6e):

Item M5-6 (Algebra and Functions) complete pattern and write rule
Item M11-11 (Statistics and Probability) find simple probability
Item M13-4 (Number and Operations) multiply a number by zero
Item M15-4 (Geometry) use ruler to draw a triangle

The greater number of DIF items related to Algebra and Functions and Data Analysis, Statistics, and Probability is consistent with SSI states making greater improvements in reform areas. Also, three of the items were categorized as problem-solving or extended problem-solving (Items M5-10, M14-6, and M5-6). Only one of the eight items was categorized as Number and Operations, a topic that remains important but does not represent the significant changes in content emphases over the past decade. Only one DIF

item favoring SSI states in 1996 favored non-SSI states in 2000: Item M6-7 (Measurement), read a ruler. This item could indicate some reversal in the constructs that were being measured. However, this item assessed a very basic skill and did not indicate that students in non-SSI states were gaining an advantage on other than a very traditional item.

Nine items indicated some change in differential performance that favored SSI states from 1996 to 2000, compared to the more sustained performance indicated by the eight items discussed above. Three DIF items favored students in non-SSI states in 1996, but then favored students in SSI states in 2000 (Table 9.6e):

Item M5-1 (Measurement) recognize best measurement unit
Item M6-1 (Geometry) draw an obtuse graph
Item M15-3 (Algebra and Functions) complete geometric pattern

Six DIF items favored students in non-SSI states in 1996, but were not DIF items in 2000:

Item M4-2 (Measurement) compare weights
Item M4-7 (Geometry) apply properties of a cube
Item M5-4 (Algebra and Functions) list different possibilities
Item M6-3 (Geometry) visualize a geometric figure
Item M8-8 (Measurement) interpret reading a gauge
Item M8-15 (Geometry) visualize a written statement

Even though some states' SSI funding ended during the period between 1996 and 2000, these items could represent a continued impact from a state's systemic reform effort. Seven of the nine items were in the content areas of Measurement or Geometry, requiring some visualization. But what distinguishes this group of nine items most noticeably is that six of the items (all except Items M5-1, M4-2, and M8-8) were categorized as problem solving. All of the problem-solving DIF items that favored students in non-SSI states in 1996 either favored students in SSI states in 2000, or did not distinguish between the two groups of students. In 2000, 12 DIF items were categorized as problem solving. Nine of those 12 favored students in SSI states (Tables 9.12 and 9.13). The lack of sustaining differentiation on these items by students in the non-SSI states from 1996 to 2000 indicates that the students in the SSI states made some improvements in the area of problem solving along with the problem solving DIF items that favored the SSI states in 2000.

In summary, the DIF analysis for grade 4 indicated some change in the underlying mathematical constructs of students from the 14 SSI states compared to students of equal abilities in the 13 non-SSI states. In 1992, there were few differences in the performance by the two groups. In 1996, some differences appeared as a greater number of Measurement DIF items favored students from non-SSI states and a greater number of Data Analysis, Statistics, and Probability items favored students from SSI states. Along with these differences, from 1992 to 1996, SSI students improved their performance on

some basic skills compared to the non-SSI students, which they sustained to 2000. Between 1996 and 2000, students from SSI states sustained their advantage on Data Analysis, Statistics, and Probability items, while increasing their advantage over students from non-SSI states on problem-solving. Students from SSI states also showed some change in performance on Measurement and Geometry items that required some visualization. The improvement on Data Analysis, Statistics, and Probability in terms of spatial visualization, and problem solving are all consistent with increased emphasis of reform mathematics in the 1990s.

Table 9.8
Frequency of DIF Items by Topic and Process Grade 4 SSI States 1992

	Conceptual Understanding	Procedural Knowledge	Problem Solving	Total by Topic
Number & Operations	2			2
Measurement			1	1
Geometry			1	1
Data Analysis, Statistics & Probability		1		1
Algebra and Functions	1			1
Total by Process	3	1	2	6

Table 9.9
Frequency of DIF Items by Topic and Process Grade 4 Non- SSI States 1992

	Conceptual Understanding	Procedural Knowledge	Problem Solving	Total by Topic
Number & Operations	1		1	2
Measurement	1	1		2
Geometry	1		2	3
Data Analysis, Statistics & Probability	1			1
Algebra and Functions				
Total by Process	4	1	3	8

Table 9.10
Frequency of DIF Items by Topic and Process Grade 4 SSI States 1996

	Conceptual Understanding	Procedural Knowledge	Problem Solving	Problem Solv + Ext	Total by Topic
Number & Operations	2				2
Measurement			1		1
Geometry	1	1	1		3
Data Analysis, Statistics & Probability	1		1	2	4
Algebra and Functions	1		1		2
Total by Process	5	1	4	2	12

Table 9.11
Frequency of DIF Items by Topic and Process Grade 4 Non- SSI States 1996

	Conceptual Understanding	Procedural Knowledge	Problem Solving	Total by Topic
Number & Operations				
Measurement	3	2		5
Geometry			4	4
Data Analysis, Statistics & Probability				
Algebra and Functions			2	2
Total by Process	3	2	6	11

Table 9.12
Frequency of DIF Items by Topic and Process Grade 4 SSI States 2000

	Conceptual Understanding	Procedural Knowledge	Problem Solving	Total by Topic
Number & Operations	1	1	1	3
Measurement	2	1		3
Geometry	1		2	3
Data Analysis, Statistics & Probability	1	3	2	6
Algebra and Functions	1	1	4	6
Total by Process	6	6	9	21

Table 9.13
Frequency of DIF Items by Topic and Process Grade 4 Non- SSI States 2000

	Conceptual Understanding	Procedural Knowledge	Problem Solving	Total by Topic
Number & Operations	3	2		5
Measurement		1	2	3
Geometry	1			1
Data Analysis, Statistics & Probability				
Algebra and Functions	3	1	1	5
Total by Process	7	4	3	14

Grade 8 Items

In the initial stages of the SSI program in 1992, the students from the 14 SSI states included in the DIF analysis demonstrated some differences in the underlying constructs of their knowledge of mathematics when compared to students of equal abilities from the 13 non-SSI states. Students from SSI states, when compared to students of equal ability from the non-SSI states, showed higher performance on some of the Geometry items, whereas the students from the non-SSI states demonstrated a slightly higher performance on Measurement items. Based on an analysis of the 1992 grade 8 State NAEP items, comparing students from SSI states with students from non-SSI states, the DIF analysis identified five Geometry items that differentiated the two groups (Tables 9.14 and 9.15 and Figure 9.7). Four of these geometry DIF items favored the students from the 14 SSI states and only one favored the students from the 13 non-SSI states. In Measurement, in 1992 there was a total of six DIF items, four favoring students from non-SSI states and two favoring students from SSI states. When considering items classified by the process required, the students from the SSI states demonstrated greater conceptual understanding, while the students from the non-SSI states were slightly stronger on procedural knowledge items (Tables 9.14 and 9.15 and Figure 9.10). Of the eight DIF items classified as measuring conceptual understanding, six favored students from the SSI states. Of the four DIF items classified as measuring procedural knowledge, all favored the students from the non-SSI states. Thus, in 1992, the State NAEP indicates that the grade 8 students from the two groups of states did vary some in their knowledge of the underlying constructs of mathematics.

Over the next eight years, as indicated on the 1996 and 2000 State NAEP, the differentiated performance of grade 8 students, when comparing students of the same ability, indicated some change in how students from one group of states understood mathematics compared to students from the other group of states. In 1996, after three or four years of the SSI program, the performance of students from SSI states was differentiated from that of students from the non-SSI states as being better on Data Analysis, Statistics, and Probability items on items classified as problem solving items. Students from the non-SSI states did not differ greatly, by more than two items, on any of

the content or process categories. Of a total of eight DIF items measuring Data Analysis, Statistics, and Probability content, six favored students from SSI states (Tables 9.16 and 9.17 and Figure 9.8). Of a total of 11 DIF items categorized as problem solving, eight of the items favored students from SSI states (Figure 9.11). In all of the other categories of items, the proportion of DIF items were nearly equal in the two groups.[1] In 2000, the students from SSI states still retained what appeared to be an advantage in working Data Analysis, Statistics, and Probability items and also in Geometry on items requiring the skill of conceptual understanding. Students from non-SSI states had an advantage on Number and Operations items and Measurement items (Tables 9.18 and 9.19 and Figure 9.9). Of the 11 DIF items related to Data Analysis, Statistics, and Probability, seven favored students from SSI states. All six Geometry DIF items favored students from the SSI states. Eleven of the 16 DIF items categorized as conceptual understanding favored students from the SSI states (Figure 9.12). Five of the seven Number and Operations DIF items and four of the five Measurement DIF items favored students from the non-SSI states.

At grade 8, tracking specific items over the three testing times—1992,1996, and 2000—can provide further insight into how the underlying mathematical construct of students in each group changed over time. Four items indicated some sustained differences between students in the SSI states and students in the non-SSI states. In 1992, students at similar ability levels performed similarly on these items, since these were not DIF items for that administration of the State NAEP. However, in 1996 and 2000, these four items were identified as DIF items that favored students from the SSI states. Two of the items were probability and statistics items, one item was a Geometry item, and one was a Number and Operations item:

Item M6-4 Apply Place Value
Item M6-13 Explain Sampling Bias
Item M9-9 Find Probability and Explain
Item M11-3 Apply Property of a Cube

All of the items measured fairly basic skills. Three of the four items were classified as measuring conceptual understanding and one was classified as problem solving with an extension. These four items, plus another three items, were DIF items that favored students in SSI states for both 1996 and 2000. The three additional items were:

Item M5-6 Complete Pattern and Write Rule
Item M6-11 Draw a Line of Symmetry
Item M8-9 Find an Average

[1]Having any DIF items indicate some difference in performance between the two groups. Having nearly the same number of DIF items in one category does not indicate that the underlying construct measured is not different, but that the differences are more refined and would require more refined analysis to detect the differences. When the total number of DIF items differ greatly between the two groups then this would indicate difference in the performance that is more pervasive through out the category.

The set of seven items represents content knowledge, where students from SSI states consistently performed higher than students of equal abilities from non-SSI states. Three of the four items in the content area of Data Analysis, Statistics, and Probability and two in the content area of Geometry, one in the content area of Algebra and Functions, and one in the content area of Number and Operations. Four were classified as measuring conceptual knowledge, while the other three were classified as measuring problem solving (including one with extension). What is interesting is that the DIF items favoring students from SSI states included those from Data Analysis, Statistics, and Probability, an area given more emphasis in reform mathematics in the 1990s, and no procedural knowledge items. This is a finding similar to that at grade 4.

A shift in the underlying construct demonstrated by students from SSI states is also evident when looking at DIF items that first favored students from non-SSI states and then were not identified as DIF items in the next administration of the State NAEP. This pattern indicates that the performance of students in SSI states became more similar to the performance of students in the non-SSI states. Over the four-year period from 1992 to 1996, during the time when the largest number of states had SSI funding, ten grade 8 items were identified as not being DIF in 1996 that were DIF items in 1992 that favored the non-SSI states. These items measured more traditional topics. Three items assessed knowledge in the content area of Number and Operations and three items assessed knowledge in the content area of Measurement. The other four items included two in the areas of Data Analysis, Statistics, and Probability and one each in Geometry and Algebra and Functions. The items were classified as requiring procedural knowledge (four items) or problem solving (four items). These ten items thus indicate that students in SSI states changed over time, performing similarly to those in the non-SSI states in more traditional areas:

Item M3-3 Multiply Two Negative Integers
Item M3-7 Read Dials on a Meter
Item M4-7 Apply Properties of a Cube
Item M4-12 Solve Story Problem (fractions)
Item M6-5 Apply Part-Whole Relationships
Item M9-5 Use Data from A Line Graph
Item M10-5 Compared Areas of Two Shapes
Item M11-12 Convert Hours to Actual Time
Item M11-17 Find Simple Probability
Item M13-10 Locate Object on a Grid

Over the next four years, 1996 to 2000, the students from the SSI states (comparing students of equal ability) performed comparably to students from the non-SSI states in the content areas of Algebra and Functions and with items categorized as measuring conceptual understanding. In 1996, six DIF items were identified that favored students from non-SSI states. In 2000, these items did not differentiate between the performance of students from the two groups of states. These items were:

Item M6-14 Graph an Inequality
Item M7-4 Complete Rational Number Line
Item M8-17 Order Fractions
Item M8-18 Convert Temperatures
Item M11-10 Identify Point on Number Line
Item M15-6 Interpret Meaning of 2x

Of these six items, four measured knowledge of Algebra and Function and two measured knowledge of Number and Operations. Five of the six items were classified as assessing conceptual understanding. Since these six items assess fairly basic skills, the convergence of performance on these items between students from the SSI states and those from the non-SSI states indicates that the students from the SSI states became more similar to students in the non-SSI states in traditional mathematics following the heavy concentration of funding for the SSI program.

Six additional items show how the performance of students from the non-SSI states became more comparable to the performance of students from the SSI states over the same period of time, 1996 to 2000. In 1996, there were six DIF items that favored students from SSI states. Four years later, in 2000, these items were not identified as DIF items:

Item M4-5 Interpret Pie Chart Data
Item M4-11 Solve Story Problem (Division)
Item M6-1 Solve a Number Sentence
Item M7-6 Apply Line of Symmetry
Item M14-2 Identify Better Survey
Item M15-3 Draw a Vector

These six items were classified as assessing knowledge from all five of the content areas. One of the six items corresponded to each of the five content areas with two items corresponding to Data Analysis, Statistics, and Probability (Items M4-5 and M14-2). Three of the items were classified as measuring problem solving, two as measuring conceptual understanding, and one measured procedural knowledge.

Thus, there was some congruence in performance between students from SSI states and those from non-SSI states over the four years, 1996 and 2000. This was the period of time during which most SSI states would have had the opportunity to implement systemic reform on their own. This also was the period of time when some states had received funding for an additional five years to assist them in going to scale. Over this time, the performance of students became more similar mainly in the area of Algebra and Functions (five items) and Number and Operations (three items). Seven of the 12 items were classified as conceptual and three were classified as problem solving. Overall, it appears that over these four years the students from the non-SSI states became more like the students from the SSI states on problem solving items over a range of content areas and the students from the SSI states became more like those from the non-SSI states on more traditional basic skills areas.

In summary, grade 8 students from SSI states showed some capacity to sustain performance in the content areas of Data Analysis, Statistics, and Probability and in Geometry when compared to students from non-SSI states of the same abilities. The increase in the number of DIF items favoring students from SSI states from 1992 to 1996 and the number of repeating DIF items favoring SSI states in 2000 substantiates this finding. Students from SSI states also sustained their performance on conceptual understanding and problem-solving items. Even though the mean performance of SSI states lagged behind the mean performance of the non-SSI states, the DIF analysis at grade 8 indicates that students from SSI states demonstrated a more favorable performance than those of similar abilities from the non-SSI states in areas emphasized in the recommendations for needed mathematics reforms, namely Data Analysis, Statistics, and Probability and problem solving. Along with this change, the DIF analysis produced some evidence that from 1992 to 1996, students from the SSI states became more like students from non-SSI states in more traditional content areas—Number and Operations and Measurement—in procedural knowledge. From 1996 to 2000, students from SSI states became more like students from non-SSI states in achieving a conceptual understanding of Algebra. This convergence on conceptual understanding of algebra in the latter part of the last decade is consistent with the increased emphasis on students taking algebra prior to grade 9. Although, this study does not identify a causal link, the data are consistent with that of grade 8 students from SSI states comparably gaining a greater understanding of mathematics in the direction of the reform emphases.

Table 9.14
Frequency of DIF Items by Topic and Process Grade 8 SSI States 1992

	Conceptual Understanding	Procedural Knowledge	Problem Solving	Total by Topic
Number & Operations	2		1	3
Measurement	1		1	2
Geometry	2		2	4
Data Analysis, Statistics & Probability	1		2	3
Algebra and Functions			1	1
Total by Process	6		7	13

Table 9.15
Frequency of DIF Items by Topic and Process Grade 8 Non- SSI States 1992

	Conceptual Understanding	Procedural Knowledge	Problem Solving	Total by Topic
Number & Operations	1	2	1	4
Measurement		1	3	4
Geometry			1	1
Data Analysis, Statistics & Probability	1	1	1	3
Algebra and Functions			1	1
Total by Process	2	4	7	13

Table 9.16
Frequency of DIF Items by Topic and Process Grade 8 SSI States 1996

	Conceptual Understanding	Procedural Knowledge	Problem Solving	Problem Solv + Ext	Total by Topic
Number & Operations	2		2	1	5
Measurement		1			1
Geometry	2		2		4
Data Analysis, Statistics & Probability	1		3	2	6
Algebra and Functions	1		1		2
Total by Process	6	1	8	3	18

Table 9.17
Frequency of DIF Items by Topic and Process Grade 8 Non- SSI States 1996

	Conceptual Understanding	Procedural Knowledge	Problem Solving	Problem Solv + Ext	Total by Topic
Number & Operations	4			1	5
Measurement		1	1		2
Geometry	1		1		2
Data Analysis, Statistics & Probability			1	1	2
Algebra and Functions	3	1			4
Total by Process	8	2	3	2	15

Table 9.18
Frequency of DIF Items by Topic and Process Grade 8 SSI States 2000

	Conceptual Understanding	Procedural Knowledge	Problem Solving	Total by Topic
Number & Operations	2			2
Measurement	1			1
Geometry	4		2	6
Data Analysis, Statistics & Probability	2	1	4	7
Algebra and Functions	2	1	2	5
Total by Process	11	2	8	21

Table 9.19
Frequency of DIF Items by Topic and Process Grade 8 Non- SSI States 2000

	Conceptual Understanding	Procedural Knowledge	Problem Solving	Total by Topic
Number & Operations	2	1	2	5
Measurement		1	3	4
Geometry				
Data Analysis, Statistics & Probability	2		2	4
Algebra and Functions	1	1	3	5
Total by Process	5	3	10	18

Discussion

In a study of randomly distributed items, it was found that about 5% of the items could have been DIF by chance. In each of the three analyzed in this study, the total number of DIF items clearly exceeded what could be expected by chance, indicating that the how students in the two groups understood mathematics did differ at least some.

An important goal of the SSI program was for students in the SSI states to learn more challenging mathematics and science advocated by the standards movement of the 1990s. This analysis sought to detect differences between SSI states and non-SSI states in what mathematics students knew and to determine whether there was any indication that students from the SSI states were learning more challenging mathematics than those in the non-SSI states. The presence of DIF items in this analysis was interpreted as an indication of the differences in the construct being measured by the State NAEP over the two periods of time from 1992 to 1996 and from 1996 to 2000. In both grades 4 and 8, we found some differences that favored the SSI states and some that favored the non-SSI states. Even though the overall performance in both the 14 SSI states and the 13 non-SSI states was nearly the same over the period investigated, the underlying construct of the performance for the SSI states was different. The difference, which was similar for both grade 4 and grade 8, was significant enough to be consistent with the finding that students in the SSI states, when compared to students of equal ability in non-SSI states, were reflecting the greater emphasis on reform topics of Data Analysis and problem solving.

At both grade 4 and grade 8, an increasing proportion of open-ended items from 1992 to 2000 distinguished the mathematics performance of students in the SSI states and from those in the non-SSI states. However, there was not an evident pattern among the open-ended items that could be used to accurately describe the differences. It suffices to note that of the areas of mathematics in which performance of the two groups differed, an increasing proportion of the mathematics was assessed with open-ended items. This is consistent with the fact that a greater number of curricula in all of the states requires students to reason and to write explanations. Along with this trend, an increasing proportion of the DIF items that favored the SSI states were multiple-choice items, indicating that students from the SSI states were performing better in comparison with students of equal ability from the non-SSI states on items measured in the more traditional ways.

At both grade 4 and grade 8, students from SSI states were able to sustain a higher performance on Data Analysis, Statistics, and Probability items in 1996 and 2000 than was apparent in 1992. Since this topic was given increased emphasis in reform documents (National Council of Teachers of Mathematics, 1989), this sustained performance on items such as explaining sampling bias or finding and explaining a probability is consistent with students performing more favorably on a topic associated with standards-based mathematics. At both grades 4 and 8, students from SSI states also had more favorable performance on items identified as requiring problem solving, a process skill given increased emphasis in standards-based mathematics. Students from SSI states performed distinctively on items from Data Analysis, Statistics, and

Probability, while reducing the differential performance on the more traditional areas of Number and Operations and Algebra and Functions. The decrease in DIF among the items from the more traditional areas provides some indication of the narrowing of the differences in the underlying constructs measured in the SSI and non-SSI groups. It is interesting to note the strength of these findings, given the diversity in SSI implementations among the states in the SSI sample.

Achievement differences are somewhat related to these differences in that the kinds of items favoring one group over another may be a cause of some of them. Other research (see Chapter 4) has pointed to a clear and continuing convergence in mathematics achievement between SSI and non-SSI states from 1992 to 1996 with essentially no difference in 2000. In the case of the SSI versus non-SSI comparisons in this study, it appears that the construct differences (albeit not necessarily the achievement gap) may be decreasing.

This study demonstrates the viability of using DIF analysis for investigating the differences among SSI-related state reforms. The analyses were performed using the same IRT model as that used by NAEP and the same NAEP item-categorization scheme and item descriptors that were used to look for patterns in DIF items. At most, three point trends were analyzed in this study. The State NAEP 2000 data make it possible to see that differential performance in 1996 was sustained over the next four years. Although the overall performance of the two groups of states, SSI and non-SSI, did not vary greatly, there is evidence in this study that performance of students from SSI states could be distinguished from the performance of students from non-SSI states and was more reform oriented.

Figure 9.1. Patterns of DIF by content.

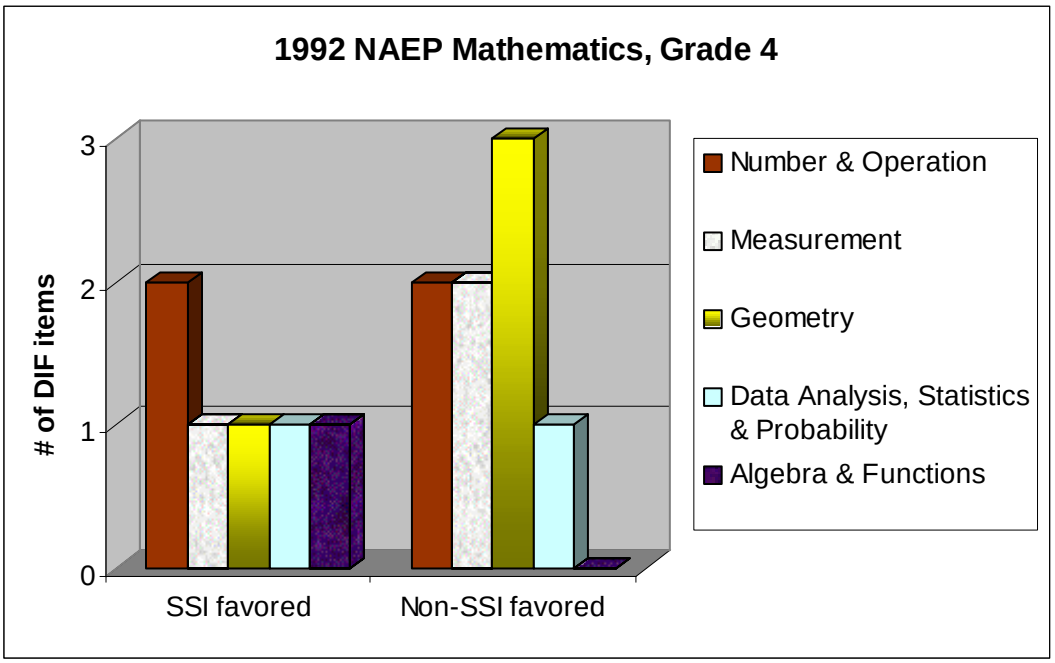


Figure 9.2. 1996 grade 4 by content.

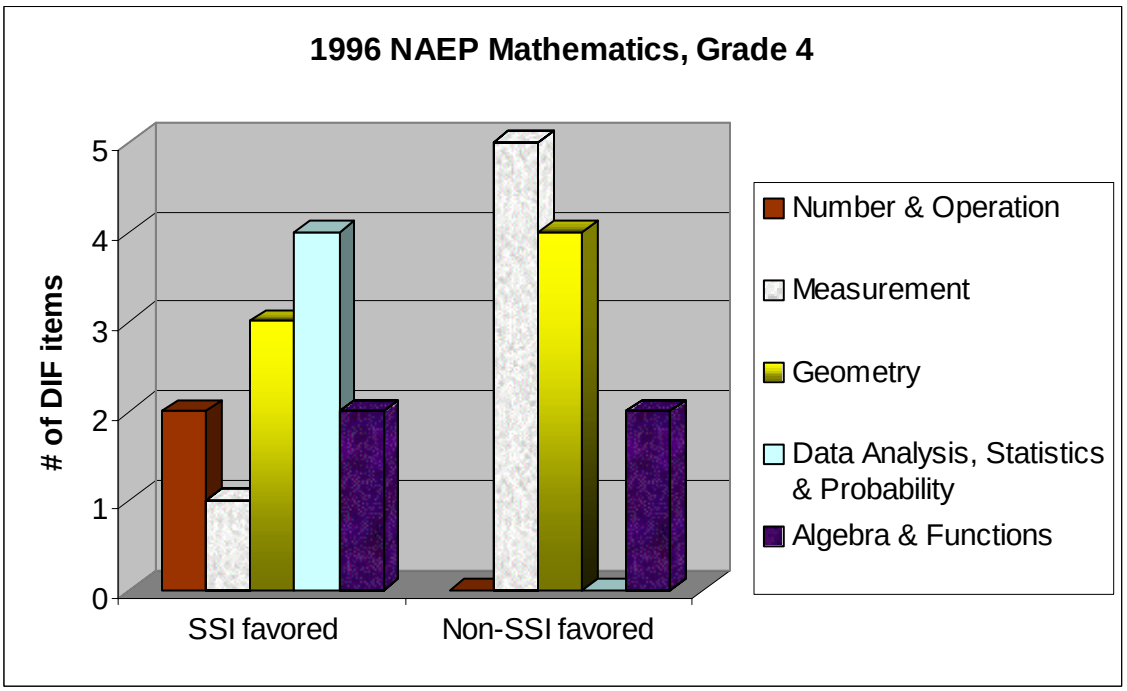


Figure 9.3. 2000 grade 4 by content.

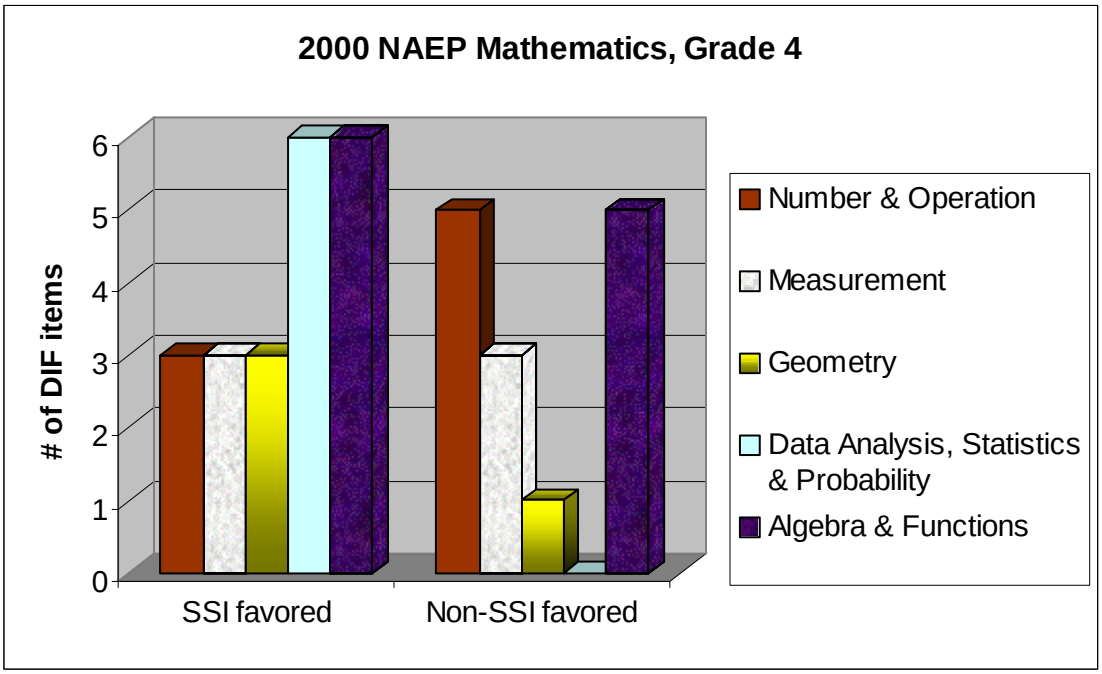


Figure 9.4. 1992 grade 4 DIF by process.

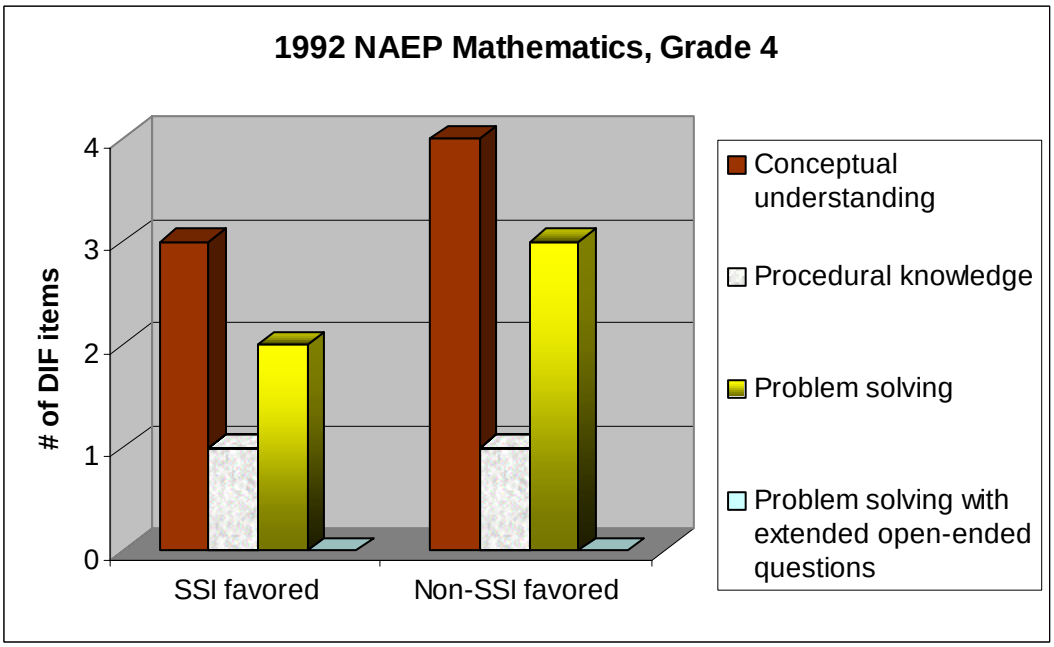


Figure 9.5. 1996 grade 4 DIF by process.

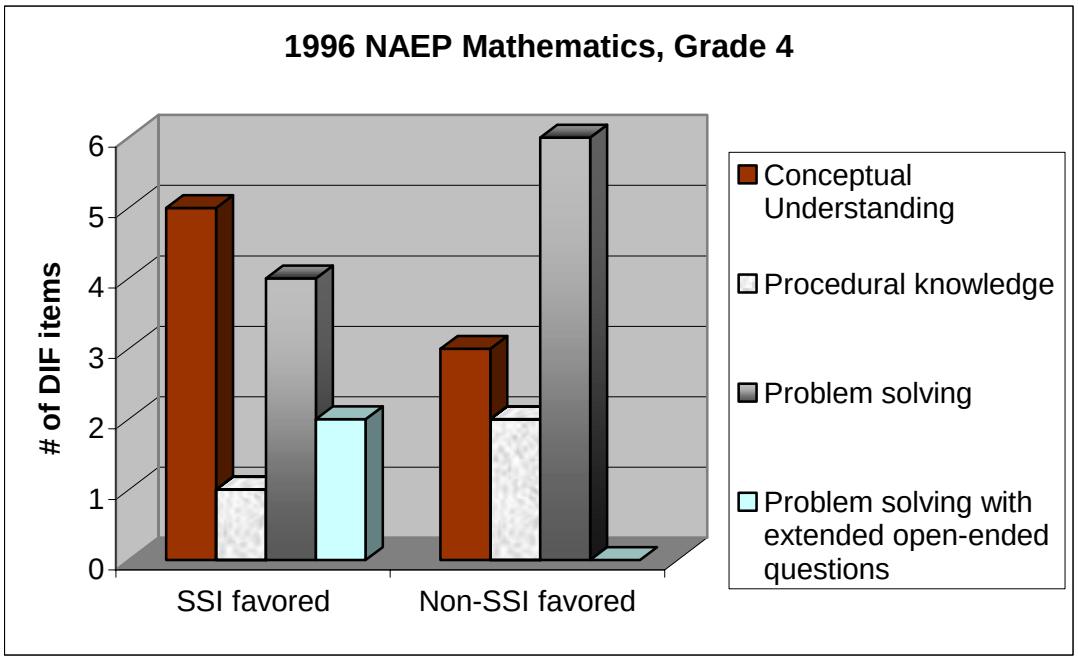


Figure 9.6. 2000 grade 4 DIF by process.

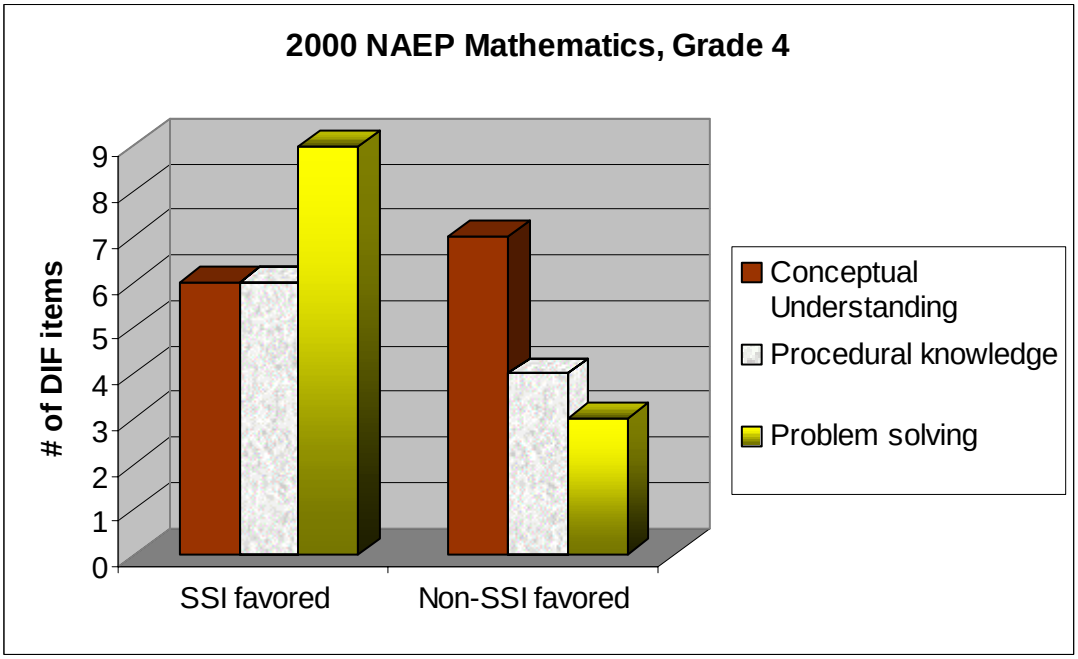


Figure 9.7. 1992 grade 8 DIF by content.

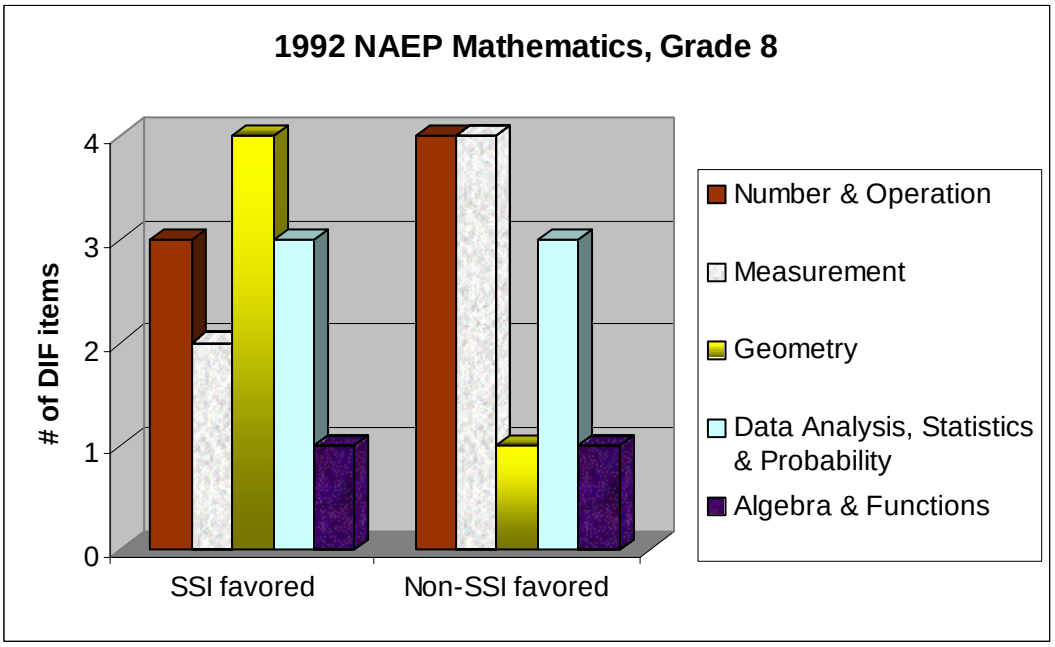


Figure 9.8. 1996 grade 8 DIF by content.

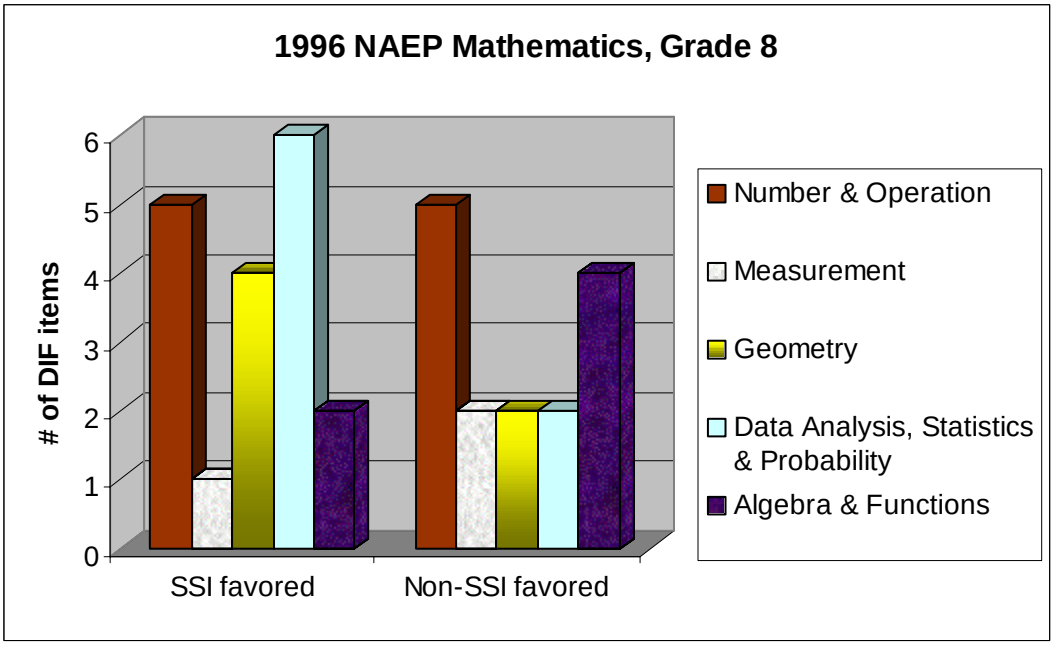


Figure 9.9. 2000 grade 8 DIF by content.

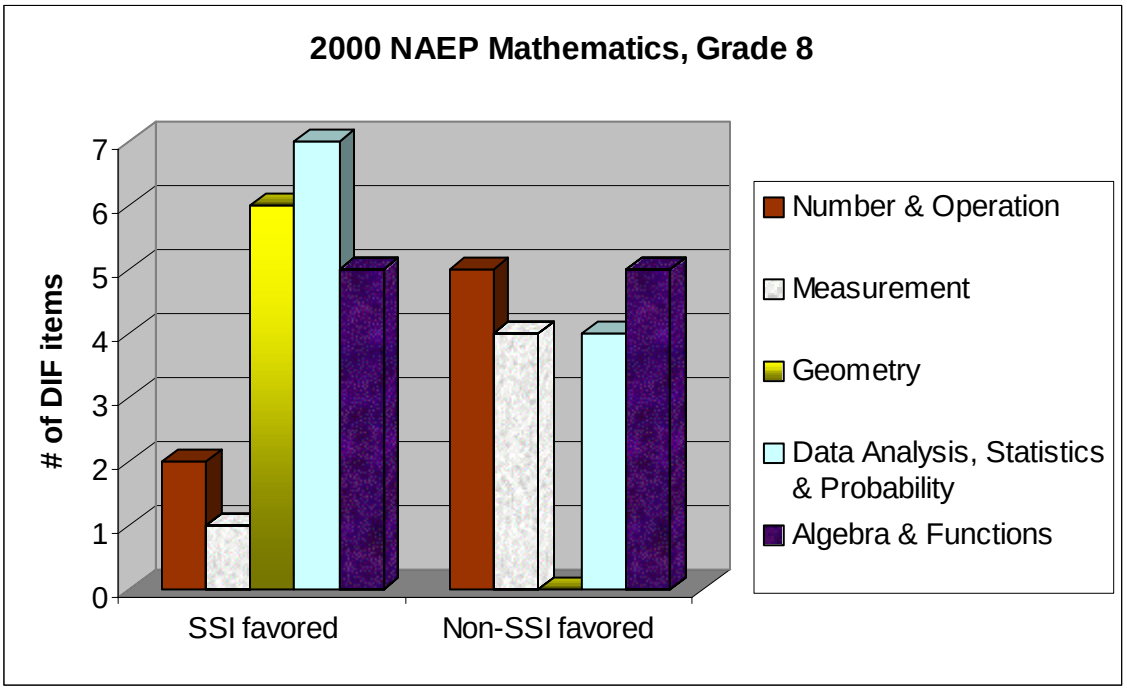


Figure 9.10. 1992 grade 8 DIF by process.

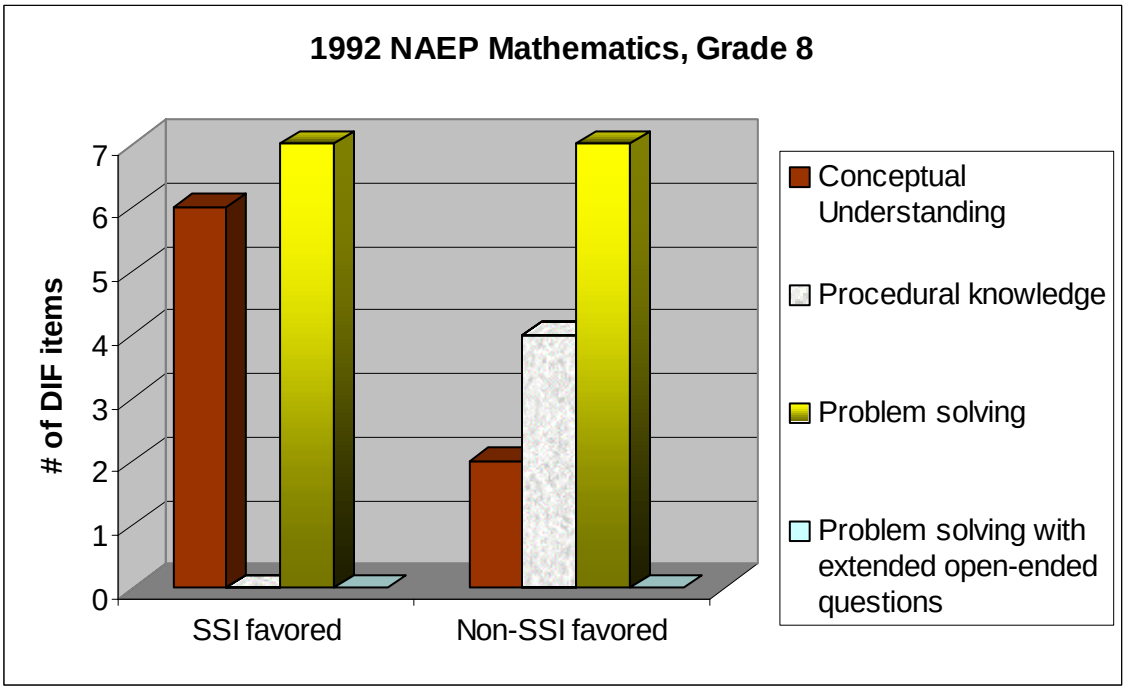


Figure 9.11. 1996 grade 8 DIF by process.

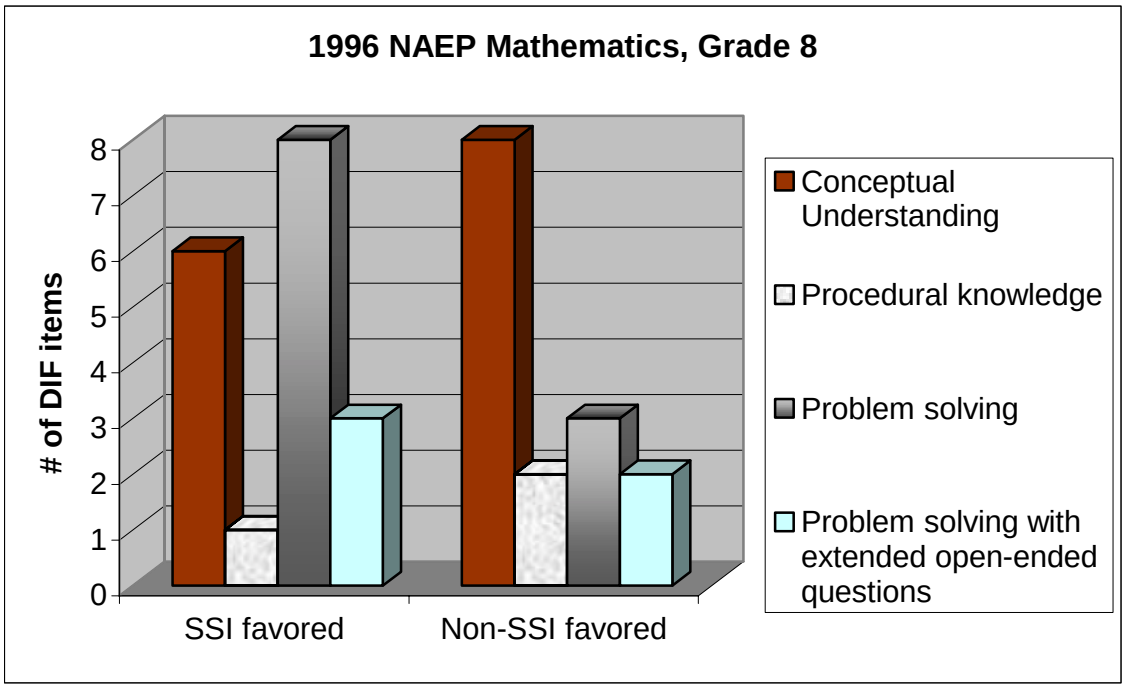
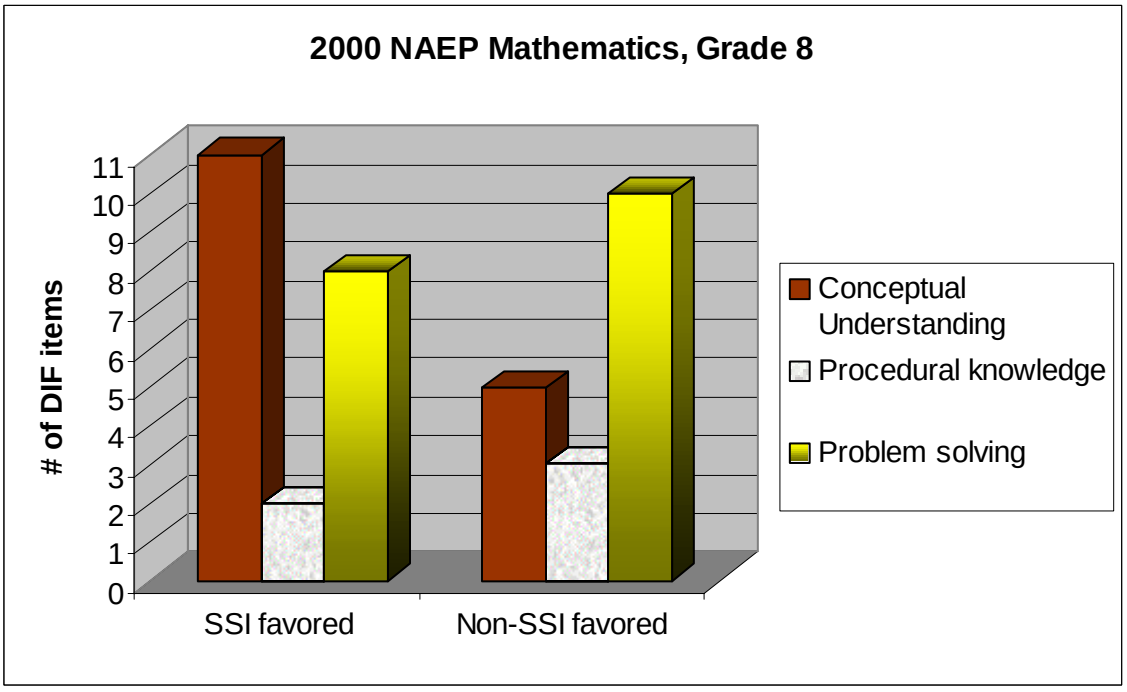


Figure 9.12. 2000 grade 8 DIF by process.



CHAPTER 10

Conclusions

Using data from the National Assessment of Educational Progress (NAEP), this study sought to provide evidence of the impact the National Science Foundation's Statewide Systemic Initiatives had on student achievement. State NAEP student mathematics achievement data from 1992, 1996, and 2000 and teacher report information on classroom practices from 1992 and 1996 were analyzed to compare student performances and practices in 14 SSI states and 13 non-SSI states. All states that consistently participated in the State NAEP assessments from 1992 to 2000 were included in the sample. These states, although not randomly chosen, represented a cross-section of the states in each group and had characteristics similar to those of all of the states in their respective groups. These states were representative of the larger group of states on variables that included average public school enrollment, per capita expenditure, percent of White students enrolled, and 1992 average mathematics achievement.

The main research question for the study was: What impact has NSF's Statewide Systemic Initiatives Program had on student learning, on student participation, and on other important variables such as classroom practices and differential performance by ethnic group? This research question was further divided into three more specific research questions:

A. What differences between SSI states and non-SSI dates were evident in mathematics achievement and student participation variables (e.g., course completion) as measured by NAEP over the period 1992-2000? What explanations exist for observable differences or for the absence of observable differences?

B. Were there improvements in statewide achievement and student participation variables for mathematics and science on multiple measures, including NAEP and state assessments, in a selected cluster of SSI and non-SSI states? What explanations are there for improvements or for no observable improvements in relation to SSI, state reform initiatives, and other activities within the states?

C. How does improvement in mathematics and science outcomes on multiple measures (e.g., state assessments and district assessments) relate to the degree of implementation of systemic reform, level of SSI participation, and other variables?

Concurrently, while Wisconsin Center for Education Research principals conducted studies that addressed these three questions, Horizon Research conducted research on the lessons that could be learned from the leaders of the SSI about designing, implementing,

evaluating, and supporting statewide systemic reform. The results and findings from their work are reported in Volume 1 of this two-volume final report.

We used a variety of analytic approaches with the State NAEP data to compare and contrast the SSI and non-SSI states. Overall achievement, as well as performance for population subgroups, was described via means and mean differences between groups. Hierarchical linear modeling was used to estimate rates of growth for each group across 1992, 1996, and 2000. Performance differences on individual NAEP items were identified using differential item functioning. Scales of reform-related instructional practices were constructed from items on the NAEP teacher questionnaire, and changes over time were examined with regression methods. Qualitative methods were used to identify differences among the SSI states and to relate these features to achievement gains. Finally, results of state assessments in three SSI states were compared to the results of State NAEP for those states. The paragraphs below outline the findings from the study of NAEP data.

Research Question A*: What differences between SSI states and non-SSI states were evident in mathematics achievement and student participation variables (course completion) as measured by NAEP over the period 1992-2000? What explanations exist for observable differences or for the absence of observable differences?*

During Phase I, the primary period of NSF's funding for the SSI program, most states improved in overall mathematics achievement at both grades 4 and 8. This was a favorable time for mathematics achievement in general, although inequities among the performance of different ethnic/racial groups remained. Across the three administrations of the State NAEP, the average mathematics composite scores of the 14 SSI states in the longitudinal analysis were nearly identical to those of the 13 non-SSI states at both grades 4 and 8. The largest difference was for grade 8 students in 1992, near the beginning of the SSI program, when students in the SSI states averaged 1.2 scale points lower on the mathematics composite score than students in the non-SSI states. For all other comparisons, the average composite scores of students from the SSI states and those from the non-SSI states varied by less than 1 scale point. At both grades, the students from SSI states had a slightly higher rate of increase in scores from 1992 to 2000 than students from non-SSI states, though the differences were not statistically significant. For grade 4, the average increase was .84 scale points per year in SSI states compared to .80 scale points per year in non-SSI states; for grade 8, the increases were .85 and .76 respectively, a slightly larger difference than at grade 4.

Lack of differences in average mathematics achievement between SSI states and non-SSI states is not surprising. States in both groups varied greatly in political context, demographics, resources, and other variables. Four years (1992–1996) may be too short an interval in which to expect significant gains in achievement from a state-level program, simply because of the massive undertaking required to reach an adequate number of teachers, have them trained, and expect them to make changes in their practices to increase student performance. Many of the SSIs reached a relatively small proportion of teachers in the state; few were able to mount an effort large enough to

change state policies to provide a context for supporting and sustaining changes in educational practices across the state.

We found some differences between SSI and non-SSI states when we looked at performance of population subgroups and by content strand. At grade 8, gains from 1992 to 2000 for White, Black, and Hispanic students were larger in SSI states than in non-SSI states. At grade 4, gains for Black students were larger in SSI states than in non-SSI states, but gains for White and Hispanic students were larger in non-SSI states. Patterns underlying the 8-year gains are different for Black and Hispanic students. For Black students, performance increased more in SSI states in non-SSI states from 1992 to 1996; it increased more in non-SSI states than in SSI states over the next four years. For Hispanic students, increases were larger in non-SSI states than in SSI states from 1992 to 1996; they were larger in SSI states than in non-SSI states over the next four years. These fluctuations may be the result of sampling error and regression toward the mean. Alternatively, they may have resulted from differences in reform strategies. SSI states, encouraged by NSF, made substantial efforts to improve mathematics achievement of minority students. SSIs in states such as Connecticut and Michigan directed resources to underperforming schools, many in inner cities with a high proportion of Black students. The noticeable gain by Black students in SSI states from 1992 and 1996, could have resulted from this focused effort. When the SSI program ended, states may not have had the resources to sustain the rate of gain for students in these schools.

Analyses of cohort gains from grade 4 in 1992 to grade 8 in 1996 found that in SSI states, Black students increased more than White students in the content strands of Number/Operations and Algebra/Functions. The relatively higher performance of Black students in these two content strands coincides with the greater emphasis placed on basic computational skills and on Algebra and Functions in the middle grades. Although not confirming, the pattern of performance among the different mathematics content strands is consistent with the impact to be expected from a reform effort. The finding that this trend did not continue over the next four years could be indicative of the difficulty in sustaining a reform effort and the fact that most of the states, such as Michigan, were not funded in Phase II.

The number of Hispanic students included in accountability systems, as in Texas, increased dramatically between 1996 and 2000 (Webb, Clune, Bolt, Gamoran, Meyer, & Thorn, 2002). For example, the percentage of grade 3 Hispanic students tested by the state in 1994 was about 29%. In 2000, the percentage of students tested in Texas who were Hispanic increased to 36%. The largest increases took place after 1996. In contrast, the proportion of Black students tested by the state of Texas remained essentially constant over this time period. The SSI in Texas spent a significant amount of its effort encouraging statewide policy changes that would result in improved performance in mathematics and science for all students. The lower 1992–1996 increase in performance by Hispanic students in SSI states, compared to those in non-SSI states, and higher increases in performance by Hispanic students in SSI states in 1996–2000 could be explained by a convergence of factors, including increased accountability, greater inclusion of Hispanic students in the accountability systems, and improved policy

structures from efforts of the SSIs. Unfortunately, we do not have sufficient information on the full implementation of the SSIs in states to completely explain the performance of Hispanic students on State NAEP.

Analyses of differential item functioning (DIF) found that when students in SSI were compared to students of equal abilities in non-SSI states, they performed more favorably on assessment items measuring Data Analysis and problem solving. These two areas in mathematics were given significant emphasis in reform documents such as the NCTM's *Curriculum and Evaluation Standards for School Mathematics* (National Council of Teachers of Mathematics, 1989). Based on responses by mathematics state leaders working with the SSIs of the 14 states included in the analysis, at least of nine of the 14 SSIs emphasized to Data Analysis as much as or more than Geometry, Measurement, or Algebra/Functions (see the State Profiles for the SSI sites on our web site, http://www.wcer.wisc.edu/SSI/Profiles/state%20profile.htm). Higher student performance on Data Analysis and problem-solving items by students from SSI states is consistent with the conclusion that these states attended more to the mathematics reforms. To the degree that SSIs advanced reform in these states, there could be a link between the SSIs and student performance. Although the design of this research can not support conclusions about causality, the results from the State NAEP and the interviews of SSI leaders support the conclusion that the mathematics curricula in SSI states emphasized Data Analysis and problem solving more than in non-SSI states. While students in SSI states performed more favorably on items from a greater number of reform areas, they performed comparably to students in non-SSI states on more traditional assessment items from 1992 to 2000. In 2000, fewer items from the more traditional content strands distinguished between students in SSI states and in non-SSI than in 1992.

We initially proposed to examine differences in course-taking patterns. However, the State NAEP did not include data on grade 12 students or high school course- taking patterns, so differences between SSI and non-SSI states on this measure were not examined.

Overall, there were few differences in student average mathematics achievement between the group of SSI states and the group of non-SSI states. Nearly all states improved throughout the 1990s. Variances in mathematics achievement among all states in both groups decreased from 1992 to 2000, indicating that over time the average performance of students among the states became more similar. Some differences in performance by subgroups of students and on specific kinds of items were consistent with the conclusion that SSIs advanced reform mathematics practices within the states and attended to the performance of minority students. However, the lack of sustained results over eight years for Black and Hispanic students in SSI states indicates the need to consider other variables in addition to SSI.

Research Question B: *Were there improvements in statewide achievement and student participation variables for mathematics and science on multiple measures, including NAEP and state assessments in a selected cluster of SSI and non-SSI states? What*

explanations are there for improvements or for no observable improvements in relation to SSI, state reform initiatives, and other activities within the states?

Obtaining accurate and complete state assessment data for the period of time that coincided with the implementation of the SSI program proved to be very difficult. For example, states changed the assessments they used, vendors retained the databases on the assessments, and data for some years of assessments were not always accessible. We were able to get some longitudinal data from the state assessments given by three SSI states—Texas, Massachusetts, and Maine.

State NAEP results did not always coincide with results from the assessments administered by these states. In eight comparisons of gain in performance over several years (from 1992 or 1994 to 1996 at grades 4 and 8 for all three states and from 1996 to 2000 at grades 4 and 8 for Texas), the State NAEP results were very similar to the state assessment results on five of the comparisons. On the other three, gains on the state assessments were larger than on the State NAEP.

These findings indicate that there is some relationship between the results from the State NAEP and those from state assessments. However, the results can vary mainly because the two types of assessments measure somewhat different content and students have a greater incentive to do their best work on the state assessments than they do on the State NAEP. Since a state assessment is aligned with curriculum practices within the state and with the SSI efforts, the state assessments should be more sensitive to measuring impact within a state attributable to an SSI than to the State NAEP. This suggests that the results from the State NAEP may be less sensitive to improvement over time. However, because the State NAEP is an external measure and, in general, divorced from state politics and the effects of "teaching to the test," the results from the national assessment may better represent the true gains by students in a state. It is impossible to disentangle these two explanations for this report. It would be possible to do a detailed content analysis of the different tests to shed some light on the difference in content coverage by each. Our analysis supports the general recommendation of the *Standards for Educational and Psychological Testing (1999)*: multiple measures are preferable to a single measure. In making use of data from state assessments and State NAEP to evaluate state programs, it is crucial to have data that can link the content topics and emphasis in the program with the content measured by each of the instruments.

(Note 1: One of the reasons for using the state data was to be able to separate out the schools/district that directly received SSI services. Do we want to talk about these results here?

Note 2: Do we want to note that these findings are based on examining mean scale scores, not the percentage of students at a given achievement level?)

Research Question C: *How does improvement in mathematics and science outcomes on multiple measures (e.g., state assessments and district assessments) relate to the degree of implementation of systemic reform, level of SSI participation, and other variables?*

One of the most challenging parts of this study was obtaining information on the degree of implementation of systemic reform by the states. For most SSI states, we interviewed key mathematics curriculum leaders in the state and for the SSI project. From their reports, we obtained information on the design of the state's SSI and its emphasis on variables relevant to the structure of the NAEP data (e.g., degree of emphasis on the five content strands—Number/Operations, Measurement, Geometry, Data Analysis, and Algebra/Functions). We also used data gathered and reported by other sources during the initial five-year funding period. We confirmed information from these sources via data collected by Horizon Research as a part of this study. We supplemented information on the SSI projects with information on state accountability systems collected and reported by CCSSO. When possible, we used information on instructional practices and teacher professional development activities from the State NAEP teacher questionnaire. This information was not available in 2000.

Results showed that SSI states were generally more oriented towards using reform practices than were non-SSI states. Teachers in SSI states reported using a greater number of reform-related instructional practices than those in non-SSI states. While teachers emphasized computational skills in all states, teachers in SSI states gave more emphasis to reasoning and communication skills than teachers in non-SSI states. Analyses found that emphasis on reasoning and communication skills was related to higher student performance in mathematics. Teachers in SSI states also reported giving students greater opportunities to use mathematical discourse. Both SSI status and use of criterion-referenced tests (CRT) were related to achievement gains across the three State NAEP administrations (1992, 1996, and 2000). The gains were the largest among states with criterion-referenced tests.

Since this was a retrospective and correlational study, we cannot attribute the use of reform practices specifically to a state's participation in the SSI program. It is possible that the states with existing reform tendencies were those that applied for SSI funding and were successful in acquiring funding. For example, in 1992, the 14 SSI states averaged slightly, but not significantly, higher than the 13 non-SSI states on the reasoning and communication skills indicator at both grades 4 and 8. By 1996, the difference was statistically significant for both grades and at grade 8 when SES was used as a covariant. However, this relationship could be due to states' disposition toward reform efforts as much as to outcomes of activities sponsored by SSI projects in states. Another plausible explanation is related to the demographics of the SSI states. The SSI program targeted states with a larger proportion of disadvantaged students. Perhaps teachers in these states were more willing to try reform practices and to participate in SSI activities, because they were dissatisfied with more traditional practices. In other words, SSI participation and reform practices could both be associated with a common factor, a higher proportion of disadvantaged students. Although we could not disentangle these alternative hypotheses, we did establish that SSI participation was related to the use of reform practices.

States with SSI programs varied greatly in student mathematics performance on State NAEP. To better understand the factors that might be related to the variability among the SSI states, we considered available data on SSI implementation in relation to

levels of performance over time. We classified the SSI states into three groups (Steady Increase, Some Increase, and Little/No Increase) based on level of mathematics performance in 1992, 1996, and 2000. Six SSI states showed steady increase at at least one grade level across both four-year periods. Four SSI states had some increase, more in one period than in the other. The remaining four SSI states had little or no increase in mathematics performance over either four-year time period. When we analyzed the SSI states in terms of these three categories, certain patterns emerged. States with accountability and assessment policies tended to increase in performance. Those states that had a strong infrastructure prior to the SSI program tended to perform at a higher level over the course of SSI funding. The approach towards reform taken by a state's SSI also appeared to be related to higher student performance. States with SSIs that attended to policy rather than only to teacher development and classroom practices performed at a higher level. When state policies did not conform with SSI policies, then reform efforts were compromised.

We also found a relationship between mathematics performance and the receipt of Phase II SSI funding. Five of the eight Phase II SSI states were among the 14 SSI states that participated in the State NAEP. The performance of these five states over the period from 1996 to 2000 accelerated more than either the nine SSI states that did not receive additional funding or the non-SSI states. Again, it is difficult to sort out whether states were selected for Phase II funding because they were higher-performing states or whether Phase II funding helped the SSI sustain an effort long enough to have a greater impact. From our perspective, there are reasons for believing that the Phase II funding may have contributed to improved performance. At grade 8, the five Phase II states performed below states with Phase I funding and non-SSI states in 1992 and 1996. However, in 2000 the average mathematics performance of these five states was higher than that of the other groups. Also, grade 4 and grade 8 students in South Carolina, one of the five Phase II states, showed essentially no growth in average mathematics performance between 1992 and 1996, but sizable growth in both grades from 1996 to 2000. South Carolina was the only SSI state of the 14 states included in the analysis that increased from essentially no growth, less than one scale point over the four years in both grades between 1992 and 1996, to gain scores of five scale points or higher between 1996 and 2000—making it clear that NSF did not take into consideration only a state's gain in performance from 1992 to 1996 as a basis for awarding Phase II funding. While we can't conclude that SSI Phase II funding was the only factor leading to improved mathematics performance between 1996 and 2000, the results support the conclusion that Phase II funding, or sustained funding, was a contributing factor.

Summary

In conclusion, the available data from the State NAEP and other sources used in this study provided a wealth of information to compare mathematics performance and reform-related educational practices in SSI and non-SSI states. Because this research was retrospective and correlational, we could not conclude that the SSI program resulted in specific increases in students' mathematics achievement. But we did identify relationships between SSI participation and mathematics performance.

SSI states and non-SSI states had comparable average scores over the duration of the study, from 1992 through 2000. The mean composite mathematics scores on the State NAEP increased steadily from 1992 to 1996 to 2000, with a slightly faster rate of increase for the SSI states. The variation among SSI states was as great as among non-SSI states. From 1996 to 2000, SSI states with Phase II funding had the largest gains, while achievement leveled off in many of the SSI states that did not receive continued funding. Teachers in SSI states used a greater number of reform practices than those in non-SSI states, and students from SSI states performed more favorably in those mathematics areas that were given greater emphasis in reform mathematics curricula—Data Analysis and problem solving. The close fit between improved performance and SSI funding suggests a possible relationship between the statewide systemic initiatives and student mathematics achievement, but it was impossible to discount other alternative hypotheses, including the possibility of selection bias.

Even though we were not able to isolate the effect of the SSI program on student performance across all SSI states, we were able to identify conditions that seemed to be associated with improved student mathematics performance in SSI states. The variation in improved student performance among SSI states appeared to be related to prior conditions in their education systems, accountability, state assessments, and duration of funding, or the capability of an SSI to create these conditions. In states that began the SSI with a relatively strong infrastructure and policy context, notable gains could result fairly quickly when the SSI worked with and strengthened the components. Without state policy and infrastructure, notable gains in statewide student achievement were unlikely to occur early in the project, but if these components were developed later gains could result.

NSF and others referred to an SSI as a catalyst for change. The results of our analyses are consistent with this characterization. Those SSIs that were able to work with other reform efforts and policies within their states were able to accomplish more. States with SSIs that were more limited in scope, or that primarily focused on changing specific sets of schools or on professional development of teachers, did not increase as much in student performance over the time studied. Furthermore, our findings are very compatible with the theory of systemic reform, which emphasizes the importance of changing multiple components in concert rather than focusing on a few isolated components.

State NAEP has been a very valuable tool in examining the effects of the SSI program. State NAEP provided a common measure for all states so we could compare states that participated in the SSI program with other states. Since State NAEP is a voluntary program, the analytic sample was necessarily limited to those states that consistently participated in State NAEP. Future analyses would be strengthened if states participated more consistently from year to year. Sampling for State NAEP is generally adequate for conclusions about the state as a whole. However, sometimes the sample for particular population subgroups is not very large resulting in relatively large confidence

intervals. State NAEP could be strengthened by using oversampling for groups of interest, as is done for national NAEP.

The State NAEP has many desirable features, including reporting results by five content strands and ability levels. However, when the results from the State NAEP are compared with results from assessments administered by states, the conclusions are not always compatible. In measuring the impact of reform, the greater the alignment of a state's assessment with curriculum emphasis, the more likely it is that the improved performance on the assessment will be related to the reform.

We have several recommendations for studying future reform based on our research conclusions. Future externally-funded large-scale reforms need to consider the entire statewide system for educational improvement and to address how their efforts will coordinate with other educational improvement efforts. Reforms are designed for the school level are not likely to result in statewide reform. Curriculum change, professional development of teachers, organizational change, and school improvement planning models individually or in concert are not adequate to bring about statewide change that leads to significant improvement in student performance. Statewide reform efforts that incorporate accountability, assessments, and other policy initiatives, along with related programmatic changes, are necessary to sustain continued improvement in student performance.

NSF's SSI program began with the assumption that effective educational reform can occur at the state level. Our results support the conclusion that state policies and statewide infrastructure can work together to improve students' mathematics achievement. When a relatively strong statewide system is in place, noticeable statewide achievement gains can result in 3 or 4 years. If statewide policy and infrastructure is relatively weak, gains may not be evident until policy and infrastructure is adequately developed. Given the challenges of statewide reform, more recent reform efforts have moved to smaller units, such as districts or individual schools. Our results suggest that even when the focus of the reform is a district or school, state policies and infrastructure are powerful influences on the success of reform.

REFERENCES

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*(1), 67-91.

Allen, N. L., Jenkins, F., Kulick, E., & Zelenak, C. A. (1997). *Technical report of the NAEP State assessment program in mathematics.* Washington, DC: National Center for Education Statistics.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: Authors.

American Psychological Association, Joint Committee on Practices. (1988). *Code of fair testing practices in education.* Washington, DC: Author.

Amrein, A. L., & Berliner, D. C. (2002). High-stakes testing, uncertainty, and student learning. *Education Policy Analysis Archives, 10*(18). Available at http://epaa.asu.edu/epaa/v10n18.

Barton, P. E. (2001). *Facing the hard facts in education reform (A Policy Information Perspective).* Princeton, NJ: Educational Testing Service.

Barton, P. E., & Coley, R. J. (1998). *Growth in school: Achievement gains from the fourth to the eighth grade* (Policy Information Report). Princeton, NJ: Educational Testing Service, Policy Information Center.

Browne, W. J., & Draper, D. (2002a). *A comparison of Bayesian and likelihood-based methods for fitting multilevel models.* Manuscript submitted for publication.

Browne, W. J., & Draper, D. (2002b). Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. *Computational Statistics, 15,* 391-420.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods.* Newbury Park, CA: Sage Publications.

Carnoy, M., Loeb, S., & Smith, T. (2001). *Do higher state test scores in Texas make for better high school outcomes?* (Report No. RR-047). Philadelphia: University of Pennsylvania, Consortium for Policy Research in Education.

Clune, W. H. (2001). Toward a theory of standards-based reform: The case of nine NSF Statewide Systemic Initiatives. In S. Fuhrman (Ed.), *From the capitol to the classroom: Standards-based reform in the states* (100th Yearbook of the National Society for the Study of Education). Chicago: University of Chicago.

Clune, W. H. (1998). *Toward a theory of systemic reform: The case of nine NSF Statewide Systemic Initiatives* (Research Monograph No. 16). Madison: University of Wisconsin, National Institute for Science Education.

Cohen, A. S., Kim, S.-H., & Wollack, J. A. (1996). An investigation of the likelihood ratio test for detection of differential item functioning. *Applied Psychological Measurement, 20*(1), 15-26.

Cohen, A. S. (1969). *Statistical power analysis for the behavioral sciences.* New York: Academic Press.

Coleman, J. S. et al. (1966). *Equality of educational opportunity.* Washington, DC: U.S. Department of Health, Education, and Welfare, Office of Education.

College Board, National Assessment Governing Board. (1997). *Mathematics framework for the 1996 National Assessment of Educational Progress.* Washington, DC: U.S. Department of Education.

College Board, National Assessment Governing Board. (1996). *Mathematics framework for the 1996 National Assessment of Educational Progress.* Washington, DC: U.S. Department of Education.

Council of Chief State School Officers, National Assessment Governing Board. (2001). *State student assessment programs annual survey. Data volumes I and II.* Washington, DC: U.S. Department of Education.

Cronbach, L. J. (1951). Coefficient alpha and internal structure of tests. *Psychometrika, 16,* 297-334.

Feuer, M. J., Holland, P. W., Green, B. F., Berthenthall, M. W., & Hemphill, F. C. (Eds.). (1999). *Uncommon measures: Equivalence and linkage among educational tests.* Washington, DC: National Academy Press, National Research Council.

Goertz, M. E., Duffy, M. C., & LeFloch, K. C. (2001). *Assessment and accountability in the fifty states: 1999–2000.* Philadelphia, PA: University of Pennsylvania, Consortium for Policy Research in Education.

Grissmer, D. W., Flanagan, A., Kawata, J., & Williamson, S. (2000). *Improving student achievement: What NAEP state test scores tell us.* Santa Monica, CA: RAND, MR-924-EDU.

Grissmer, D.W., & Ross, J. M. (Eds.). (2000). *Analytic issues in the assessment of student achievement.* NCES 2000-050. Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement, Educational Resources Information Center. ED 443-890.

Joint Committee on Testing Practices. (1988). *Code of fair testing practices in education.* Washington, DC: American Psychological Association.

Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory: Principles and applications.* Boston, MA: Kluwer-Nijhoff Publishing.

Haney, W. (2000). The myth of the Texas miracle in education. *Education Policy Analysis Archives, 8*(41). Available at http://epaa.asu.edu/epaa/v8n41.

Heck, D. J., Weiss, I. R., Boyd, S., & Howard, M. (2002). *Lessons learned about planning and implementing Statewide Systemic Initiatives in mathematics and science education.* Presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA, April 2, 2002.

Kane, J. (2002). *Using information from national databases to evaluate systemic educational reform.* A paper presented at the annual meeting of the American Evaluation Association, Arlington, VA.

Kenney, P. A., & Silver, E. A. (1998). *Content analysis project–State and NAEP mathematics assessments. North Carolina–NAEP study. Final report.* Washington, DC: National Assessment Governing Board.

Kim, S.-H., & Cohen, A. S. (1998). Detection of differential item functioning under the graded response model with the likelihood ratio test. *Applied Psychological Measurement, 22*(4), 345-355.

Klein, S. P., Hamilton, L. S., McCaffrey, D. F., Stecher, B. M., Robin, A., & Burroughs, D. (2000a). *Teaching practices and student achievement: Report of first-year findings from the 'Mosaic' study of systemic initiatives in mathematics and science.* Santa Monica, CA: RAND.

Klein, S. P., Hamilton, L. S., McCaffrey, D. F., & Stecher, B. M. (2000b). *What do test scores in Texas tell us?* (Issue Paper). Santa Monica, CA: RAND.

Laguarda, K. G., Goldstein, D. S., Adelman, N. E., & Zucker, A. A. (1998). *Assessing the SSIs' impact on student achievement: An imperfect science.* Menlo Park, CA: SRI International.

Laguarda, K. G., Breckenbridge, J., & Hightower, A. (1994). *Assessment programs in the Statewide Systemic Initiatives (SSI) states: Using student achievement data to evaluate the SSI.* Washington, DC: Policy Studies Associates, Inc.

Linn, R. L., Baker, E. L., and Herman, J. L., & Koretz, D. (2002). *Standards for educational accountability systems* (Policy Brief 5). Los Angeles: CRESST, National Center for Research on Education, Standards, and Student Testing.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 29*(2), 4-16.

Lord, F. M. (1980). *Applications of Item Response Theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Loveless, T. (2002). *The 2002 Brown Center Report on American Education: How well are American students learning?* Washington, DC: The Brown Center on Education Policy.

Loveless, T. (2001). *Draft 2004 Mathematics framework for NAEP* (Governmental Studies). Washington, DC: National Assessment Governing Board, The Brookings Institution.

Loveless, T., & Diperna, P. (2000). *The Brown Center Report on American Education: How well are American students learning? Focus on Math Achievement 1*(1). Washington, DC: The Brookings Institution.

Maine Department of Education, Department of Educational and Cultural Services, Division of Educational Assessment. (1989). *Guide to the Maine Educational Assessment: 1988 – 1989*. Augusta, ME: Authors.

Maine Department of Education, Department of Educational and Cultural Services, Division of Educational Assessment. (1990). *Guide to the Maine Educational Assessment: 1989 – 1990*. Augusta, ME: Authors.

Maine Department of Education. (1997). *Learning results*. Augusta, ME: Author.

Maine Department of Education, Maine Educational Assessment. (no date). *Comparing the old Maine Educational Assessment (MEA) and the new MEA, Maine's Learning Results Performance Assessment*. Augusta, ME: Authors.

Maine Department of Education, Maine Educational Assessment. (no date). *Timeline of the MEA and Learning Results*. Augusta, ME: Authors

Maine Department of Education, Maine Educational Assessment. (Date?). Released test items: 1991–1999. Augusta, ME. Authors.

Massachusetts Department of Education. (1995a). *The Massachusetts Educational Assessment Program: 1994 statewide summary*. Malden, MA: Authors.

Massachusetts Department of Education. (1995b). *The Massachusetts Educational Assessment Program: Description of test content and reporting categories.* Malden, MA: Authors.

Massachusetts Department of Education. (1996a). *1996 MEAP statewide summary: Massachusetts Educational Assessment Program*. Malden, MA: Authors.

Massachusetts Department of Education. (1996b). *Mathematics Curriculum Framework: Achieving mathematical power*. Malden, MA: Authors.

Massachusetts Department of Education. (1996c). *Mathematics Curriculum Framework: Achieving mathematical power*. Malden, MA: Authors.

Massachusetts Department of Education. (1998a). *Massachusetts Comprehensive Assessment System: Release of Spring 1998 test items, mathematics, grade 4*. Malden, MA: Authors.

Massachusetts Department of Education. (1998b). *Massachusetts Comprehensive Assessment System: Release of Spring 1998 test items, mathematics, grade 8*. Malden, MA: Authors.

Massachusetts Department of Education. (1999a). *Massachusetts Comprehensive Assessment System: Release of Spring 1999 test items, mathematics, grade 4*. Malden, MA: Authors.

Massachusetts Department of Education. (1999b). *Massachusetts Comprehensive Assessment System: Release of Spring 1999 test items, mathematics, grade 8*. Malden, MA: Authors.

Massachusetts Department of Education. (1999c). *Massachusetts Comprehensive Assessment System: 1998 Technical report*. Malden, MA: Authors.

Massachusetts Department of Education. (2000). *Massachusetts Comprehensive Assessment System: 1999 Technical report*. Malden, MA: Authors.

Matthews, J. "Computational Skills Slipping, *"The Washington Post,"* September 3, 2002.

Mehrens, W. A. (2000). Defending a state graduation test: GI Forum et al. v. Texas Education Agency et al: Measurement perspectives from an external evaluator. *Applied Measurement in Education, 13*(4), 387–401.

Mid-continent Research for Education and Learning (McREL). (2003). Database of content standards (http://www.mcrel.org/standards/index.asp). Aurora, CO.

Miles, M. B., & Huberman, A. M. (1984). *Qualitative data analysis. A sourcebook of new methods*. Newbury Park, CA: SAGE Publications, Inc.

Mullis, I. V. S., Dossey, J. A., Owen, E. H., & Phillips, G. W. (1993). *NAEP 1992 mathematics report card for the nation and the states. The Nation's Report Card.* (Report No. 23-ST02.) Washington, DC: The National Center for Educational Statistics.

National Assessment Governing Board. (undated). *Mathematics framework for the 1996 and 2000 National Assessment of Education Progress.* NAEP Mathematics Consensus Project. Washington, DC: U.S. Department of Education.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics.* Reston, VA: Author.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics.* Reston, VA: Author.

National Science Foundation. (1997). *Statewide systemic initiatives.* Washington, DC: Author.

North Central Regional Educational Laboratory. (1996). *State student assessment programs database.* Oakbrook, IL: Author.

Perda, D. (2000). Personal communication.

Phillips, M. (2000). Understanding ethic differences in academic achievement: Empirical lessons from national data. In D. W. Grissmer & J. M. Ross (Eds.), *Analytic issues in the assessment of student achievement* (pp. 103-132). Washington, DC: U.S. Department of Education, RAND.

Rankin. S. C., & Hughes, C. S. (1987). *The Rankin-Hughes Framework: Developing thinking skills across the curriculum.* Westland, MI: Michigan Association for Computer Users in Learning, pp. 1-13.

Raudenbush, S. W. (2000). Synthesizing results from the NAEP Trial State Assessment. In D. W. Grissmer & M. Ross (Eds.), *Analytic issues in the assessment of student achievement* (pp. 3-42). Washington, DC: U.S. Department of Education, RAND.

Raudenbush, S. W., & Bryk, A. S., Cheong, Y. F., & Congdon, R. T. (2000). *HLM 5: Hierarchical linear and nonlinear modeling.* Chicago: Scientific Software International.

Raudenbush, S. W., Fotiu, R. P., & Cheong, Y. F. (1999). Synthesizing results from the Trial State Assessment. *Journal of Educational and Behavioral Statistics. 24*(4), 413-438.

Raudenbush, S. W., Fotiu, R. P., & Cheong, Y. F. (1998). Inequality of access to educational resources: A national report card for eighth-grade math. *Educational Evaluation and Policy Analysis, 20*(4), 253-267.

Roeber, E., Bond, L., & Braskamp, D. (1997). *State student assessment programs annual survey. Data volumes I and II.* Washington, DC: Council of Chief State School Officers.

Roussos, L., & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20,* 355-371.

Shaughnessy, C. A., Nelson, J. E., & Norris, N. A. (1997). *NAEP 1996 mathematics cross-state data compendium for the grade 4 and grade 8 assessment.* Washington, DC: National Center for Education Statistics.

Shields, P. M. Marsh, J., & Adelman, N. (1998). *Evaluation of the National Science Foundation's Statewide Systemic Initiatives (SSI) Program: The SSI's impacts on classroom practices.* Arlington, VA: National Science Foundation.

Spiegelhalter, D. J., Thomas, A., & Best, N. G. (2000). *WinBUGS Version 1.3 user manual.* Cambridge, UK: University of Cambridge. Medical Research Council Biostatics Unit.

Texas Education Agency, National Computer Systems, Harcourt Educational Measurement, Measurement Incorporated, and BETA, Inc. (1999). *Technical Digest for the Academic Year 1998-1999.* Austin, TE: Authors.

Texas Education Agency, Student Assessment Division. (1999). *TASS Coordinators' Manual.* Austin, TE: Authors.

Thissen, D. (1991). MULTILOG (version 6.0) [computer program]: *Multiple, categorical item analysis and test scoring using item response theory.* Chicago, IL: Scientific Software, Inc.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of Item Response Theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity*, (pp. 147-169). Hillsdale, NJ: Erlbaum.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning*, (pp.67-113). Hillsdale, NJ: Erlbaum.

Turnbull, B. J., Grissmer, D. W., & Ross, J. M. (2000). Improving research and data collection on student Achievement. In D. W. Grissmer & J. M. Ross (Eds.),

Analytic issues in the assessment of student achievement (pp. 299–315). Washington, DC: U. S. Department of Education; RAND.

U.S. Department of Education. (2001). *No Child Left Behind Act of 2001*. Reauthorization of the *Elementary and Secondary Education Act*. Washington, DC: Author.

U.S. Department of Education. (1997). *Individuals with Disabilities Education Act Amendments of 1997.* Washington, DC: Author.

U.S. Department of Education. (1994). *Improving America's Schools Act of 1993: The reauthorization of the Elementary and Secondary Education Act and other amendments.* Washington, DC: Author.

Webb, N. L., Clune, W. H., Bolt, D., Gamoran, A., Meyer, R. H., Osthoff, E., & Thorn, C. (2002). *Models for analysis of NSF's Systemic Initiatives Programs—The impact of the Urban Systemic Initiatives on student achievement in Texas, 1994–2000.* Madison: University of Wisconsin, Wisconsin Center for Education Research.

Webb, N. L., Kane, J., Kaufman, D., Yang, J.-H. (2001). *Study of the impact of the Statewide Systemic Initiatives Programs: Technical report to the National Science Foundation on the use of State NAEP data to assess the impact of the Statewide Systemic Initiatives.* Madison: University of Wisconsin, Wisconsin Center for Education Research.

Wenglinski, H. (2002). How schools matter: The link between teacher classroom practices and student achievement performance. *Educational Policy Analysis Archives, 10*(12).

Wenglinski, H. (2000). *How teaching matters: Bringing the classroom back into discussion of teacher quality.* Princeton, NJ: Educational Testing Service.

Zucker, A. A., Shields, P. M., Adelman, N. E., Corcoran, T. B., & Goertz, M. E. (1998). *Statewide Systemic Initiatives Programs: A report on the evaluation of the National Science Foundation's Statewide Systemic Initiatives Program.* Menlo Park, CA: SRI International.