10. CONCLUSIONS

The main goal of the project, *Systemic Initiatives: Student Achievement Analysis Study*, was to develop an analytic framework for studying the degree to which systemic reform contributes to improved student achievement and other outcomes. In pursuing this goal, we sought to provide information to the National Science Foundation that will enable NSF, its education constituencies, and the education research community to address the following questions:

1. How can the data submitted to NSF by systemic initiatives be used to evaluate systemic reform?
2. How does the precision of analysis depend upon the qualities of student assessment data?
3. What statistical models best fit the data linking systemic initiatives to student achievement?
4. What are the lessons learned about the kind of databases and analyses that are most effective for evaluating and understanding systemic reform?

To achieve the main goal of the project and to produce information on the specific research questions, we formed a multidisciplinary team that met at least monthly for more than a year to discuss the specific issues of analyzing student achievement data and of attributing gains in student achievement to state and urban systemic initiatives. The research team included:

Norman Webb—Curriculum, Assessment, and Evaluation of Systemic Reform
William Clune—Policy Analysis, Evaluation of Systemic Reform
Adam Gamoran—Sociology
Daniel Bolt—Educational Psychology
Robert Meyer—Economics
Christopher Thorn—Political Science and Technology
Eric Osthoff—Educational Policy Analysis, Evaluation of Systemic Reform
Jianbin Fu—Educational Psychology graduate student, database analyst
Kyo Yamashiro—Educational Policy Studies graduate student
Darwin Kaufman—Assessment and state education systems

In addition to these, Douglas Bates, Statistics, served as a consultant on the project on database management and analysis. Jung-Ho Yang, Educational Policy Studies, did additional analysis and consulted on the project. Bridget Gleeson, Economics, conducted some of the analysis. Margaret Powell was the editor and Lynn Lunde the project assistant.

Early in the project, we decided to focus on Texas, using data from the Texas Assessment of Academic Skills (TAAS). We did this for a number of reasons:

1. We had acquired TAAS system data, with individual student scores for grade 3 through grade 8 for 1994 through 2000.
2. Texas was an SSI state.
3. Texas also had three Urban Systemic Initiatives, providing data for most of the period of our study.

The TAAS data allowed us to analyze student achievement for multiple USI districts and over a period of time. The data set included individual identification of students by school, district, and demographics. The data set did not include any USI participation information by school or district. In preference to doing a superficial analysis of data from a number of states, we chose to look more deeply at data from one state. We felt we could best meet the goals of the project and produce the most useful information to NSF by addressing the rich database provided by Texas. It was clear that few states have data as rich or as broadly based as Texas. In analyzing Texas data, we were able to identify the specific features of databases that are needed to relate student achievement to systemic initiatives' activities. We were also able to study the effects of the Urban Systemic Initiatives (USIs), in addition to the systemic initiative operated at the state level.

In order to isolate the effects associated with the Texas SSI would require comparing Texas with other states on comparable variables. Such an analysis would require identifying a control group and common measures, such as the National Assessment of Education Progress. We have done this type of analysis in another study (Webb, Kane, Kaufman, & Yang, 2001).

Use of Data Sets Submitted to NSF

The SSIs submitted data to NSF in a variety of forms, including printouts of raw data, printed summary tables, and electronic data files. Data in print form are almost useless unless a considerable amount of effort is made to abstract the necessary information. Electronic data files with the appropriate documentation that identifies the data structure, variable labels, and contextual information (e.g., groups excluded from the data set) can be useful provided that the systemic initiative activities of specific groups are identified by participation level. Student achievement data that are not linked to other data, including demographic data and participation data in the systemic initiative's activity, will be virtually useless for attributing improved student performance to the systemic initiative.

Our analyses were confined primarily to available data, similar to data that systemic initiative sites make available to NSF. The biggest problem that confronted us was obtaining data on the independent variables, the degree of implementation of the systemic initiative. Although we tried, we were for the most part unsuccessful in getting from districts data that described the USI participation level of teachers by school. We also were able to compare student performance by USI districts without distinguishing in any way the quality of the USI implementation. Other desirable features of data and databases that are needed in studying the impact of the systemic initiatives include:

1. An identifiable control group
2. Testing students in consecutive years
3. Vertical scaling of scores over grades
4. An assessment that measures the full range of student knowledge without, ceiling effects
5. Assessment data linked with student, school, and district demographic and program data
6. Alignment of assessments with district standards and USI goals
7. A means of determining student attrition rates in the population and the selective exclusion of students from testing

Precision of Analysis

How precise the analysis of data can be will depend on a number of factors. Clearly, the quality of analysis will depend on the quality of the data. The level of analysis is very important in determining the precision of the analysis. The preponderance of information from an analysis is produced at the level with the greatest variation. In the case of systemic reform, this would be at the school level. If schools could be classified by their degree of participation in and implementation of a systemic initiative, then comparing performance by schools would produce the most precise and most powerful information. We attempted one such analysis, but we had very weak data on the classification of schools by their degree of participation in the urban systemic initiative. To adequately link growth in student performance to the systemic initiative requires data on student performance by student and by the teacher or teachers who taught each student. Using the TAAS data, we were successful in obtaining scores for individual students and the school and district each student attended when tested. We also were able to acquire a data file from Dallas that gave, for some years, the identifying number for a teacher, the duration and type of his/her professional development, and the teacher's school. This allowed us to do some computations at the school level, including producing a robust indicator of professional development at the school that related to the urban systemic initiative and student performance at the school. We were unable, however, to relate a teacher's participation in professional development to student performance. In the absence of being able to link teachers to students, the best we could do was to produce very general information on student performance for a school and a district. Most of our analyses were at the district level, the major focus of systemic reform. However, there was less variation among districts and a lower number, resulting in less precise results, than if there were 50 or more USI districts in the state.

Another factor in the precision of analyses is the proportion of the population that is tested and included in analysis. In 1999-2000, overall about 10% of the students enrolled were not tested. We found that nearly twice as many African Americans and Hispanic students as White students were not tested. Of those tested, a higher percentage of African Americans and Hispanic students than of White students were excluded from the accountability system. This indicates that there was some racial bias in the assessment results. We considered mobility as one reason why some students may not have been tested. Over the period of the study, we found that inclusion rates of those tested in one year and then tested in the next year increased. This would suggest that the test results in the latter years were less biased than in the earlier years. Test bias due to mobility or taking the test in two consecutive years was mainly related to low performance. The students who were the worst performers in one year were less likely to be tested the next year. Thus, test results may not accurately describe the performance of a population or the true effects of the program because of what seemed to be the systematic exclusion of students. Mobility is a significant problem in large-urban districts. When Gamoran tried to estimate growth in student achievement over time in Dallas elementary schools, he found that only 35% of the students had three or more data points. Thus, test bias from exclusion of students can affect the precision of analyses needed to produce information that is generalizable to the population.

Standard error of measurement of the assessment instruments is another important factor that contributes to the precision of analysis. Measurement error becomes more of a factor when scores are analyzed over time. As Meyer has shown, uncorrected measurement error in prior achievement can result in bias parameters in the analytic model. The bias due to measurement error will be more significant at the extremes of the distribution. In developing his models (Section 8), Meyer explored different methods for estimating the measurement error characteristics of the two scales employed, a maximum likelihood approach and a finite sample approach. His analysis indicated that the finite sample approach produced much better estimates of error variances, particularly at the extremes. Bolt (Section 5) applied the IRT latent ability metric rather than the TLI because of a ceiling effect. However, as Meyer reports, even though the continuous IRT metric is not subject to the ceiling effect, the large number of relatively easy items on the test forms resulted in much less precision in the ability estimates at the high end of the latent ability metric.

In summary, we have identified three general areas that can influence the precision of analyses of student achievement data in studying the impact of systemic reform: (1) the extent to which teachers, schools, and districts participated in the systemic initiative over time; (2) what students were excluded from the testing and analyses; and (3) the standard error of measurement in the assessment instruments. We have addressed each of these areas in the models we produced, but the existing data were insufficient for us to overcome all of the bias. However, our procedures for generating the models are sound and do demonstrate how the problem of analyzing systemic reform can be approached.

Fit of Analytic Statistical Models

There is no one best model for analyzing the link between systemic initiatives and student achievement. Each model is based on specific assumptions made necessary by the incompleteness of available data or other constraints. Even under ideal laboratory conditions, assumptions are made about what instrument to use, what an appropriate control group is, and what data to collect. The three models presented in this study are all viable approaches to relating a district's participation in an Urban Systemic Initiative with student achievement. All three models produced evidence that the USI district had at least a small positive effect on student achievement. Because of the lack of more specific information on USI activities, it was impossible for any of the models to produce definitive answers as to whether the effects are directly related to USI participation. To gather such information would require resources and time that exceeded what was available for this project.

As a set of three analyses, all of the models provide information about students' growth over time and compare the performance by students in USI districts with those in other districts. In this respect, the results from one model serve as a replication of those from the other models. However, all three researchers made different assumptions about growth that influenced which groups of students were tracked over time. Bolt examined changes in school means on TAAS at a given grade level (e.g., grade 5 in 1994, 1995, 1996, etc.). He assumed that this approach would more effectively control for teacher effects because the same teachers are more likely to teach the same grade in successive years. Within-grade analysis also is more comparable with the nature of TAAS and the TLI scores that are equated within grades. His model implies that the variation among different cohort of students is due to program effects rather than to other factors.

Similarly, Gamoran used nearly all of the students in the database to estimate the growth intercepts and slope. In this very robust model, students with any two scores, even those whose scores are not for consecutive years, can be used to estimate the parameters. To develop estimates for a model with a quadratic term requires students with four data points in order to estimate the three parameters intercept, slope (linear term), and changed (quadratic term). Students with fewer than four data points contribute to estimating the lower order terms. Thus, Gamoran's growth model included students who left or entered the system during the period investigated. He also included students who changed schools within the system, state, or district. Students who were retained in a grade and had two scores for the same grade were deleted from the database used in the analysis.

Meyer's value-added analysis examined students' performance in a grade by taking into consideration prior achievement from the year before. His analyses included students that had test scores for two consecutive years (e.g., grades 3 and 4, grades 4 and 5, etc.). This was a more restrictive requirement than that used by the other two researchers. The advantage of this approach is that greater precision could be given to improved student performance that can be attributed to a school year. The other two researchers computed the intercept term, or the initial performance at grade 3, that was used to compare the starting points among districts, but their models computed the difference between any two grades using equations developed to fit all of the points over the seven years, rather than just between two years.

All three models considered change in student scores over time. The type of analyses that were performed depended on a rich database, such as that produced by Texas. Most state databases do not include test scores by student for multiple years. At best, other states would only have school means over time. We did not conduct an analysis using school means to demonstrate the limits on analysis, if data were only available by school. Similar procedures, as demonstrated here, could be used to model change over time using school data. All three models and approaches to defining contrasting groups for comparison are based on specific assumptions and have advantages and disadvantages. We found that it was informative and productive to engage in developing alternative models concurrently. The interaction among the researchers influenced the development of each of the models.

Findings Produced by the Three Models

Based on the three analytic models, we drew the following conclusions:

1. Texas Assessment of Academic Skills (TAAS) scores improved from 1994 to 2000.

2. USI districts began with lower scores, but raised scores faster than non-USI districts.

   A. There is some evidence that USI districts improved faster than non-USI districts, but this may be due to the scoring metric used (e.g., TLI rather than Rasch). Meyer found positive USI effects over time for most grade levels. Bolt found a positive USI effect at grade 7.

   B. Texas USIs scores started below those of non-USI districts and produced smaller annual gains than non-USIs in 1994.

   C. There is no evidence that USI districts lost ground compared to non-USI districts from 1994-2000.

3. Achievement gaps between minority and White students and between advantaged and disadvantaged students narrowed statewide on the TLI. However, this finding was highly dependent on the scaling metric used. There was much less narrowing of gaps when the Rasch scale, which is more sensitive to gains at the extreme ends of achievement, was used rather than the TLI. There was some evidence from one model that the gap between White students and African American students narrowed more in USI districts than in the contrast districts.

4. There is no difference in the rate at which achievement gaps are narrowing in USI and non-USI districts.

5. Attrition rates of students not tested declined over time.

   A. TAAS attrition rates in students not being tested went down over time.
   B. There was no difference in the TAAS attrition rate between USI districts and large urban districts that are not USIs.

6. There are large differences in TAAS attrition by demographic group.

Lessons Learned

A number of lessons learned about understanding and evaluating systemic reform pervade this report. In interpreting assessment data and working with the complications in analyzing data for the purpose of drawing conclusions about systemic reform, we learned or verified:

1. Analysis of the academic growth of students over time is productive.
2. Careful attention needs to be given to the sample of content in the set of test forms given over time to assure scores are comparable.
3. How test scores are scaled and whether they are scaled only within grade or vertically across grades will have an impact on what kind of analysis can be performed.
4. Quadratic and higher-order models are needed to represent the complexity in data associated with analyzing the change in system data over time.
5. Attributing gain in scores by subpopulations to a systemic initiative has to be sensitive to variations in scores as a result of using equivalent forms over time.
6. Information on student and teacher mobility is essential if change in student assessment scores over time is to be attributed to any large-scale program.

We confirmed lessons that have been learned elsewhere about the design of large-scale studies. Some of these basic lessons include:

1. A control or contrast group is essential for considering systemic reform impact.
2. Data on teacher, school, and district participation in a systemic initiative should be maintained and continually updated from the very beginning of the initiative.
3. Specific information about the design and implementation of individual systemic initiatives is necessary in order to isolate possible causal links between reform activities and student achievement patterns.
4. In tracking achievement over time in large systems, it is necessary to attend to policy changes that affect test scores, such as requiring all Special Education students to be included in the assessment system.
5. Any analysis of assessment data over time needs to be accompanied by an analysis of the change in those tested over time.

Finally, we learned about the process for building models and the value of working with an interdisciplinary group. Through developing three models concurrently, each model approached was improved. Each model had benefits and tradeoffs. However, the efficacy of the models could not compensate for the relative lack of data on the implementation of the USIs.