

## 1. INTRODUCTION

In this project, *Systemic Initiatives: Student Achievement Analysis Study*, we have developed analytic frameworks for studying the degree to which systemic reform contributes to improved student achievement as well as other education outcomes. Using grades 3 through 8 data from the 1994 to 2000 Texas Assessment of Academic Skills (TAAS), we have developed prototype analyses to illustrate approaches that can be used to study the impact of a systemic initiative on student learning. Although the three models for analyzing systemwide data have produced some specific findings, the more instructive results from this project are, first, the identification of the pitfalls and problems in doing such analyses and, second, the recommendations for designing and structuring assessments and program data at the outset to better achieve the desired analysis. Faced with using extant data, we have had to make a number of choices and assumptions in developing our models and performing the analysis. Most critically, data on the implementation of the systemic initiatives and the degree of saturation of the initiative in schools and districts were either not available or were very sparse.

Four research questions guided the project's work:

1. What are the most effective ways to use the datasets submitted to the National Science Foundation (NSF) by the Systemic Initiatives (SIs) for evaluating the success of systemic reform and understanding its dynamics, both within and across the SIs?
2. How is the analytical precision required to answer research question #1 above dependent upon specific qualities of the student assessments and data, including reliability, validity, and comparability of the tests?
3. What analytic statistical models best fit the data linking the systemic initiatives to student achievement and how can these be demonstrated in a prototype analyses?
4. What lessons can be learned and communicated to the field from this project about the kind of database design and analysis that is more or less useful for evaluating and understanding systemic reform?

Of the four research questions, we focused most of our attention on the third. In developing the analytic statistical models and solving the many problems that arose in the process, we gained the insight and information needed to answer the three other questions.

Shortly after the beginning of the project, May 1, 2000, we made a decision to focus on data from only one state, Texas. There were a number of reasons for doing this. TAAS had been in place since 1994, which meant that in Texas we had student-level data in mathematics for each year and each grade from grade 3 through grade 8. This afforded us the opportunity to follow the same cohort of students for up to seven years of schooling, from grade 3 in 1994 through grade 8 in 2000, and other cohorts of students for five or six years. Texas was in the second cohort of states to receive NSF funding for a statewide systemic initiative (SSI) in 1992. By 1999, four of the state's largest urban areas had also received urban systemic initiative (USI)

grants—Dallas (1993), El Paso (1993), San Antonio (1995), and Houston (1999). In 1999, Dallas and El Paso received continued funding through NSF's Urban Systemic Program (USP), the follow-up to the USIs. Texas was the only SSI state with multiple sites funded by the USI, other school districts in urban areas without USI funding, and a very large assessment database. Within Texas, we focused on Dallas. The USI in Dallas, along with that in El Paso, were among the first USIs funded. We contacted the leaders of the El Paso USI, but they were overloaded with researchers investigating their USI and did not have the time or resources to work with us. Since Houston did not become a USI until 1999, the year before our last year of TAAS data, we selected districts in the Houston area as the non-USI districts in our study. The choice of the elementary and middle grades as our research focus permitted us to track the efforts of the SSI and USI, since these grades had student test scores available.

Although Texas had joined the second cohort of states funded through the SSI program in 1992, the state initiative had a very slow start. Two or three years later, the leadership requested and received permission from NSF to restart the SSI. By the 1996-97 school year, much of the Texas SSI's work was devoted to enhancing the existing state infrastructure in mathematics and science education. In the early stages, the work of the Texas SSI centered around developing state mathematics and science frameworks and the Texas Essential Knowledge and Skills (TEKS) standards that would eventually impact all schools in the state. Leadership teams, or Action Teams, formed with the help of the SSI, worked on writing these documents and designing ways of implementing them. Four other projects had a direct influence on individual districts and schools around the state. These included: 1) a Title I project; 2) TEXTTEAM's professional development of school staff; 3) implementation of the reform mathematics curriculum, *Connected Mathematics Project*, in selected schools; and, 4) the availability of Advanced Placement courses in selected high schools. Most of these projects centered on mathematics, which was included in the state's accountability system, and given priority by schools. The SSI also attended to science education in the state, but science did not have the same leverage as mathematics in the schools since it was not tested by TAAS.

We received TAAS data from the Texas Education Agency (TEA) in August, 2000. The TAAS is a criterion-referenced test in reading and mathematics that is administered in the spring to students in public schools in grades 3 through 8 and in grade 10. In the spring of 1994, the passing standards in reading and mathematics at grades 3 through 8 were aligned with the exit-level standard in order to measure student achievement across time. Students in grades 3 through 8 and in grade 10 who achieved a Texas Learning Index (TLI) score of 70 or higher met minimum expectations in reading and mathematics. In this study of SIs in mathematics and science, we only focused on the mathematics scores for grades 3 through 8 from 1994 through the spring of 2000. The Texas Education Agency was very cooperative in providing us with the large dataset and responding to our questions. However, for security reasons, we were unable to obtain the TAAS Item Response Theory (IRT) parameter values from the TEA. The lack of these parameters impeded a portion of the analysis, but we were able to perform some analyses by estimating parameters and using what we have called pseudo-parameters.

In order to study the impact of the USIs, we had to identify 1) districts and schools that had participated in USI activities, or would have been influenced by the USI, and 2) an appropriate comparison group of schools and districts. Data on the schools that had participated

in the Dallas USI were not readily available. In order to identify the effects of the systemic initiative, it was important to be able to distinguish between schools and districts that participated in a USI from those that did not, or to determine the degree to which schools and districts participated in a USI. We had limited resources for collecting additional data for study of the independent variable. A district was identified as a USI district if the district had received NSF funding and was identified as a USI in the proposal. Knowledgeable informants in Dallas and data on participation in professional development activities were used to help us determine the degree to which Dallas schools had participated in USI activities.

Two of the models developed in this project use different and contrasting groups for comparing the USI districts and schools with non-USI districts and schools. In one model, performance of students in USI districts is compared to the performance of students in all other districts in the state of Texas. This is a stringent criterion for comparison, but is reasonable if all districts are expected to perform equally well. In another model, student performance in USI districts is compared to the performance of students in other large urban or urban-fringe districts. For this comparison, we made the assumption that the USI districts should be compared only to other districts that would be eligible to receive USI funding. Both the El Paso and the San Antonio USIs included school districts from the surrounding area. Thus, we made certain assumptions about a school's participation in a USI and an appropriate comparison group. Each of these assumptions needs to be considered when interpreting the findings and the differences in the models developed here.

Three different approaches were used to analyze the data, each based on different assumptions, each using different techniques, but all achieving similar findings. Daniel Bolt used a hierarchical linear model to examine trends in the TLI within schools and districts at a fixed grade level from 1994-2000. This analysis permitted the estimation of average growth curves at the school, district, and district type (USI or non-USI) level, as well as the quantification of the variance of growth curves within USI and non-USI districts. In addition to this trend analysis, he used Item Response Theory (IRT) to trace growth at a more refined level by comparing changes in the three-item domains being used (concepts, operations, and problem solving). Strong similarities in growth patterns across the different domains can indicate that improvement was not centered on any one area but was spread among a number of areas. A unique feature of this approach is that the model fit can be directly assessed. Adam Gamoran employed a multilevel growth model that 1) monitored trends for changes in the grade levels from one year to the next and 2) tracked changes in the cohorts of students as they moved through the school system over time. He included in his model student characteristics such as gender, race/ethnicity, and free/reduced-cost lunch status. He contrasted USI districts with eligible USI districts and, within Dallas, students in schools that implemented the USI rigorously with schools that implemented the USI less faithfully. Robert Meyer developed a value-added model by comparing the change in scores by the same group of students from one grade to the next for both USI students and non-USI students. He incorporated into his model other factors that included race/ethnicity, free/reduced-cost lunch status, and prior achievement. Meyer's model controls explicitly for measurement error in prior achievement, using a new method that is substantially more accurate than conventional methods. One benefit of his approach is that it produces estimates of district, state, and USI program effects by grade and year.

This report continues with discussion of the critical issues encountered as we studied the impact of a large-scale USI program on student achievement. In this section, we report on the substantive questions of interest faced by those designing an evaluation study that analyzes the relationship between student achievement and an SI program. This is followed by a review of how others have approached the challenge of attributing change in student achievement to program effects. Then, the methodology for our study is described along with how we resolved the problems we faced, including how to clean the data, collect information on school participation, and define the control groups. An important question encountered in conducting longitudinal analysis is, How does the population being studied change over time? Christopher Thorn studied the attrition rate of students tested from one year to the next. This is reported in the next section and leads into a section on each of the three models. A section that compares and contrasts the models follows the discussion of the three models. This is followed by conclusions and recommendations.