3. HOW OTHERS HAVE ADDRESSED THE PROBLEM

Tracking changes in student achievement in large school districts and in states and attributing these changes to specific activities or initiatives such as NSF's Systemic Initiative (SF) Program is fraught with problems. There are a few studies that have made the attempt, but none have fully accomplished a definitive study identifying the effects on statewide or district student achievement that can be attributed to a single cause or to a set of causes. There are a number of reasons for this, but the most basic is that it is very difficult, if not impossible, to conduct such a study without expending major financial resources and securing cooperation from a large number of people within the systems being studied. At the same time, there is more testing of K-12 students in the United States now than there has ever been. Because of the large amount of testing that is taking place, there are vast quantities of data on students available to the nation, the states, and districts. All of this testing provides a massive amount of information on the overall performance of students, how students in our nation compare to the performance of students in other countries, how the performance of students in the nation and in the states has changed over time, how the performance of different groups of students compare, and the general factors that relate to higher achievement. We know less about factors that link large-scale initiatives to improved performance. Important reasons for this lack of attribution are: 1) the fact that it is not feasible to randomly assign students, schools, districts, and states to these initiatives; 2) it is difficult to define adequate control groups after progressing years into the program; and, 3) there are no baseline data.

A number of researchers have analyzed data from the National Assessment of Education Progress (NAEP). Webb, Kane, Kaufman, and Yang (2001) analyzed state NAEP data for 1990, 1992, and 1996 to study the impact of the National Science Foundation's Statewide Systemic Initiatives (SSI) Program. In 1990, NAEP reported state level data for the first time. The data structure allows some comparison among states at grades 8 in mathematics beginning in 1990 and for both grades 4 and 8 in mathematics beginning in 1992. Because the SSI program began in 1992, there are at least some state data prior to or concurrent with the initiation of each state's SSI. However, states' participation in the State NAEP is voluntary and not every SSI state or non-SSI state participated in this national assessment program. Thus, any analysis of State NAEP data is limited to those states that had students tested.

For the first three testing years—1990, 1992, and 1996—a total of 17 SSI states of a possible 26 (including Puerto Rico) and a total of 11 non-SSI states participated in NAEP all three years. In comparing these states, some conclusions could be reached. The SSI states were lower-performing states when the SSI program began and improved at a slightly faster rate than the non-SSI states, but not to the extent that the SSI states outperformed the non-SSI states by 1996. A statistical model using an empirical Bayesian method was used to compute how much variance could be explained by SSI participation. At both grades in 1992, SSI participation accounted for more of the variance, 8-10%. In 1996, SSI participation accounted for less of the variance, 4-5%, at both grade 4 and grade 8. The decrease in the variance accounted for by the SSI indicates that the performance of SSI states was more similar to that of non-SSI states. This is meaningful because the SSI states as a group had a higher proportion of minority students.

Both SSI and non-SSI states maintained the achievement gap between Black and White students over the three testing years. There was some evidence that the achievement gap between Black and White students in SSI states declined on specific subtopic scales—geometry and algebra/functions—whereas the gap increased in non-SSI states on all five subtopic scales for both grades 4 and 8. This study has some benefits. The study incorporates a single instrument for comparing multiple states, it tracks achievement over time, and it uses baseline data. It has some specific limitations: not all SSI states participated in NAEP data; the control group is restricted to only non-SSI states that participated in NAEP; the activities of the SSIs in most states had reached only a small proportion of teachers in the state; and, even though the NAEP assessment instrument measured a wide spectrum of content, it is possible that the instrument did not include a concentration on items related to the content emphasis of each SSI.

Klein, Hamilton, McCaffrey, Stecher, Robyn, and Burroughs (2000) in the Mosaic study investigated possible outcomes of the SSIs and the Urban Systemic Initiatives (USIs) using a research design that required some inference to relate their findings to the Systemic Initiatives. They examined the relationship between teaching practices and student achievement in mathematics and science at six sites, including up to four SSI states and up to seven urban centers. The researchers assumed that if "reform teaching," defined by attributes identified in the literature, could be shown to be related to higher student achievement, it could be inferred that participation in a reform initiative, such as the Systemic Initiatives, would enhance teaching reform and consequently lead to higher student achievement.

The researchers purposefully chose to model the relationship between teaching practices and student achievement rather than to compare the performance of students on the basis of the level of their teachers' participation in reform because of the number of issues related to doing the latter. Those from the sites studied were asked to identify schools that implemented reform practices and schools that implemented traditional practices. Ten schools from each group were selected to participate, primarily to assure a group of teachers with a range of practices. In the analysis of the study, teacher reports of their classroom practices on a questionnaire were used to classify their practices.

Whereas the Webb et al. (2001) study compared student achievement from states in which a proportion of teachers had additional professional training and schools had participated in an SSI, Klein, Hamilton, McCaffrey, Stecher, Robin, and Burroughs (2000) chose to compare student performance by the type of teaching the students had received within the context of educational systems that had participated in an SI. There was no assurance that the teachers using reform practices had a specific link to the Systemic Initiative. This study focused on instruction at the classroom level rather than at the school level or higher. In the first year of the study, the only year for which results were reported, the results indicated a small, but statistically significant, positive relationship between teachers' use of reform practices and student achievement in both science and mathematics and on both multiple-choice and open-response items. Teachers' use of traditional practices, measured with a less reliable instrument than that used to measure reform practices, was found in general to be unrelated to student achievement. These differences were much smaller than those found for covariates, including ethnicity and socioeconomic status (SES).

Klein et al. (2000) used a linear regression model in the analysis, controlling for student background characteristics and previous test scores. These researchers recognized a number of limitations to their study, including no random selection of students or teachers; no information on what influenced teachers in how they taught, such as professional development associated with the Systemic Initiative; questionnaires were used to classify the type of practices and not observations; and the study was confined to one school year, without allowing any opportunity to follow the development of teachers' practices over time and the relationship to student achievement.

A limited amount of data is available that was produced by the SSIs either from their state assessments or by focus studies that were specifically designed to evaluate the impact of the SSI. Laguarda, Goldstein, Adelman, and Zucker (1998) reported on the seven SSIs that were most likely to generate the most credible evidence that student achievement had improved. Four of the states used data from their state assessments in mathematics. Two states and Puerto Rico used mathematics assessments administered specifically to study the impact of the SSI. Five of the SSIs also reported student outcomes in science. For five of the seven SSIs, the treatment groups were students taught by teachers who had received professional development from the SSI. For the two other states, SSI participation was defined by those schools that identified themselves with the state program. In those schools, a lead teacher may have been trained to work with other teachers, some groups of teachers may have attended a professional development activity, or the principal may have simply pledged to participate. Three of the SSIs selected comparison groups that took race and level of poverty into consideration. Another SSI controlled for these factors using a regression model. Only two of the states attempted to control for prior academic achievement. The sample size in these schools varied considerably. Two of the states used a small sample of 12 or fewer classrooms. One used 12,000 students in grades 4 and 8 in 216 public and private schools. A state that had only baseline data did not report any findings. The other six SSIs reported, for at least one grade level, some positive results—instances where the SSI group performed better than the comparison groups by one or more percentage points. All or some of these studies had a number of limitations, including small sample sizes, unreported information about variances among schools and students, no random assignment between treatment and control groups, a short duration for the study (generally one year), no longitudinal analysis, and a lack of control for confounding variables.

A follow-up study to the Ohio SSI considered the continuation of the impact of the initiative after the end of NSF funding in 1996. Ohio had been very successful in obtaining funding for systemic initiatives, including funding for the state in 1991, three cities as part of the Urban Systemic Initiatives in 1993, and five counties as part of the Appalachian Rural Systemic Initiative in 1994. Providing professional development to mathematics and science teachers was the main strategy of the SSI. The state continued the work of the SSI after 1996 through a state funded project called *Discovery*. The study of the impact of the initiative used a number of different studies (Kahle, 1999). One compared the performance of students of teachers who had voluntarily received professional development with the performance of students of those who volunteered but had not participated. A total of 2,374 students participated in this study, with those taught by the SSI-trained teachers outperforming the others by 2 to 7 percentage points (Supovitz, 1996). In a later study, Ohio Proficiency Test data for 1997 and 1998 were used to compare the performance of middle/junior high school students in schools with a critical mass,

51% or more, of science and mathematics teachers that had participated in the SSI (*Discovery*) training. Schools were further identified by those with policies aligned with reform practices. In those schools with aligned reform policies, enrolling a total of over 6,000 students, schools with a critical mass of *Discovery* teachers (51% or higher) made significant gains in the percentage of students passing the Ohio Proficiency Test in both mathematics and science, compared to schools that had 25% or fewer *Discovery* teachers. The percentages were nearly the same in schools where the policies varied. The schools with a high percentage of *Discovery* teachers had a higher percent of minority students and those eligible for free or reduced-price lunch than the schools with the low percentage of *Discovery* teachers (Kahle, 1999). This study, primarily produced for policy makers, used the proportion of trained teachers to represent the variation in the treatment, along with aligned policy. Some limitations of the study are: 1) it is difficult to distinguish between the effects associated with aligned policies in schools and effects associated with having a critical mass of SSI teachers, and 2) the scores compared in the two years were produced by different groups of students.

Klein, Hamilton, McCaffrey, and Stecher (2000) questioned the validity of the gains on TAAS scores by Texas students over the period 1994 to 1998, raising the issue of what meaning can be given to high-stake state tests or whether such tests can be used these as indicators of improved system quality. These researchers compared the differences in scores between White students and students of color on TAAS with the State NAEP results for 1992 and 1996 with standardized achievement tests administered in 20 schools as part of the Mosaic study described above (Klein, Hamilton, McCaffrey, Stecher, Robyn, and Burroughs, 2000). They computed the effect size for the difference in achievement between white and black students and used this metric to compare the performance of Texas students on the different instruments. In one analysis, the study compared the performance of grade 8 students in mathematics in one year with grade 8 students four years later and the performance by grade 4 students in mathematics and reading in one year with grade 4 students four years later. That is, the performance of different cohorts of students was compared when the students were in the same grade.

In a second analysis, using only the NAEP data, the change in mathematics performance of a single cohort of students (students tested at grade 4 in 1992 and four years later at grade 8 in 1996) was compared for Texas and the nation. In the cohort analyses, students from the same group were tested four years later, but they were not exactly the same students. A similar analysis was done in reading, comparing the change in performance of grade 4 students in 1994 with grade 8 students in 1998. These analyses indicated that in mathematics and at grade 4 on NAEP, the gap between White and Black students, using the effect size metric, increased from 1992 to 1996. The gap between White and Black students at grade 8 on NAEP also increased over the same time period. However, using the TAAS and comparing 1994 with 1998, the gap between White and Black students at grade 4 decreased substantially, as did the gap at grade 8. The researchers hypothesized that the discrepancy between NAEP and TAAS results could be due to the preparation of students that narrowly focused on improving their performance on the TAAS, but did not generalize to the nationally administered NAEP. Another explanation given for the discrepancy is that the TAAS is too easy for some students, that it does not reflect accurately how much better some students are than other students. The results from the 20 schools where students were given tests in mathematics and science that included open-response questions indicated that the correlation varied substantially between the TAAS and non-TAAS in

mathematics, depending on whether students were used as the unit of analysis or the school was used as the unit of analysis. Regardless of the unit analysis, the TAAS tests had a very low correlation with SES. The researchers indicated that the study was not done under ideal conditions and had some limitations. The scores for the cohort group, and not the exact same students, were not compared longitudinally because the NAEP was not designed to do this. The effect size for the differences between the four years used to compare TAAS and NAEP in mathematics did not cover the same periods of time; TAAS data were used from 1994 to 1998, and NAEP data from 1992 to 1996. It is also possible that other factors, such as student demographics, may change that could make a difference when results are not compared for the same time period. This was not a problem for reading because the same time periods were used for both TAAS and NAEP. A strength of this study includes trying to draw meaning about student performance in a state using more than one measure of performance.

A number of recent studies have examined the relationship of a variety of different factors to performance by states on the basis of NAEP data (Grissmer et al., 2000; Raundenbush et al., 1998; Raudenbush et al., 1999; and Wenglinski, 2000). All of these studies found a strong relationship between home and family characteristics and student mathematics achievement. Beyond these relationships, a number of other variables showed some correlation with student achievement, including pupil-teacher ratio, teacher background variables, student enrollment in algebra, and instructional practices. Wenglinski (2000) found, after controlling for student, family, and home characteristics that teaching higher-order thinking skills were positively related to mathematics achievement as measured by NAEP. Grissmer et al. (2000) used up to seven NAEP scores to do a regression analysis of the relationship of the normalized test score and other variables. Both random-effect and fixed-effect models were used. One conclusion is that Texas students performed well compared to most other states, particularly on the 1996 assessments—seemingly contradicting the finding of Klein, Hamilton, McCaffrey, and Stecher (2000) reported above. The variables in the model did not account for all of the variation in performance. The researchers indicated a state's participation in a Systemic Initiative was a likely factor for explaining the variance in performance among the states. Grissmer et al. (2000) increased the number of performance data points by using up to seven achievement scores for each state and then modeled the average state scores as a function of other variables. This approach homogenizes achievement by assuming that change in reading performance at any grade is given equal weighting as change in mathematics performance at any grade. Also, variance among states can be computed, but no attention was given to variance among schools or students within states.

In summary, a number of studies have analyzed state data with implications for describing the effects of statewide programs. A common approach researchers have used is to develop some form of regression model to examine the relationship among input variables and student performance. State NAEP data have been used in many of these studies as the dependent measure because this database has certain very desirable traits, such as the fact that the same instrument has been administered in the same way in a number of the states. Many of the studies based on the State NAEP data have used only the state mean scores and have not used the complete database for computing variances among students or considering differences in performance among different topics within a content area. Many of the studies have identified correlates to achievement. Some of these have then made inferences about statewide programs.

For example, Klein, Hamilton, McCaffrey, Stecher, Robyn, and Burroughs (2000) studied the relationship between reform instructional practices and student achievement, making the inference that since the SSI program emphasizes reform instruction, if there is a relationship between these practices and student achievement, then there is a likelihood that the SSI program will have an impact. Some within-state studies have contrasted teachers who have participated in SSI activities in the state with those that have not as another possibility for inferring that an SSI is likely to have an impact on student learning. These models, however, ignore studying one of the primary tenets of the SSI program—i.e., to effect statewide change. Some of the researchers have looked at student performance over multiple years, but have not studied the change in performance over time in the same group of students. Rarely have studies used multiple measures of student performance or have used different analytic approaches based on different assumptions that can be used to replicate findings.

References

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.

Kahle, J. B. (1999). *Discovering from* Discovery: *The evaluation of Ohio's systemic initiative.* In N. L. Webb (Ed.), *Evaluation of systemic reform in mathematics and science. Synthesis and proceedings of the Fourth Annual NISE Forum* (Workshop Report No. 8). Madison: University of Wisconsin, National Institute for Science Education.

Klein, S. P., Hamilton, L. S., McCaffrey, D. F., Stecher, B. M., Robin, A., & Burroughs, D. (2000). *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, M. B. (2000). *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.

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.

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.

Supovits, J. (1996, December). *The impact over time of* Project Discovery *on teachers' attitudes, preparation, and teaching practice. Final report.* Chapel Hill, NC: Horizon Research, Inc.

Webb, N. L., Kane, J., Kaufman, D., & Yang, J.-H. (2001). *Study of the impact of the Statewide Systemic Initiatives Program.* Madison: Wisconsin Center for Education Research.

Wenglinski, H. (2000). *How teaching matters: Bringing the classroom back into discussions of teacher quality.* Princeton, NJ: Educational Testing Service.