2. ANALYZING THE IMPACT OF THE STATEWIDE SYSTEMIC INITIATIVES

Introduction

A number of complications arise when applying measures of student achievement to a study of the impact and progress of urban and state systemic initiatives. An important reason for engaging in this project was to identify what some of these complications are and the approaches that can be used in studying the impact of an initiative that takes these complications into consideration. The main goal of our study, Systemic Initiatives: Student Achievement Analysis, was to develop frameworks for analyzing the degree to which systemic reform has contributed to improved student achievement and other outcomes. The study has produced prototypes and models that illustrate how inferences can be made on the basis of student data about program impact. This project is a response in part to the difficulty the National Science Foundation (NSF) has had in obtaining data and reports from SSI states and commonwealths that could be used to evaluate how its ambitious initiatives have influenced student achievement.

Although there had been prior efforts by NSF to obtain data from SSI states, a special attempt to acquire these data from the states was made in 1996. On July 8, Dr. Luther Williams, then-Assistant Director of Education and Human Resources, NSF, sent a letter to all SSI jurisdictions requesting student achievement data. Each SSI state and commonwealth had been requested to submit student achievement data in mathematics and science from a period prior to the project's implementation, along with relevant current data, in graph or table form, for three groups of students, referred to as cohorts. Cohort I students were those whose schools had participated in SSI activities for at least three years. Cohort II students were those whose schools had participated in SSI activities for two years or less, and Cohort III were those whose schools had yet to participate. This effort was an attempt by NSF to acquire comparative data that could be used to evaluate the impact of the SSI program on student achievement. This request acknowledged two important classes of datasets that would be used to study the impact of the initiatives—data from prior to the implementation of the initiative and data from groups with different levels of participation in the initiative. However, the letter did not request the detailed data needed to fully measure whether there had been improvement in student performance and whether that improvement could be attributed to the systemic initiative. In order to do this, more information would have to be acquired about the initiative, other concurrent reforms, the psychometric qualities of the test, the demographic composition of the groups, and several other factors.

Twelve states responded to Dr. Williams' letter, nine within one month and three others within three months—one state produced Cohort I data, nine produced Cohort II data, and only two states produced Cohort III data. The turn-around time for the request forced all of the states, both SSI states and non-SSI states, to use existing data and reports available at the time, which had been prepared for other purposes. States submitted data in a wide range of formats, from listing assessment results and number of years served by school to scaled scores for participating and non-participating schools to a narrative report showing the differentials between participating and non-participating schools. The wide variation in how the data were reported

and the nature of the data that were reported made it almost impossible to judge the impact of the SSIs.

In subsequent years, the National Science Foundation continued to request achievement data from the SSI states and the Urban Systemic Initiatives (USIs). By 1999, NSF had received a range of state assessment data from a number of the 25 states and Puerto Rico that had received funding through the Statewide Systemic Initiatives (SSI) Program and from the USIs (Urban Systemic Initiatives). These data generally existed in a written form broken down into tables specific to the needs of each state at the time, rather than to NSF's objective of deriving meaning from the data about the SSI program. While some of the data were in tables, other data were presented in graphs. The significant variance in what was submitted by states signified how states differed in the achievement data available, number of years of testing, grade level tested, statistics reported, type of tests (criterion-referenced, norm-referenced, portfolios), SSI participation, inclusion or exclusion of specific groups (e.g., special education, private schools); context for testing, and conditions for assessment. NSF's attempts to acquire data were laudatory, but produced a collection of data that was almost impossible to use in drawing any conclusions about an individual state, much less for studying the aggregate. The difficulty in acquiring data and analyzing data does not remove the need for having measures of impact and for studying what can be attributed to highly visible, funded programs such as the systemic initiatives. NSF has been under tremendous pressure from Congress and others to produce evidence of the effectiveness of the SI program, which had been launched in 1991, was expanded in two subsequent award cycles by the mid-90s, and had expended hundreds of millions of dollars.

This report presents the findings of a research team that confronted the problem of how to judge the impact of a systemic initiative within a state using existing state assessment data. Although our analyses targeted the systemic initiative program, we addressed the more global questions on conducting analyses of state assessment data and on what state assessment data would provide an adequate base for evaluating the impact of a statewide or urban systemic initiative.

Difficulty in Interpreting Assessment Data

States and districts report statewide and district test scores annually. The general perception is that if the test scores for the current year are higher than those in the previous year, then education has improved. If there is a positive slope in the multi-year trend of average scores or the percentage of students who have obtained proficiency, then this is considered good and a public success. The more astute observer will be interested in the disaggregation of the data by race/ethnicity, gender, and poverty, but will be satisfied with evidence that these different groups show improvement. What is more difficult is to explain, or attribute the causes of, a positive trend to specific factors in general and to relate a rising trend to a specific program such as the systemic initiative. Repeated administration of equivalent forms of the same tests generally result in improved scores with each administration. In addition to educational factors behind these effects, such as improved instruction that impacts student performance, there are many other factors that influence the rise or fall in test scores. There are technical factors related to how the assessment data are analyzed and reported; there are content factors regarding what student

knowledge is tested; and there are factors regarding the population being tested that can be influenced by social and administrative practices.

Technical Factors

Most national, state, and district assessment data are reported for specific grade levels. These trend analyses portray the performance of students in a grade, say grade 4, for each year. However, the students in grade 4 change each year, so that the multi-year trend represents how a different group of students performed each year: The common factor is the grade level at which students were tested, but not the students per se nor even the teachers or the curriculum. Just having a different group of students tested the next year at grade 4 knowing more could result in an increase or variation in test scores at grade 4. Improved grade 4 scores could be considered an artifact of education in the primary grades (K-3), rather than of curriculum and instruction in grade 4. An alternative way to report longitudinal assessment data is to report the scores of the same students as they progress through the grades. Such cohort analysis would describe the progress of students from one grade to the next and could be used to measure the value added for the work in each grade. Such an analysis would be more sensitive to program changes or changes in instruction in a specific grade or grade range. However, cohort analysis would require testing at multiple grades, increasing the cost of the assessment program and raising other technical issues.

Equating of test forms is a psychometric technique used to draw inferences about student test scores on one form with the student test scores on another form. Equating test forms requires using statistical adjustments so that the scores on two or more alternate forms are placed on a common scale (Feuer, Holland, Green, Bertenthal, & Hemphill, 1999, p. 105). A number of methods can be used for equating or linking test forms. Statistically, equating test forms does not imply that exactly the same content is tested on both forms. Test forms can be statistically equated and produce student scores that can be placed on the same scale. But without careful attention to the sample of content tested on each form, the scores from each form may not be comparable in measuring students' content knowledge. Most states use equated test forms over a series of years to measure student achievement at a grade level, for example grade 4 in 1996, 1997, and 1998. Some states will administer tests to students in every grade and use vertically equated test forms so that students' scores can be compared from one grade level to the next. Most commercial testing companies vertically equate tests so that students who take tests in a sequence of grades can be placed on the same scale. However, commercial testing companies do not have a large number of multiple forms for any one grade. This means that if a state or district uses vertically equated forms, then after two or three years the same forms would have to be used again. Since 1994, Texas has led most other states in administering new test forms each year in grades 3 through 8. The Texas Assessment of Academic Skills (TAAS) test forms are equated for a grade level, that is, horizontally equated, but not from one grade to the next, or vertically equated.

Scaling of test scores is used to enhance the interpretation of test scores by placing scores from different tests or test forms onto a common scale or producing scale scores designed to support criterion-referenced or norm-referenced score interpretations. As with the Fahrenheit and Celsius scales for measuring and reporting temperature, a variety of scales are used to report

student achievement scores. It is essential that we know the characteristics of the scale being used, particularly for the extreme scores, when interpreting scale scores. For example, the National Assessment of Educational Progress (NAEP) uses a scale from 0 to 500. Item-response theory (IRT) scale-score models are used to vertically scale all of the items at grade 4, 8, and 12 on the same 0 to 500 scale. Scaling for the State NAEP assessment in mathematics was done separately for each of the five mathematics content strands for grade 4 and grade 8 (Johnson & Allen, 1997). Another example is the Massachusetts assessment. Starting in 1988, the Massachusetts assessment used a scale ranging from 1000 to 1600 and with an initial mean of 1300 and a standard deviation of 100. Grade 4 items and grade 8 items were not vertically scaled. Yearly comparisons could be made for a grade level, but not between the two grades. In 1998, Massachusetts' new assessment used a scale from 200 to 280. A third example is Maine's state assessment, which is scored on a scale from 0 to 400, with the mean set at 250 and a standard deviation of 50.

Texas employed a one-parameter IRT-equating procedure to develop the Texas Learning Index (TLI), used since 1994 to report TAAS scores. The TLI ranged from 0 to 100, with a score of 70 in each grade in 1994 taken to represent the criterion for students meeting the standard. One problem that arises in using the TLI, formed using IRT equating procedures, is that there is very little discrimination between students who score at the higher end of the scale. Even though students' raw scores may differ by one or two items, their TLI score is the same, thus forming a very skewed distribution at the upper end—meaning that the high-scoring students reached the "ceiling" of the test. This may not be a problem if the education policy in the state is primarily oriented toward improving the achievement of students who are performing at the lower end of the scale. Another scale can be created using a Rasch model, a one-parameter IRT model where the probability of a correct response to an item is a function of latent examinee ability, denoted θ, and the difficulty of the item, denoted b. The θ (theta) scale differentiates among ability levels at the extremities, thus creating a more normal distribution. It is possible when comparing the performance among different groups of students (e.g., race/ethnicity) or between different programs for there to be different findings, or a different magnitude in findings, depending on which scale, such as the TLI scale or θ scale, is used. Simply comparing the performance between two or more groups without knowledge of the scaling procedures used could lead to erroneous conclusions.

How the 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. If the tests are only equated within a grade and not between one grade and another, then it is possible to do a trend analysis at each grade level, but not a cohort analysis, where the performance of the same group of students are compared across grades. NAEP scaling practice allows an analysis of the growth in performance for the group of students from grade 4 in 1992 to grade 8 in 1996 (Webb, Kane, Kaufman, & Yang, 2001; (Barton & Coley, 1998). However, the structure of the TAAS data, Massachusetts data, and data from other states, makes cohort analysis more difficult and misleading if the scales at the different grades are interpreted as being the same. In developing statistical models to analyze student achievement over time, it is important to know whether the analysis assumed that the tests administered at each grade level can be considered as a repeated measure or not.

Ideally, a new test form will be administered each year the test is given with a random sample of items from the content domain. However, this has monetary implications for states: it is generally too expensive. Each time the same test form is given in repeated administrations, there is a strong likelihood that students will have some prior knowledge of items on the test. Even when new forms are given at each administration of the test, if the test is designed to measure the same content domains and objectives and if similar types of items are used, there is a strong possibility that students will experience some "teaching to the test," particularly if the stakes are high. Such practices reduce the generalizability of the test and what the test scores represent. Improvement in scores may be due to increasing familiarity of students with the item format rather than with the underlying content the item is intended to measure. Understanding the context of how the tests are developed and the repeated administration of tests is important in using test scores to judge the effectiveness of programs and in distinguishing superficial learning from substantive learning of content. Using multiple measures of students' knowledge is one way to counter this problem.

Simply looking across time at increases in mean scores or the percent of proficient students can misrepresent the complexity of the data. Frequently, linear growth models based on change in mean scores over time ignore the common principle that the rate of growth over time generally decreases in the higher grades. Quadratic and higher- order models are needed to represent both the increase in mean scores over time and over grades and the rate of change (positive or negative) in these means. In addition to the order of the model—i.e., quadratic vs. linear—it may be necessary to incorporate other variables into the model to fully explain the change over time as well as the impact of the program being evaluated. Prior achievement is a very important covariant, along with demographic and contextual variables—including socioeconomic (SES) status, parents' level of education, free or reduced-cost meal status, gender, and race/ethnicity. More sophisticated growth-modeling procedures also will consider factors related to the test itself, varying from standard error of measure and differential effect to growth related to different subscales. The main point here is not to give an exhaustive list of all of the important variables to be considered in constructing causal models, but to acknowledge that such variables do exist and must be incorporated into models if we are to achieve a more complete understanding of the relationship between growth in student achievement and a program such as the systemic initiatives program.

Content Factors

In order to adequately interpret assessment results and judge how these results relate to the impact of a systemic initiative, some knowledge is required about what the tests are measuring, how the content on the tests compares to the state standards, and what the relationship is between the content and grade-level emphasis of the systemic initiative. Along with technical factors, information about what is measured by the assessments, how the assessments are structured, and what assessment results are reported are necessary for us to assure that there is a link between the content focus of the initiative and the assessments.

Alignment

Alignment between the assessments, state standards, and the systemic initiative should be assumed, but is one content factor that, if ignored, may result in drawing invalid conclusions. Knowledge of the degree of alignment between these components is necessary for evaluators to be sure that drawing a link between the assessment and the initiative is conceptually sound. A high degree of alignment among assessments, standards, and curricula requires that these components be in agreement on a number of criteria, including categorical concurrence, depth-of-knowledge consistency, range-of- knowledge correspondence, and balance of representation (Webb, 1997). It is possible that a state assessment only measures a small range of the content being covered by the initiative and therefore represents a very poor measure of the full impact of the initiative. There may also be incongruities between the assessment and the initiative's content focus if the state assessment adheres to a state accountability policy that requires raising the scores of the lower 25% of the population. Because schools are judged on the basis of increases in scores at the lower end of achievement by the student population, the state assessment may include a higher proportion of assessment items at a lower level of depth-of-knowledge and only weakly address the other levels. Again, this may lower the effectiveness of the state assessment in measuring a systemic initiative that is targeting the improvement of mathematics and science performance by all students.

Equivalence of Items in Repeated Testing

Testing student performance from one year to the next requires administering the same test or equivalent tests each year. Test security and teaching to the test are only two of the problems associated with using the same form for more than one testing time. To guard against issues associated with repeated use of the same form, test developers use test equating to make statistical adjustments so that the scores from each form will be comparable (Feuer et al., 1999). There are different methods used for equating tests—methods that will vary depending on whether the test forms are developed using the same assessment framework, the same test specifications, or the same sample of items. Even statistically equated test forms can produce some variation in student performance among forms because different items are used. Such variation is likely in subtest scores. Equivalent forms can produce valid statistical information on large population of students, but there can be variation due to tests that use equivalent forms when scores are reported by subpopulations that include district, schools, and individuals. This increases the problem of attributing gain scores, when using new tests each year, to a specific cause such as a systemic initiative because it is difficult to distinguish whether some of the variation in scores from year to year may be due to using a new test each time. True effects are more apparent if gains in student scores are tracked over a number of years and there is a clear trend in improved scores.

Changes in the Population Tested

Technical and content factors heavily contribute to the meaning that can be derived from assessment results for studying program impact, but these are not the only factors. Other critical factors pertain to those who are tested and to general changes in the population. To adequately interpret assessment data and to make inferences about trends over time requires taking note of

system policies that may exclude or provide alternative conditions for testing specific groups of students.

Inclusion and Exclusion of Students Tested

Not all students are tested on statewide tests. This was particularly true in the early1990s and before. Administrative rules have been issued that allow schools to exclude from testing special groups of students, including some or all special education students and limited-English-speaking students. A few states, such as Wisconsin, allowed parents to choose to exclude their children from state assessments. Thus, we see that state or district assessment scores over time can be affected by simply changing the policy dictating what students are, or may be, excluded from taking the test and how assessment scores are reported. Some states will compute the percentage of students at proficiency or basic, based on those students who are enrolled in school rather than on the students who took the test. This is done to increase the incentives schools have to test all of their students and to make efforts to have students absent on the testing day make up the test. Other states will compute these indicators only on the students tested. In still other situations, students may have been tested, but their test scores may not have been entered into the database. Recently, there has been more pressure for schools to test and to include all students in the accountability system to encourage teachers and principals to address the learning needs of every child. At the same time as the standards for learning and course demand increases, this could result in a higher percentage of students dropping out of school and being excluded from the accountability system. This is a greater issue at the high school level than at the elementary and middle school levels. Assessment scores can increase simply because lower-performing students are dropping out and are no longer in the accountability system. An excessive number of students dropping out because of increased standards could be one negative consequence of a systemic initiative. Systemic reform may also be responsible for changing policies to include students in the assessment system, which in turn could actually lower test scores for at least one or two testing times because a larger number of students are being tested. Interpreting assessment data and relating change in student scores to a program in the absence of complete information about inclusion policies can produce erroneous interpretations of the data and inhibit making an association between the results of the program and the assessment.

Population Shifts

Over a period of five to ten years, a state or district may experience a change in economic conditions or an influx of immigrants. In California, the percentage of Hispanic students enrolled in public schools increased from 28% in 1986 to 41% in 1998 (U.S. Department of Education, 2000, Table 44). In Texas, the percentage of students enrolled in bilingual education in public schools increased from 8.9% in 1991 to 12.5% in 1999. The closing of a major industry or the rapid expansion of industry in an area both can result in changes in school enrollment in a district. Any dramatic shift in populations can influence state and district assessment scores either positively or negatively, particularly if these shifts in populations are related to economic conditions. This implies that along with tracking change in assessment scores over time, it is important to track or factor into the analysis other conditions as well, to be assured that these changes are not simply the result of shifting population, but are true effects that can be related to district program changes.

Student and Teacher Mobility

Mobility of students between schools generally affects the interpretation of school-level data more than the interpretation of district- or state-level data. Families of students will stay in the same area, but for a variety of reasons may move students from one school to another. In high-mobility areas such as low economic areas or areas close to military establishments, it is not unusual for a high percentage of students to leave school during an academic school year or enroll in different schools during the course of the academic year. The high mobility of students between districts or in and out of a state is more rare, but still can be a factor in analyzing longitudinal data on the same students over time. A high percentage of mobile students reduces the number of students that can be included in a longitudinal analysis to only those students who attended the school implementing or participating in a program such as a systemic initiative. However, if the initiative addresses more systemic issues, as intended, mobility of students among schools within the USI district or among districts within a state is less important as an issue.

In some schools, there also can be high turnover in teachers. The mobility of students and teachers tends to confound the problem of trying to detect effects attributable to a specific program. Evaluating the intensiveness of a program at a school experiencing high teacher turnover can be problematic if the primary means for training teachers are workshops offered at certain fixed times on an annual basis. For example, in order to reach a larger number of schools, a district may offer summer workshops for teachers from some schools one year and then during the next year offer workshops for teachers at other schools. Or teachers may volunteer to attend summer workshops that could result in a wide variation in the number of teachers with training at any one school and thus the concentration or effect of the program. Unless schools have some way of expanding the impact of these workshops and initiating new teachers into the instructional mode being offered, the effect of the workshops can diminish in time. Such dampening of program effects over time, of course, depends on many other factors in addition to teacher turnover, as has been noted above.

To have 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. Such information is needed to document the percentage of students who have been exposed to the program, the degree of this exposure, the proportion of the faculty at schools who had been trained by the program, the likelihood that student test scores could have been influenced by the program, and the attrition rate of students in the cohorts used to establish the longevity of program effects over the test years.

Summary of Factors that Impact Interpretation of Data

Simple descriptive statistics of the change in student test scores over time can be deceiving if not inaccurate in portraying growth in student learning that might be related to a specific program such as a systemic initiative. Effective systemic reform will need to tap into a number of resources and on-going programs to reach the magnitude and breadth required to have a measurable impact on student learning across the system. This implies that a number of compound factors working in conjunction with each other can influence change in students' test

scores. Studying the impact of a systemic initiative on student test scores requires accounting for the major competing factors that can influence student learning in addition to and in conjunction with efforts of a systemic initiative. One important task in substantiating the impact of a systemic initiative is to establish how meaningful student test scores are over time. Another important task is to establish the effects, statistical or other, that can be attributed to the systemic initiative through multiple analyses.

We have found that three general categories of factors need to be considered to ascertain both the meaningfulness of student test scores over time and the effects that can be attributed to a systemic initiative. *Technical factors* pertain to the psychometric quality of the assessment instruments, testing designs, and analysis procedures being used to produce valid measures of important student content learning over time. Technical factors include comparing scores of different cohorts of students over time, the degree to which different test forms are equivalent, selection of a scale for reporting assessment results, and the possible contamination of data. In the absence of technical information on the assessments, their designs, and analyses, it is very difficult to determine whether a change in student test scores has occurred over time, much less whether any change can be attributed to a systemic initiative. *Content factors* concern the substance that is measured by the assessment instrument and the depth to which students' knowledge of subject matter is measured. How meaningful changes in test scores are over time and whether any change can be attributed to a systemic initiative will depend on content factors such as the alignment of the assessments with standards and the content emphases of the initiative, the representativeness of items on different test forms, and changes in content coverage from one year to the next. *Population factors* are those factors related to the group or groups of students within the system targeted by the initiative, the group of teachers who are the potential implementers of what is being advanced by the initiative, and the changes in these groups. Population factors can be affected by changes in policy, such as including or excluding specific groups of students in testing such as LEP students or special education students; critical shifts in the community population served by the education system; high mobility of students and teachers within the system and in and out of the system; and the selective exclusion of students tested from the database. Important shifts in populations or inclusions or exclusions of students in testing can have a significant impact on the meaningfulness of scores over time—comparing scores at one time with scores from previous times, but with a different composition of students who are tested.

References

Feuer, M. J., Holland, P. W., Green, B. F., Bertenthal, M. W., & Hemphill, F. C. (1999). *Uncommon measures: Equivalence and linkage among educational tests.* Washington, DC: National Academy Press.

Johnson, E. G., & Allen, N. L. (1997). Theoretical background and philosophy of NAEP scaling procedures. In N. L. Allen, F. Jenkins, E. Kulick, & C. A. Zelenak (Eds.), *Technical report of the NAEP 1996 state assessment program in mathematics.* Washington, DC: National Center for Education Statistics.

Barton, P. E., & Coley, R. J. (1998). *Growth in school: Achievement gains from the fourth to the eighth grade* (ETS Policy Information Report). Princeton, NJ: Educational Testing Service, Policy Information Center.

Texas Education Agency. (2000). *Pocket Edition: Texas Public School Statistics.* http://www.tea.state.tx.us/perfreport/pocked/index.html

U.S. Department of Education. (2000). *Digest of education statistics 2000.* http://nces.ed.gov/pubs2001/digestdt044.html

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

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education.* (Research Monograph No. 6.). Madison, WI: Council of Chief State School Officers and National Institute for Science Education.