7. USING MULTILEVEL GROWTH MODELS TO ASSESS THE IMPACT OF THE URBAN SYSTEMIC INITIATIVES IN TEXAS, 1994-2000

Adam Gamoran

Monitoring trends in student achievement over time can be a powerful tool for assessing the impact of the Systemic Initiative. Researchers are often faced with a dilemma, however, in deciding whether the impact of reform is best indicated by monitoring trends *for a particular grade level* over time, or whether it is best to focus on trends for *a particular group of students* over time.

Multilevel modeling can address both of these aims. With a multilevel model of test scores over time, nested within students, nested within schools, we can observe the progress of successive cohorts at a particular grade level, and changes over time for a focal cohort of students, with the same statistical analysis. In this report, we demonstrate this approach using test scores from Texas grades 3-8 during the time period of 1994-2000. The data set includes all elementary and middle schools in Texas, a total of over 5,000 schools, over 3 million students, and over 10 million test scores. However, the analyses used in this demonstration are based on selected subsets of the data as described below.

The Multilevel Growth Model

The multilevel growth or longitudinal model may be thought of as a system of equations in which the independent variables from one level of analysis (e.g., race at the student level) become dependent variables at the next level of analysis (e.g., the average racial gap at the school level). This approach is referred to as a "growth" model because the first level consists of test scores for a particular student over time. These test scores may be described by an intercept—in our case, that is 3rd grade achievement in 1994, the first year of our data—and a slope, which reflects changes in test scores over time. Since the Texas Learning Index (TLI) is normed within grade levels, changes in TLI for a student over time reflect changes in relative position rather than absolute growth. At the level of test scores, we can model not only test score trends over time, but also by grade. This allows us to monitor separate trends for grade levels from one year to the next and for cohorts of students as they move through the school system over time. Thus, the level 1 equation is:

$$(achievement)_{hij} = \pi_{0ij} + \pi_{1ij}(Year) + \pi_{2ij}(Grade) + \omega_{hij} \qquad (1)$$

where h = test scores, i = students, j = districts in district-level analysis, or schools in the analysis of Dallas schools.

For convenience this equation lists only linear trends, but a variety of non-linear specifications can also be examined. In preliminary analyses with statewide data, we found clear evidence of a declining rate of growth as students reached 7th and 8th grades, so we have taken that into account in modeling test score trends. In most of the analyses, we include quadratic (squared) terms for year and grade in the model to allow for curviliearity in the achievement growth trends.

Although Year and Grade vary independently in the sample as a whole (for example, in 1994 the data contain 3rd graders, 4th graders, 5th graders, etc., and these same grades are represented in subsequent years), the two indicators are *not* independent within students because students move ahead one grade each year (except for a very small number of retainees, a factor not addressed in this study). Consequently the effects of both year and grade are "fixed" as equal across students, net of the observed student characteristics specified at the next level of analysis: sex, race/ethnicity, and free/reduced-cost lunch status. (Technically, either year *or* grade, but not both, could be freed to vary across students. Model parameters were not sensitive to this specification change, so we fixed both year and grade.)

Students are the second level of analysis. Here, the model considers whether initial student achievement and changes over grades and time vary according to student characteristics, such as gender, race/ethnicity, and free/reduced-cost lunch status:

$$\pi_{0ij} = \beta_{00j} + \beta_{01j}(Race/ethnicity) + \beta_{02j}(Free/reduced\_lunch) + \beta_{03j}(sex) + \varepsilon_{0ij} \quad (2)$$

$$\pi_{1ij} = \beta_{10j} + \beta_{11j}(Race/ethnicity) + \beta_{12j}(Free/reduced\_lunch) + \beta_{13j}(sex) \quad (3)$$

$$\pi_{2ij} = \beta_{20j} + \beta_{21j}(Race/ethnicity) + \beta_{22j}(Free/reduced\_lunch) + \beta_{23j}(sex) \quad (4)$$

There are no error terms in equations 3 and 4 because parameters are fixed across year and grade as noted above. That means these parameters do not differ among students within schools except according to sex, race/ethnicity, and free/reduced lunch status.

The main research question bears on whether achievement trends over time, and achievement growth for individual students across grades, are affected differentially according to whether their school or district participated in the Urban Systemic Initiative (USI). Thus, the third level of analysis is district or school:

$$\beta_{00j} = \mu_{000} + \mu_{001}(USI) + \rho_{00j} \quad (5)$$

$$\beta_{10j} = \mu_{100} + \mu_{101}(USI) + \rho_{10j} \quad (6)$$

$$\beta_{20j} = \mu_{200} + \mu_{201}(USI) + \rho_{20j} \quad (7)$$

Similar equations test the impact of the USI on the racial, ethnic, gender, and free-lunch gaps within schools. For example the equation for changes over time in the gap between students who do and do not receive free or reduced-price lunch would be written as:

$$\beta_{22j} = \mu_{220} + \mu_{221}(USI) \quad (8)$$

Assessing the Effects of the Systemic Initiative

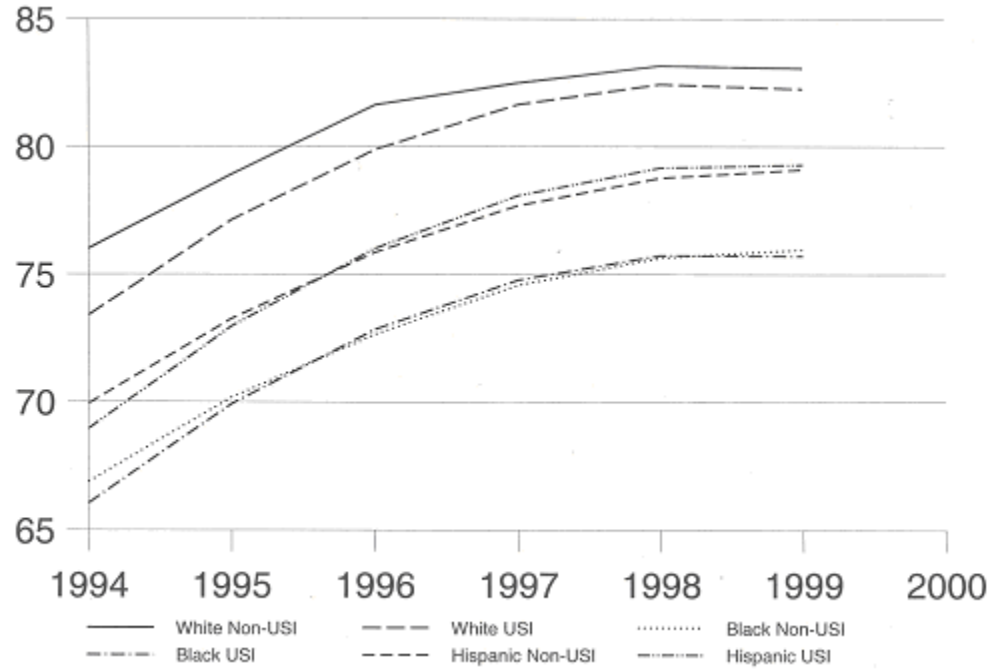
The first challenge in using state test data to assess the impact of the Systemic Initiative is to find an appropriate contrast. We need to compare students whose schools participated in the Initiative to similar students whose schools did not. Thus, state test data cannot tell us much about the impact of the Statewide Systemic Initiative, because it does not offer a contrast between participating and non-participating states. State test data are better suited to assessing the Urban Systemic Initiative (USI), because we can compare districts that do and do not participate within the same state. Moreover, we may compare schools within a district based on the intensity of their involvement in the Initiative.

USI vs. Non-USI District Comparisons in Texas

We identified 13 Texas districts that participated in the USI, including Dallas, San Antonio, and El Paso, and several adjoining districts. (Houston's participation in the USI came too late to be reflected in our analyses.) We constructed a data set consisting of 59 districts that were eligible to participate in the USI based on the USI criterion that specified inclusion of large urban districts with high concentrations of disadvantaged students. Because several adjoining districts participated in the San Antonio and El Paso USIs, we included other districts that were contiguous with these and other large urban districts on the presumption that they could have participated in the USI, but did not. (In fact, several of these adjoining districts did later participate as part of the Houston USI.) Thus, the first contrast that allows us to assess the effects of the Systemic Initiative consists of trends in student achievement in USI districts, compared to other districts that were eligible but did not participate in the USI. We took a random 2% sample of students from these 59 districts, and included all of their test scores. This totaled 39,265 students and 101,040 test scores from the 59 districts. Analyses were performed using HLM 5.0 software.

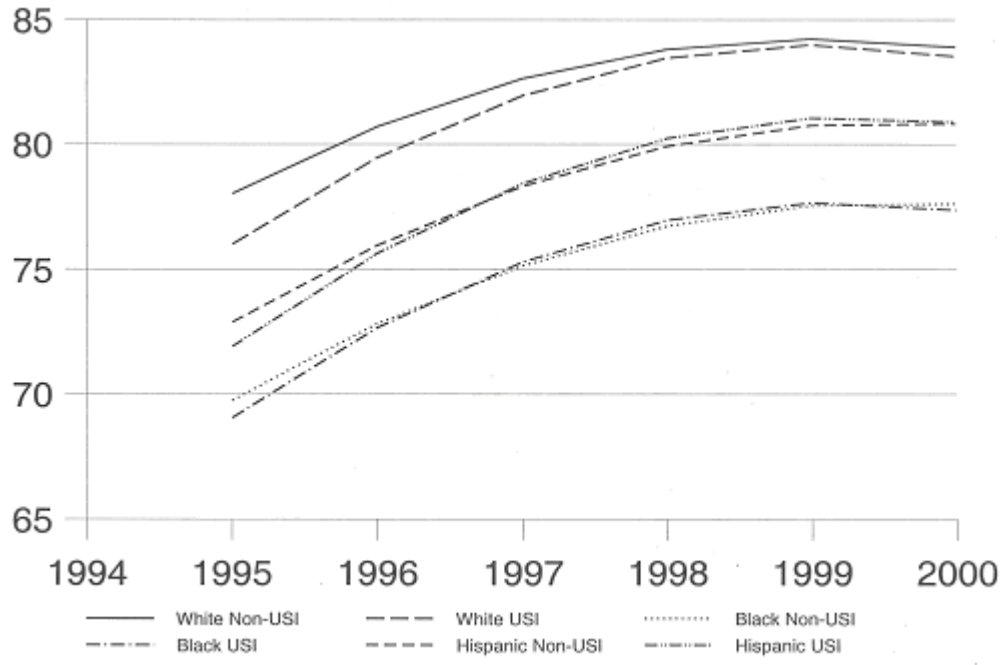
Illustrative results from this analysis are portrayed in Figures 7.1 – 7.3. These figures show trends in student test scores for three successive cohorts: students who were enrolled in 3rd grade in 1994, 1995, and 1996, respectively. The first cohort completed 8th grade in 1999 and the second in 2000, so their progress is fully observed in our data. The third cohort only reached 7th grade in our data, because they completed 8th grade in 2001. The figures portray trends separately for Black, Hispanic, and non-Hispanic White students in USI and non-USI districts. (For ease of viewing, only results for girls not on free lunch are presented, but the analysis controls for gender and free-lunch status so the lines would be identical, albeit at lower levels, for boys and for those on free lunch. Not shown also are results for a small number of students who reported other ethnicities, or who reported different ethnicities in different years.)

Figure 7.1. USI vs. non-USI districts: 1994 grade 3 cohort.



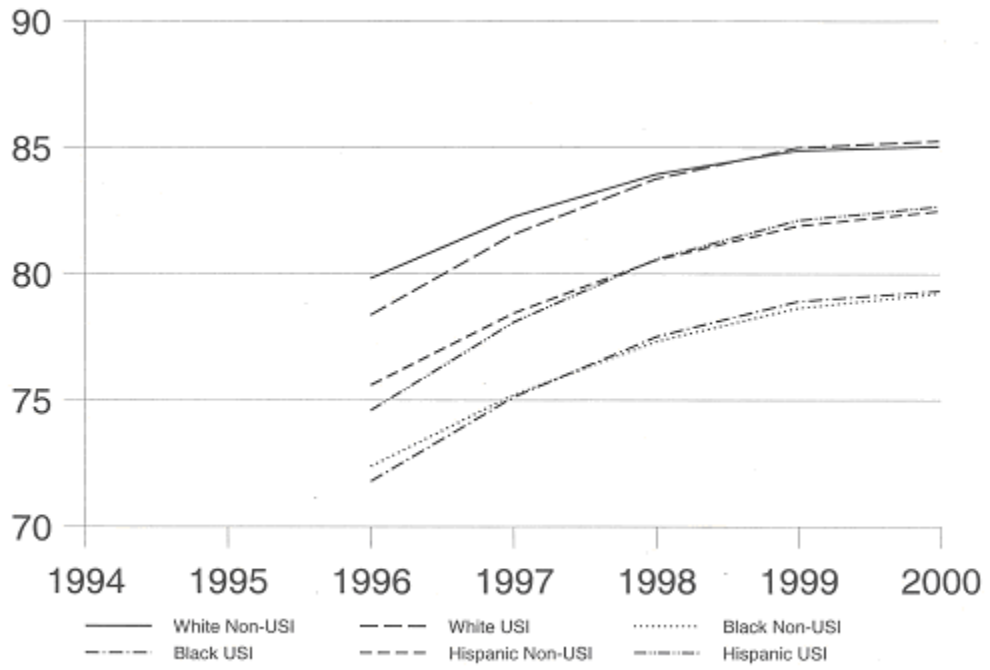
Note: Data for non-disadvantaged females; same patterns hold for other groups.

Figure 7. 2. USI vs. non-USI districts: 1995 grade 3 cohort.



Note: Data for non-disadvantaged females; same patterns hold for other groups.

Figure 7.3. USI vs. non-USI districts: 1996 grade 3 cohort.



Note: Data for non-disadvantaged females; same patterns hold for other groups.

Three important points are evident in the graphs. First, students in USI districts began 3rd grade with lower test scores than students in non-USI districts. This is evident in all three graphs and for all three demographic categories. Second, students in USI districts make faster progress than their counterparts in non-USI districts, particularly in the early grades. For example in Figure 7.2, White USI students scored about 2.5 points below their USI counterparts in 1995, but the gap had shrunk to a fraction of a point by the time they reached 6th grade in 1998. In the same chart, Hispanic and Black USI students started out about 1 point below their non-USI counterparts, respectively, but by 1998 the USI achievement level for these groups was actually above that of similar students in the non-USI districts. The differences in these trends are all statistically significant, as judged by significant coefficients for initial differences between USI and non-USI districts and by significant coefficients for differential growth across grades and years for USI versus non-USI districts. Third, there is a clear curviliearity in the achievement trend. This downturn is also statistically significant, as evidenced by significant parameters for the quadratic terms in the model. The downturn appears slightly greater in USI than in non-USI districts, but this difference is not statistically significant.

With a single cohort of students, it would be impossible to tell whether the downturn reflects a period of time, or, differences across grade levels. The multilevel growth model allows us to compare successive cohorts so we can try to distinguish changes over time from changes across grade levels. Comparison of Figures 7.1, 7.2, and 7.3 suggest that it is the transition to 8th grade, rather than changes in the late 1990s, that is responsible for the downturn. The drop-off in TLI scores that appeared among 1999 8th graders (Figure 7.1) occurs in 2000 in the second cohort of students (Figure 7.2), when they reached 8th grade, not in 1999, when they were in 7th

grade. In the third cohort (Figure 7.3, students who would reach 8th grade in 2001) the drop-off does not occur. Inspection of a single figure shows growth for a cohort, and comparison across figures shows trends over time; the multilevel growth model reveals both of these aspects of change.

Overall, the results indicate that students in USI districts made more progress over time than students in non-USI districts. This held for all demographic groups. The USI districts were about equally effective for all demographic groups; race/ethnic gaps declined in USI and non-USI districts at about the same rate.

Implementation of the USI in Dallas Elementary Schools

A second approach to assessing the impact of the USI via state test data using multilevel growth models is to examine achievement trends for students in schools that implemented the guidelines of the USI rigorously, compared to other schools in the same district that exhibited less faithful implementation of USI. For a school-level analysis, it is necessary to separate elementary and middle schools, because most students change schools at the elementary-middle transition. In Dallas, this transition occurs between grades 6 and 7.

To estimate a growth model, at least three test scores per student are needed for most respondents. In principle, analyses of middle schools could use test scores from grades 6, 7, and 8, with the grade 6 score serving as a baseline for subsequent growth. In practice, however, we found too few students with three test scores per middle school to estimate the growth models. Missing test scores may reflect a variety of conditions, including high rates of student mobility, student absences on testing days, or changes in the criteria that established who was to be tested.

At the elementary level, we have access to four test scores (grades 3, 4, 5, and 6), and the growth model was feasible. Even here, however, only about 1/3 of the students had three or more test scores, so our models are somewhat tenuous. Table 7.1 presents the frequency distribution for number of test scores per student. In total, we examined 211,049 test scores for 120,775 students in 133 elementary schools in Dallas. Because of the low number of test scores per student, only linear trends were estimated, but that is reasonable because the curvilinearity mainly occurs as students enter middle school. We also simplified the model by fixing time and grade trends to be equal across schools, except for the effects of implementation.

Table 7.1
Availability of Test Scores from Dallas Elementary Schools, 1994-2000

Frequency of Scores per Student	Percent of Students
1 test score	40.0%
2 test scores	24.4%
3 test scores	18.2%
4 test scores	17.1%
5 or 6 test scores	0.3%

Note: Sample size = 211,049 test scores, 120,775 students, and 133 Dallas elementary schools.

Assessing the extent of implementation—the key independent variable for this analysis—was a major challenge. From public records, we had only one indicator, the time at which the school joined the USI: 1995, 1997, or 1998. Presumably, achievement growth would be steeper in the schools that joined the USI earlier because there would be more time for the effects to operate on student performance. Working with the Dallas school system, we also obtained information on teacher participation in an important USI workshop that occurred during the summer of 1998. We constructed a variety of indicators, including the number of sessions attended, the number of mathematics sessions attended, and the number of teachers participating.

Unfortunately, the results show few consistent effects of USI implementation on trends in student achievement among the Dallas elementary schools. This does not mean the Dallas USI was ineffective—the *district* effect was assessed in the district comparisons above (see Figures 7.1 – 7.3). Rather, it suggests that variation in our measures of implementation was not associated with variation in achievement trends. This could be because the USI effects were similar across the Dallas elementary schools, regardless of their point of entry or participation level in the 1998 seminar, or because our measures of implementation are too weak to capture the effects.

Table 7.2 provides an example of results from these analyses. The coefficient of -.046 for mathematics sessions attended on the 1994 3rd grade intercept, compared to a standard error of .241, indicates there was no meaningful difference at that time between schools whose teachers varied in the average number of workshops attended.

Table 7.2
Selected Coefficients from Analysis of Effects of USI Implementation

	Coefficient	Standard Error
Effects on 1994 3rd grade Intercept		
Math Sessions attended by teachers	-.046	.241
Effects on Change across Grade Levels		
Math Sessions attended by teachers	-.044	.021
Effects on Change over Time		
Math Sessions attended by teachers	.050	.016

Note: Sample size = 211,049 test scores, 120,775 students, and 133 Dallas elementary schools.

The coefficient of -.044 for the effect on changes across grades is statistically significant (standard error of .021), but its practical importance is small: for example, a school whose teachers attended all five mathematics sessions would increase by .22 points per grade level per year *less* than a school whose teachers attended none of the sessions (5 sessions x -.044 points per session = -.22 points). One might conclude that schools whose grade-to-grade progression was weaker tended to send more teachers to the workshops, but the differences are very small. Similarly, the coefficient of .050 for the effect of mathematics sessions attended on year-to-year changes is statistically significant but practically small: A school whose teachers attended all five mathematics sessions would increase student test scores by .25 more per year than a school whose teachers attended none (5 sessions x .050 points per session = .25 points).

Conclusions

These findings demonstrate that the analysis of state test data with multilevel growth models is a viable strategy for assessing the effects of Systemic Initiatives. The major barrier to assessing impact is not the dependent variable (test data), but the independent variable—the lack of powerful indicators of USI implementation. Despite substantial effort to gather retrospective information about implementation, we found little evidence of an association between implementation and impact. Future evaluations of Systemic Initiatives should gather systematic, reliable information about implementation while the Initiative is under way in order to gauge variation in implementation and to see whether that variation is associated with student achievement.