9. CONTRAST AND COMPARISON OF THE THREE ANALYTIC MODELS

Models for analyzing program effects in large systems can differ in very subtle ways. The three models presented here by Bolt, Gamoran, and Meyer are similar in many respects, but differ in the criteria used to select students to include in the study, the statistical analyses performed, and the attention given to the nature and content of the assessment. It is not surprising that each of the models and approaches used by the researchers reflect their academic backgrounds. Gamoran, a sociologist by training, used a multilevel growth or longitudinal model that took into consideration students' socioeconomic status when comparing performance, over time, of student cohorts divided into different subgroups by race/ethnicity. His analysis looked for global differences over time. Bolt, a psychometrician by training, also used a multilevel model, but he employed statistical means to select a match group, conducted a dimensionality analysis of the test forms, and used test-equating techniques to link tests and domains (subtests) across years. Meyer, an economist by training, used a multilevel hierarchical quadratic growth or cross-sectional model, as did the other two researchers, but he was more interested in looking at the increase in scores over one year to compute the value-added to students' progress. He used a state-dependent or piecewise model. For him, the scale used for reporting student test scores was a critical consideration because of the variation at the extremes of the distribution of scores.

All three models, in addressing the effects that could be attributed to the Urban Systemic Initiatives (USIs) in Texas, produced similar results. Performance of USI districts was significantly lower at the outset in 1993 and 1994 when compared to other Texas districts. The USI districts had a higher rate of gains than did non-USI districts over the seven years studied. The effects were greater in the lower grades and during the earlier years. There was little evidence that the USIs had a differential effect on Hispanic students and Black students compared to White students. Bolt's model did identify noticeable differences for Blacks at grades 3 through 5, but these differences were not apparent in grades 6 through 8. All models depicted a clear curvilinearity in the performance of students over time, indicating a lowering of growth by students on the Texas Learning Index (TLI) scale in the later years.

In the remainder of this section, a comparison of the three models is developed in greater detail. The purpose of this discussion is to illustrate the similarities of the three models and to show how they differ. Each model presents a viable approach for studying the impact of systemic initiatives on student performance. Clearly, each approach has to be considered in its own right, for what it tried to do and how well it accomplished this. In evaluating the three approaches, the results produced by each model can be compared to those of the other two, thus highlighting what is valuable in each and what can be questioned. The main features of the three models are summarized in Table 9.1 at the end of this section.

Major Purposes of Each Model

All three researchers sought to identify effects that could be attributed to districts participating in the NSF Urban Systemic Initiatives program. Bolt sought to determine the influence of the USI at the school level for grades 3 to 8 over the years 1994-1999. Gamoran wanted to determine whether a school's or district's participation in the USI affected

differentially the achievement trends of students over time and the growth of individual students across grades. He used results from grades 3 through 8 over the years 1994-2000. Meyer also desired to look at the contribution of the USI on growth by students over time, but he wanted to examine yearly increments and the value that was added to students' performance each year. He examined conditional performance gains for grade 4 to 8 over the years 1995-2000. His analysis began with grade 4 because his model needed to use prior-achievement as a variable. For grade 4, the grade 3 performance was used as prior-achievement. (There was no prior-achievement variable included in the database for grade 3.)

Each researcher went beyond simply answering his central research question. Gamoran attempted to look at the degree of implementation of the Dallas USI in a school, as indicated by the number of teachers who had participated in professional development, and the relationship of this factor to student performance. In this attempt, he tried to establish a direct relationship between a specific activity of a USI in one district to school performance over time. Meyer had concerns about using only the TLI scale because it was not vertically equated and inhibited the comparison of student scores across grades. The Texas Learning Index was not designed to register increases in achievement beyond a maximal level and thus student scores had a ceiling at a value of about 93. He did analyses using both the TLI and the Rasch scale to contrast how results varied depending on the scale used, particularly at the extreme values. In addition, as a separate analysis, he looked at the change in the number of students tested over time by racial groups. Bolt finessed the problem of the TLI not being vertically equated by using a procedure that approximated parameters that could then be used to equate the test scores across the years. He then did a dimensionality analysis of test forms at each grade level to determine whether student performance over time varied on the three domains that were covered in each mathematics test.

As a set, the three analytic models provide information about students' growth over time and compare the performance by students in USI districts with those in non-USI districts. In this respect, the results from one of the models serve as a replication of those from the other models. However, all three researchers made different assumptions about growth that influenced what groups of students were tracked over time. Bolt examined changes in school means on the 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 cohorts of students is due to program effects rather than 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 four data points on a student in order to estimate the three parameters—intercept, slope (linear term), and change in the slope acceleration (quadratic term). Students with fewer than four data points contribute to estimating the lower-order terms. Thus, Gamoran's growth model included students who had left the system 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, a very small percentage, were deleted from the database used in this analysis.

Meyer's value-added analysis examined students' performance in a grade by taking into consideration the achievement of the previous year. 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 single 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 for two years.

Each researcher developed models that asked slightly different questions. The question the researcher asked influenced the assumptions each made and the structure of the model each constructed. All three researchers addressed tracking the effects of USI districts. Each one contributes a somewhat different view about what those effects are.

The Underlying Structures of the Three Models

Comparison Group

All three models incorporated the 13 USI districts in Texas—i.e., all of the districts in the state that had received funding from the National Science Foundation's Urban Systemic Initiative up to 1999. The three researchers differed in what they used for a control group. To adequately analyze the impact of a program, it is important to isolate or differentiate what other factors influenced the program effects. Ideally, at the beginning of the USIs, students would have been randomly assigned to either districts or schools participating in the program or to those not participating in the program in order to obtain comparable control groups. However, this was impossible for a number of reasons, one being that parents wanted some say in where their children attended schools. When, in the absence of random assignment, a researcher has to use existing data, one of the issues that has to be addressed is how to identify an adequate control group that can be used to counter other reasons why there is, or is not, an effect. Fortunately, the researchers had the data set for the entire state of Texas from 1994 to 2000 for grades 3 through 8 to use. This very large data set gave the researchers some flexibility in selecting control groups.

Both Bolt and Meyer contrasted performance of the 13 USI districts with all the other districts that included rural schools, suburban schools, and other urban schools. Comparing USI districts with all districts in the state is a fairly rigorous criterion that implies that students in the urban schools should be doing as well as those in the state overall. The use of all non-USI districts as the control group was less a function of the particular analytic model being used and more for convenience and ease in working with the data. What is not known is whether the non-USI districts were engaging in any practices that were similar enough to those of the USI districts to lessen the differences between the two groups. We do know that all schools and districts within the state functioned under the same accountability system, tied to a common set of curriculum standards.

In addition to using all non-USI districts as a control group to model district-level growth, Bolt used a statistical procedure to form a matched group of non-USI districts that were more similar to the USI districts in performance. It was clear from the data that the non-USI districts included a number of districts that performed much better than the USI districts. Because the TLI scale was not an interval scale, it was impossible to be sure that changes in scores over time would be the same at different points on the scale. This, along with ceiling effects, meant that sustained growth in scores by high performing districts may be apparent. To counter this effect, Bolt formed a group of non-USI districts that matched the USI on average performance over the period 1994 to 1999. Then he compared the scores for each grade by year for all three groups—USI, non-USI matched group, and non-USI total group. Even though the matched-control group was formed to be like the USI on average performance, the pattern of performance across the years could reflect differences in treatment between the USI districts and non-USI districts.

Gamoran was more selective in determining the control group used in his growth model. He identified Texas school districts that were eligible to participate in a USI as determined by the criteria used by NSF—a large urban district, or a district from the surrounding area with a high concentration of disadvantaged students. Fifty-nine school districts in Texas met these criteria, including all 13 USI districts. Gamoran used the 46 other districts, including Houston Independent School District, as a control group to compare student performance over 1994 to 2000. Whereas Bolt used comparable average performance to match school districts, Gamoran used eligibility. Neither of these methods is without problems. Even though a district is eligible to be a USI—in fact, Houston Independent School District joined the USP in 1999—there is no guarantee that the district is comparable in all respects to those districts that joined the program. Some interested party had to take the initiative necessary to receive NSF funding for the USI districts. Districts without a USI may be missing this leadership and initiative that could also have an impact on student performance.

The method the researcher used to select a control group was independent of the model used to analyze the data. Any of the three methods for selecting a control group—non-USI districts, a matched group of USI districts based on average TLI scores, and non-USI eligible districts—is viable. Each method has advantages and disadvantages that need to be identified. What is critical is that each model in some way compared the USI districts with a control group and was able to compute an initial point, or the intercept term, for the equations.

Types of Analysis

All three models used multiple levels and controlled for similar variables. The hierarchical structure of school systems, for students, schools, and districts, lend themselves to the use of multilevel models. Bolt used three levels—school, district, and district type—while Gamoran and Meyer used two levels—student and district. In the analysis Gamoran did on Dallas schools, he also used two levels—student and school. Both Bolt and Gamoran used growth models. Bolt examined growth within each grade over time, whereas Gamoran analyzed growth by cohorts of students over grades and time. Meyer computed change over one year's time and plotted this one-year change for each grade over the six years.

All of the models controlled or accounted for demographic variables of sex, race/ethnicity, and free and reduced-cost lunch status (the one indicator included in the data base used to represent economic disadvantage). In addition to these variables, Meyer controlled for measurement error in the test, a feature not included in the models developed by the other two researchers. Because Gamoran analyzed the growth in performance by cohorts, grade and year were not independent: as a student in a cohort progressed in grade, he or she also progressed in year. Therefore, Gamoran treated both grade and year as fixed variables.

Rarely will a student's learning increase at a constant rate over a long period of time. This can be thought of in many ways. Students have more to learn in the earlier grades and will tend to show more improvement in these grades and less improvement as they continue in school. Or, the test itself can be limiting. If the accountability system and tests are geared to measure a basic skill level students have reached by the end of grade 8, many students will achieve this one, two, or even three years before the end of grade 8. If the test only measures the basic skills, then the test will not provide any information about what the students have done beyond the basic skills. In other words, the students' scores reach the "ceiling" of the test. Such variables as these will cause the growth in scores for groups of students to decline over time. Attending to these and other concerns, both Bolt and Gamoran incorporated quadratic terms into their models. Meyer only considered growth between two points in time, so it was not appropriate to use a quadratic term in his model.

Data on the performance of students on the TAAS were reported using the Texas Learning Index scale. This scale is derived by using a Rasch model that estimates the student score and then predicts the raw score that a student would have received on the original 1994 TAAS. The scale itself is a linear transformation based on the mean for the base year (1994) and normed so that the scale has a standard deviation in the base year equal to 15 and a passing cut point equal to 70. (Meyer explains the TLI at some length in Section 8.) All three of the models used the TLI scale scores of students. However, the way in which the TLI scale was created, by tying all of the scores back to 1994 scores and only equating scores within a grade, placed limitations on the how the scale could be used to study growth over time. Gamoran only used TLI scores. Even though he looked at growth over time, in comparing USI districts with eligible non-USI districts, it is reasonable to assume that the scale would have affected all of the districts in the same way without giving one type of district advantage over another. Bolt and Meyer attended more to the problems that may arise in using the TLI. Bolt used the TLI scores to examine growth in performance within grades—in part, because the TLI scores were equated within grades. He used the Rasch latent-ability metric with a pseudo-linking procedure to study domain scores and total test scores across years. One reason Bolt used the Rasch latent-ability metric was because it was less susceptible than the TLI scale to ceiling effects. Meyer also was concerned with the ceiling effect and did his analyses with both the TLI and a Rasch scale. Both Bolt and Meyers found some differences in the results, depending on which of the two scales was used.

Each of the three researchers employed different estimation methods in producing their results. Gamoran used the well-known *Hierarchical Linear Modeling* software (HLM 5.0) developed by Bryk and Raudenbush (1992) because he was working on only two levels. Bolt used a similar approach, but because he used three levels and the variance of school means was

assumed known, he turned to the Gibbs' sampling procedure that was computed using the *WINBUGS* software (Spiegelhalter et al., 2000). To estimate the maximum likelihood functions for the multigroup models, he used the *LISREL* software. He estimated the IRT models using the *BILOG* software and did the dimensionality analysis on items with the *TESTFACT* program. Meyer estimated the ordinary least squares and other computations, using *Statistics Analysis Software* (SAS). He used Fortran software to compute the Rasch estimates and the finite sample errors.

Bolt and Gamoran produced measures of statistical significance to judge the differences between groups of districts. Bolt found that the intercept terms for the average growth curves were statistically significant at the .05 level for both the USI and non-USI districts. Gamoran also found that the initial differences were statistically significant as judged by the significant coefficients in the model. Meyer computed the standard errors for the USI effect estimates. Because of the low number of USI districts (N = 13) (also a factor in Bolt's analysis), the standard errors were relatively large and the separate estimates by year and grade were not significant. But the overall trend, showing that there were some USI effects over the six years and five grades analyzed support at least modest achievement growth by USI districts over the achievement growth of all of the non-USI districts.

The structure of the three models varied by the number of levels considered, by the assumptions made, the control groups used, the analytic framework applied, and the software employed. However, each approach presents a viable means for studying the effects related to a district having a USI. There were some critical problems that it was difficult for any of the models to overcome, including establishing significant differences 1) when the sample included only 13 USI districts and 2) when there was a lack of information on the independent variables related to the degree to which school personnel participated in USI activities, or were influenced by the USI.

Results from the Three Models

Even with some variation in the analytic framework, all three models as well as additional analysis support similar findings. All three researchers found that in the early years, 1994 and 1995, students in the USI districts performed significantly lower than those in non-USI districts, even when the control group was restricted to districts eligible to be a USI district. From all three models, it was evident that achievement by students in USI districts gained at a somewhat faster rate than that of students in the non-USI districts, thus narrowing the gap between USI districts and non-USI districts. A contributing factor to this narrowing is the ceiling effect of the TLI, although Meyer found similar results using a Rasch scale, which was unaffected by students reaching a maximum score. Bolt found, based on correlations between common coefficients, slightly more variation in the growth patterns among schools in USI districts compared to schools in non-USI districts.

Two of the three models produced differences in the details related to the effect on performance by Hispanic and Black students. Gamoran found that in both the USI and non-USI districts, the racial/ethnicity gaps declined at about the same rate in both USI and non-USI districts. Bolt found little differences between the two groups of districts in the change in

performance by Whites and Hispanics. He did find that Black students in USI districts performed noticeably lower than non-USI Black students in grades 3 through 5, but essentially the same in grades 6 through 8. He did not find any significant differences between USI and non-USI districts by race in the intercept, linear, or quadratic parameters.

Bolt and Meyer conducted additional analyses that produced some interesting findings. In analyzing the tests, Bolt found that the TAAS had multiple dimensions that corresponded with the three mathematics content domains used to develop the tests. This analysis indicates that the TAAS presents some psychometric complexity that is more challenging than simple computation. Meyer developed models on the number of students tested by race/ethnicity and economic disadvantaged. His models explained more than 75% of the variation in the number of students tested for 1) Hispanic students, 2) disadvantaged students, and 3) non-disadvantage students. From 4 to 6% of the variance in the number of Hispanic students and disadvantaged students tested was related to the year the student was tested and the cohort. This indicates that for these populations the number of students tested varied some over the period investigated. The exact cause of this variation was not determined. There was very little variation in the number of White and Black students tested. But from 1994 through 2000, there was a significant increase in the number of Hispanic students tested and in the number of disadvantaged students tested.

Table 9.1
Summary of the Comparison and Contrast of the Three Models

Feature	Model		
	Meyer	**Bolt**	**Gamoran**
Model Conceptualization			
1. Research Question	What is the contribution of USI to student achievement over time, considering post-assessment on pre-assessment over one year and considering sex, economic disadvantaged, race, and measurement error?	What is the influence of the USI at the school and district levels for grades 3-8 over the years 1994-99?	Are achievement trends over time, and achievement growth for individual students across grades affected differentially according to whether their school or district participated in the USI and by the degree of implementation by school within Dallas?
2. Total Population or Sample	1. Total population 2. Compared 13 USI districts with all the other districts.	1. Compared 13 USI districts with all other non-USI districts. Compared 13 Texas USI districts with 834 to 1052 non-USI Texas districts (number varied by grade). 2. Formed a matched group of non-USI districts comparable in terms of average TLI scores, 1994-99. Computed for each grade, a mean TLI score across years within each school. Computed distribution across all USI schools. Assigned probability for selection to each non-USI school. Created a sample of non-USI schools with TLI distribution to USI schools. Compared 102 (grade 7) to 376 (grade 3)	59 Texas school districts eligible to participate in the USI based on the USI criteria of large urban districts with a high concentration of disadvantaged students. Of these, 13 Texas districts had participated in the USI. A 2% random sample of students from the 59 districts was used in the analysis.

Table 9.1 (Continued)

Feature	Model		
	Meyer	**Bolt**	**Gamoran**
Model Conceptualization			
		schools in USI districts with 417 (grade 7) to 1,144 (grade 4) schools in non-USI districts (22% to 49% of non-USI schools across grades). 3. Random sample of 1,000 examines at each grade (3-8) in 1994 for dimensionality analysis of TAAS forms.	
3. N of Total Student sample	*N* = 220,000 to 260,000 per grade (grades 3-8)	N=790,000 tested each year, nearly 10 million student records over 7 years.	1. 39,265 students 101,040 test scores 2. 133 Dallas schools 12,775 students 211,049 test scores
4. Conditions for Inclusion	Two consecutive test scores for a student.	Student tested at least once, random selection, and school results.	Consecutive test scores for all years for student.
5. Baseline Data	Base periods were year 1995 for time, grade 3 for grade, and 1989 for cohort.	Grade 3-8 achievement levels in 1994.	Grade 3 achievement in 1994.
6. Grades	4-8	3-8	3-8 A total of 10 cohorts (up to 6 grades over 7 years and with at least 2 years of tests).
7. Unit of Analysis	Two levels, student and district.	School, district, and district-type.	1. Student and school levels. 2. Student and district levels for USI analysis.
8. Control Group	All districts without a USI.	All districts without a USI by grade and year.	46 eligible districts that had not participated in a USI by 1999.
Technical Features			
9. Covariants	Sex, race/ethnicity, and	Three racial groups (Whites, Hispanics, &	Sex, race/ethnicity, and

Table 9.1 (Continued)

Feature	Model		
	Meyer	**Bolt**	**Gamoran**
Model Conceptualization			
	free/reduced lunch status.	Blacks).	free/reduced lunch status.
10. Scales and Test	TLI & Rasch on total test score.	TLI & Rasch latent-ability metric for total test scores; Rasch ability metric for three test-item domains.	TLI on total test score.
11. Type of Analysis	1. State-dependent growth model (piecewise) 2. Two-level estimates—student and district. 3. Linear regression analysis. 4. Value-added.	1. Hierarchical quadratic growth model: Three levels: school, district, and district type. 2. Dimensionality analysis of test forms at each grade level. 3. Psuedo-linking procedure to equate domain scores and total test scores with respect to Rasch latent-ability metric across years. Used IRT equating procedures to link ability metrics for all forms from 1995-99 back to 1994 ability metric in each grade. Growth was assessed with respect to changes in the Rasch ability metric. 5. Used a multivariate multigroup (USI versus non-USI) latent-growth model applied to the Rasch latent-ability scores. A separate multigroup model was fit for each grade.	1. Multilevel Growth. 2. System of equations where independent variables at one level of analysis become the dependent variables at the next level of analysis. 3. Monitor separate trends for grade levels and for cohort of students. 4. Controls for gender and free lunch status. 5. By analyzing multiple-cohort, can distinguish between effects related to year or grade.

Table 9.1 (Continued)

	Meyer	Bolt	Gamoran
		Technical Features	
12. Estimation Methods	OLS: ordinary least squares EV1: errors and variables method based on exact finite sample estimates (eliminates bias due to measurement error) EV2: . . . based on maximum likelihood estimates. IV: Instrumental variables method, with twice-lagged achievement. IV-H: . . . with weighting to control for heteroscedasticity.	Used Gibbs' sampling, as implemented in the software program WINBUGS, to estimate multilevel models. For multi-group models, maximum likelihood estimation using the LISREL software was conducted. For estimation of IRT models, the software Program BILOG was used, while the program TESTFACT was used for dimensionality analysis. Both BILOG and TESTFACT use a marginal maximum likelihood procedure for estimation.	Three-level HLM.
13. Compounding factors	1.Error in student achievement method used. 2. Data set only includes a limited number of control variables.	Data set only includes a limited number of control variables.	Data set only includes a limited number of control variables.

Table 9.1 (Continued)

	Meyer	Bolt	Gamoran
		Model and Variables	
14. Variables/ Parameters	1. Assumed variables are measured without error, except pre-test was allowed to have error. 2. Post-achievement predicted using prior achievement, student characteristics, school/district indicators.	1. Intercepts, linear coefficient, and quadratic coefficient for school, district, district variances, USI, non-USI, USI/non-USI variances, race in multilevel models. 2. Intercept is expected TLI score in 1994; linear coefficients are the increase in TLI scores per year starting in 1994; quadratic coefficients are the rate of change. 3. Total TLI score, item difficulty, and domain score.	Parameters are fixed across both year and grade, student achievement, grades, year, race/ethnicity, free/reduced-cost lunch, sex.
15. Software	SAS	WINBUGS (Spiegelhalter et al., 2000), TESTFACT, BILOG	HLM 5.0
16. Order of Equation (Linear/Quadratic)	Linear	Linear and Quadratic	Include quadratic terms for year and grade to allow for curvilinearity in the achievement growth trends.
17. First-Level Equation	$\boldsymbol{Y_{igt} = \gamma_g Y_{ig\text{-}1t\text{-}1} + \beta_g' X_{igt} + \Sigma\, \alpha_{jgt} S_{ijgt} + \varepsilon_{igt}}$ $\boldsymbol{Y_{igt}}$ = Achievement time t $\boldsymbol{X_{igt}}$ = Student characteristics $\boldsymbol{S_{ijgt}}$ = school/district	$\overline{X}_{ijk} = \alpha_{jk} + \beta_{1jk} i + \beta_{2jk} i^2 + e_{ijk}$ X_{ijk} = Mean TLI score i = Year	$\boldsymbol{(achievement)_{hij} = \pi_{oij} + \pi_{1ij} (Year) + \pi_{2ij} (Grade) + \omega_{hij}}$

Table 9.1 (Continued)			
	Meyer	**Bolt**	**Gamoran**
Precision			
17. Measurement Error	Controlled for measurement error in test.		
18. Statistical Significance		Judged for the parameter estimates of the multilevel and multigroup models.	1. Judged by significant coefficients for initial differences between USI and non-USI districts, and significant coefficients for differential growth across grades and years for USI versus non-USI districts. 2. Considers statistical significance of USI implementation and achievement effects.
Results, Outputs, and Limitations			
19. Benefits	1. Model only requires data on one student for two consecutive years. 2. Accounts for differences in achievement growth associated with demographic factors, along with test measurement error.	1. Studying change in a fixed grade level across years will more effectively control teacher effects because the same teachers are more likely to be teaching the same grades in successive years. 2. Changes in TLI can be more confidently interpreted as changes with respect to a constant test performance metric because TAAS is equated across years within grade.	1. Multilevel modeling can address both monitoring for a particular grade level over time and trends for a particular group of students over time. 2. Add demographic factors.

Table 9.1 (Continued)			
	Meyer	**Bolt**	**Gamoran**
20. Tradeoffs		This approaches does not control for the change in students that comprise the grade levels in each school from year to year. Thus, change in improvements in instruction will be confounded with changes in students tested.	1. Assumes achievement growth is captured by linear and quadratic terms. 2. Assumes same achievement growth for students in same school, grade, and year, given gender, race/ethnicity, free-reduced-cost lunch.
21. Output of Analysis	Compute departure from the slope (effect) trajectories on the 3 variables (t, g, & c) in the base year. Number of students tested by year, race	Model-based estimates of trend lines within grades comparing USI and non-USI across years; model-based trend lines for each district comparing grades across years; parameter estimates for initial score (1994), linear change, and growth; estimates of variances.	1.Separate trends for black, Hispanic, and non-Hispanic white students by USI and non-USI districts. 2. Coefficients and standard error of effects on change across grade levels and over time. 3. Weak data on effect of Dallas implementation.
22. Findings	**USI Effects:** 1.USI contributed modestly to student growth and increased over time. 2. Effects generally were larger in the earlier rather than larger grades.	1. From exploratory analysis of trends within grades from 1994 through 1999, in grades 3 and 4 the sharpest growth was in the first two years and then the rate of growth decreased. 2. Grades 5-8 had more constant growth across all years. Grades 6-8 had little growth in first year, but sharp growth the next year (1996).	1. Decline in rate of achievement growth at grades 7 and 8. 2. Students in USI districts began grade 3 with lower test scores than students in non-USI districts. This was true for all three demographic groups.

Table 9.1 (Continued)

	Meyer	Bolt	Gamoran
22. Findings (continued)	**Test Taking Patterns:** 1. Model explains more than 75% of the variation in number of students tested for Hispanic, disadvantaged, and non-disadvantage students. 3.7% of variance in number of Hispanic students tested is related to time and cohort. 5.9% for disadvantaged students. 3. The effect of economic disadvantage varied by year and grade, but generally remained constant between 1995 and 2000. 3. Rasch scale, compared to TLI scale, produced slightly different results in the change of the effect of the economically disadvantaged.	3. Although non-USI districts performed higher than USI, there was a slight narrowing of the differences in all grades in the later years. 4. Variation among 13 USI districts in TLI scores over time by grade. 5. Higher TLI scores at grades 5 and 6 may indicate the TLI metric varied across grades. 6. Little differences in USI and non-USI for Whites and Hispanics, but more noticeable differences for Blacks at Grades 3-5, which disappear in Grades 6-8. 7. USI and non-USI districts differed—statistically significant ($p<.05$) on intercept parameter, but not on differences in linear and quadratic coefficients. This was true for the total test and the three domains. 8. USI districts compared to non-USI matched districts at grade 7 showed higher relative gains in 1998 and 1999. (Could be due to small N for USI.) No differences were found for any other grade. 9. No significant differences by race in the intercept, linear, or quadratic parameter between USI & non-USI districts. 10. Tests contained some multidimensionality. Three factors corresponded to the three test domains.	3. Students in USI districts made faster progress than their counterparts in non-USI districts, particularly in the earlier grades. 4. There is a clear curvilinearity in the achievement trend. 5. The USI districts were about equally effective for all demographic groups. 6. Race/ethnic gaps declined in USI and non-USI districts at about the same rate.

Table 9.1 (Continued)			
	Meyer	**Bolt**	**Gamoran**
22. Findings (continued)		11. When the IRT scale was used, the growth between USI and non-USI was more parallel than when TLI was used, suggesting the susceptibility of the TLI metric to ceiling effects. 12. Based on correlations between common coefficients, there appeared to be slightly more variation in the growth patterns among schools in USI districts compared to non-USI districts.	
23. Limitations	1. It is highly problematic to limit models of achievement growth to students with multiple data points over time. 2. It is misleading to compare average test scores over time at a given grade level. 3. It is best to present estimates based on two estimating methods (OLS and EV1), rather than only a single method.	1. Non-USI districts were performing much better than USI-districts in 1994. 2. This, along with the same changes in TLI units, could represent non-comparable amounts of growth at different regions of the TLI scale. 3. TAAS TLI scores are likely subject to a ceiling effect. 4. Used a pseudo-linking process to equate test forms within domains. 5. Using the IRT scale means there is less precision in ability estimates at the high end of the latent ability metric.	1. Ignores retainees. 2. Non-linear growth may reflect a ceiling effect. 3. Non-comparability of TLI scores across grade levels limits interpretation of gain scores.

References

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage Publications.

Spiegelhalter, D. J., Best, N. G., & Gilks, W. R. (2000). *Win BUGS version 1.3 [computer program*]. Robinson Way, Cambridge CB2 2SR, UK: MRC Biostatistics Unit, Institute of Public Health.