6. TREND ANALYSIS USING HIERARCHICAL LINEAR MODELING AND ITEM RESPONSE THEORY (IRT)

Daniel Bolt

This report examines the influence of the Urban Systemic Initiative at the school (as opposed to student) level for grades 3 through 8 over the years 1994-1999. For each school, changes in mean TAAS scores at a given grade level over time are the basis for assessing USI effects. This type of analysis has both advantages and disadvantages that are related to what it does and does not control for when compared to an analysis that studies academic growth at the student level over time. By studying change in a specific grade level across years, we can assume that teacher impact is more effectively controlled for, since the same teachers are more likely to be teaching the same grades in successive years. Moreover, because the TAAS is equated within grade but not across grades from year to year, changes in TLI scores can be more confidently interpreted as changes with respect to a constant test performance metric.

However, using this strategy means losing control over other factors. The most important factor not controlled in this analysis is the turnover of students that comprise the grade levels in each school from year to year. To whatever degree the students being tested each year change in terms of their ability distribution, this change will be confounded with the improvements in the instruction being assessed.

This report is presented in two parts. The first examines changes in TAAS performance in USI and non-USI districts using the TLI (Texas Learning Index) as the outcome measure. The second part, which is concerned with studying change at the item level, uses an Item Response Theory (IRT) latent-ability estimate as its outcome measure.

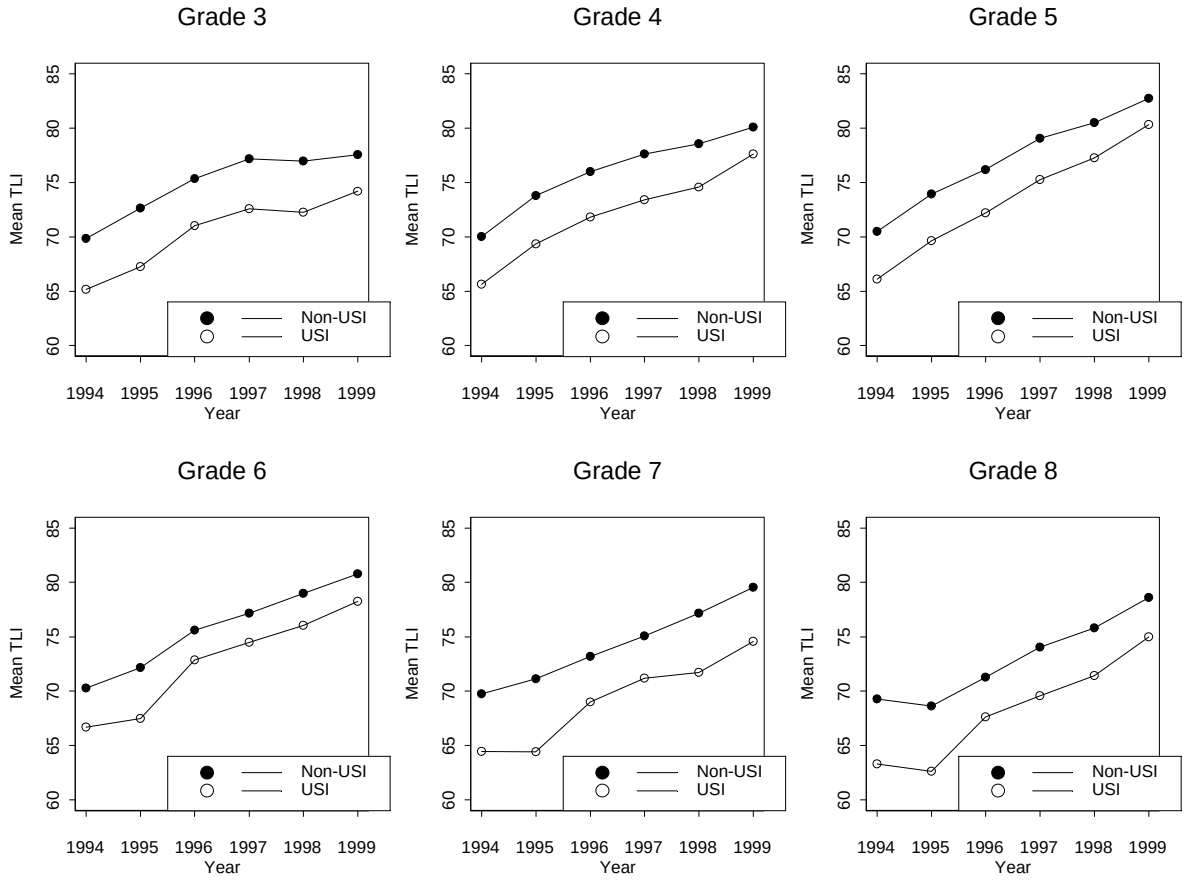
Modeling Growth in Mean TLI Scores at the School and District Level

Graphical Inspection of the Data (A)

Figure 6.1 illustrates changes in mean TLI scores within USI and non-USI districts for each grade from 1994-1999. Each point represents an average of the school means within USI and non-USI districts for a given grade and year. In both USI and non-USI districts, there is a fairly consistent increase in mean TLI scores across years within each grade. For grade 3, and to a lesser extent grade 4, the sharpest growth appears to occur early (1994-1996), and then slows in later years (1997-1999). For grades 5 through 8, growth appears to be more constant across years. In grades 6 through 8, there appears to be little growth from 1994-1995, and then a sharp increase in 1996. This change is most noticeable in the USI districts.

A more important issue relates to the consistently higher scores received across all grades and years within non-USI compared to USI districts. There appears to be a slight narrowing of the differences between USI and non-USI in all grades in the later years.

Figure 6.1. Plot of mean TLI scores comparing USI and non-USI districts, grades 3 through 8.



Figures 6.2 and 6.3 illustrate the average school means within each USI district (the coding used for the USI districts and the number of schools at each grade level included in the analysis are reported in Appendix A) for each year and grade. Figure 6.2 displays results for grades 3 through 5, and Figure 6.3 for grades 6 through 8. In these figures, unique changes within each district at a specific grade level can be monitored. For example, in USI District 7, there appears to be a sharp increase in scores in grades 3 through 5 in 1996. What also seems apparent from these graphs are differences in the TLI metric across grades. For example, across all districts, grades 5 and 6 appear to receive consistently higher scores than the other grades. While this may reflect disproportionately greater gains at these grade levels, an alternative interpretation may be that the TLI metric varies somewhat across grades.

Figures 6.4, 6.5, and 6.6 are similar to Figure 6.1, but break down average district school means according to race. The figures correspond to results for Whites, Hispanics, and Blacks, respectively. Among Whites, non-USI districts appear to perform slightly better than USI districts, although the margin of difference generally stays below 2 TLI points. Similar results exist for Hispanics, with the possible exception of grade 6, where

Figure 6.2. Plot of mean TLI scores in USI districts, grades 3 through 5.

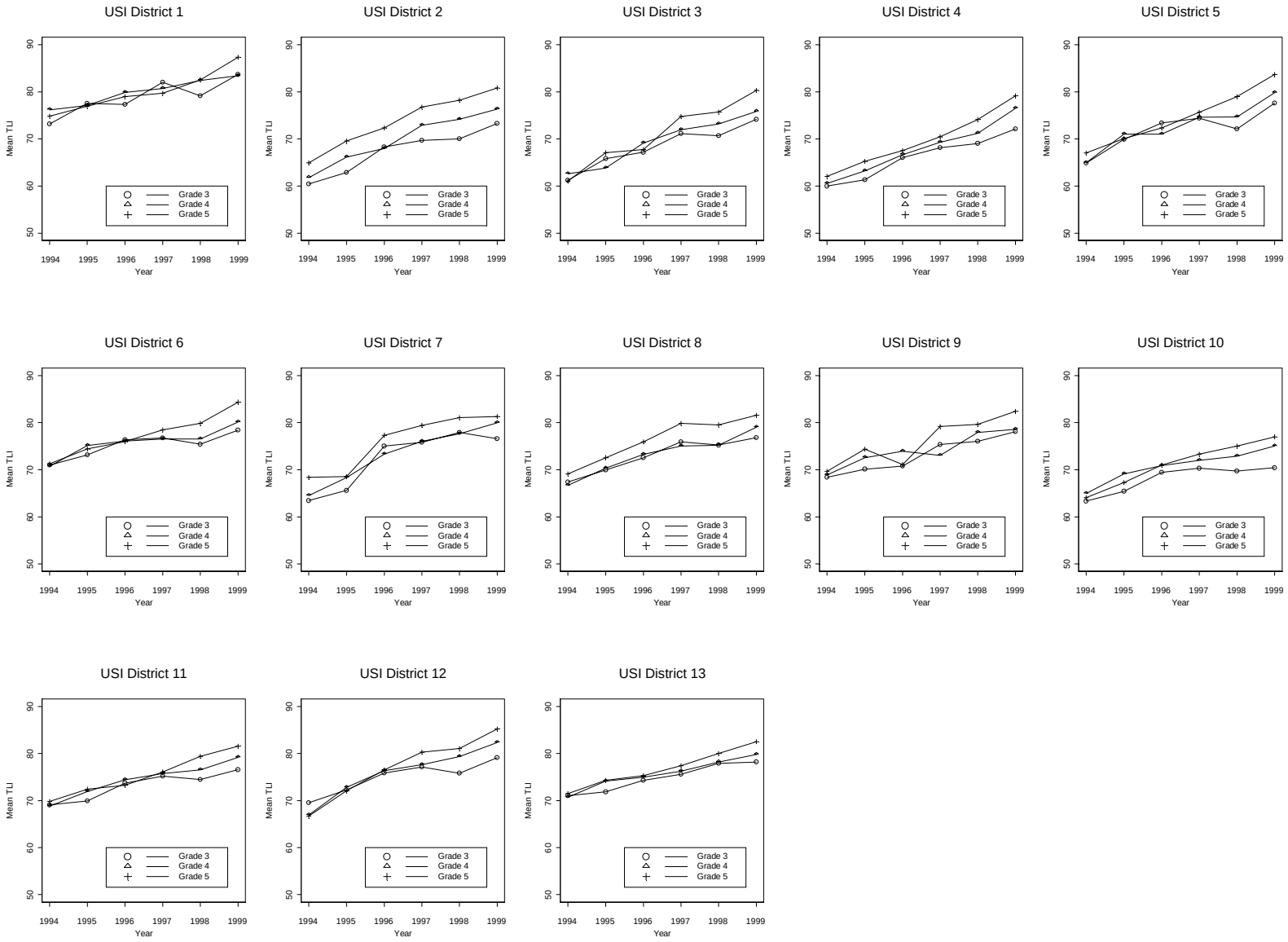


Figure 6.3. Plot of mean TLI scores in USI districts, grades 6 through 8.

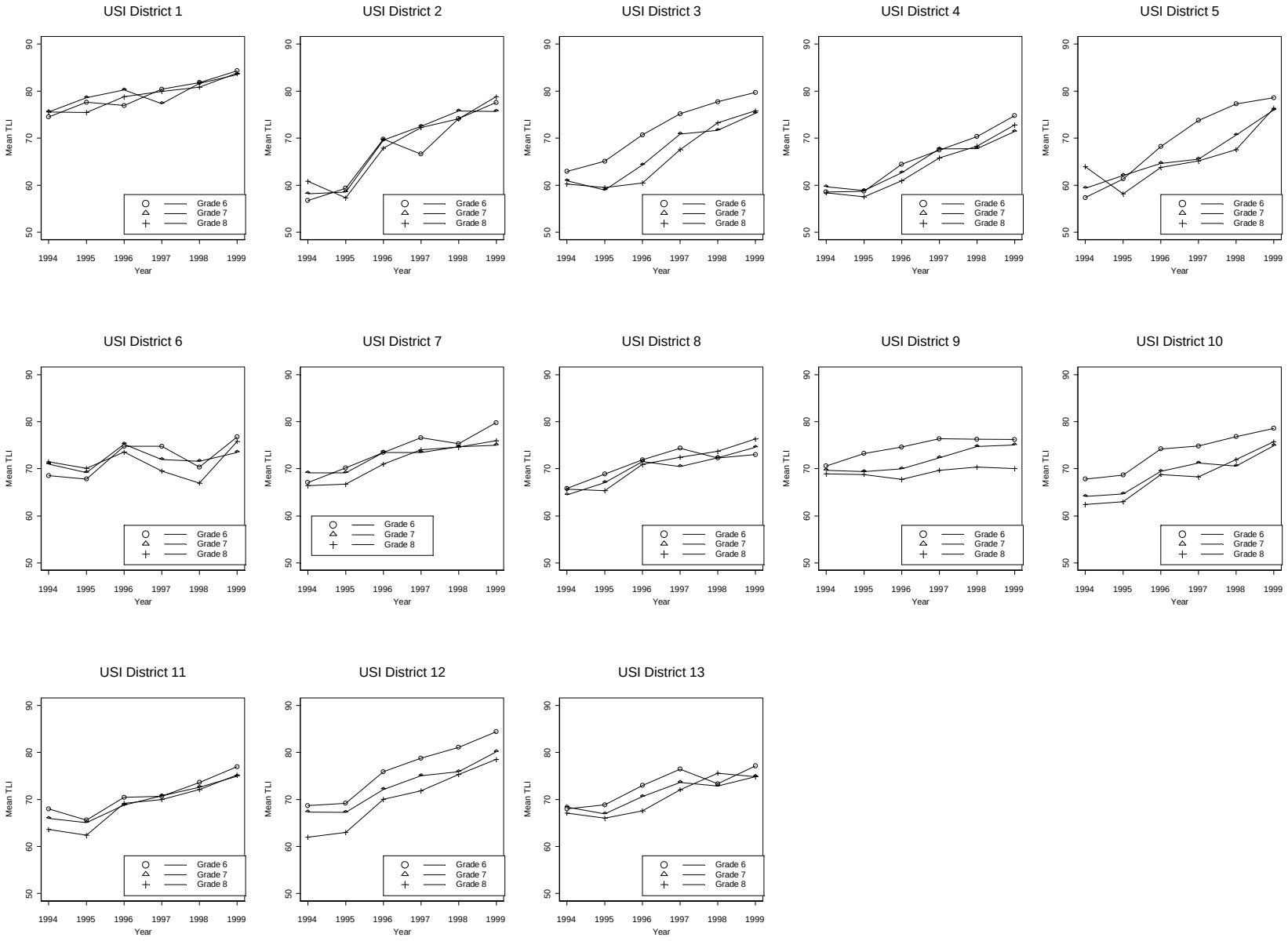


Figure 6.4. Plot of mean TLI scores comparing USI and non-USAI-districts, grades 3 through 8 White students.

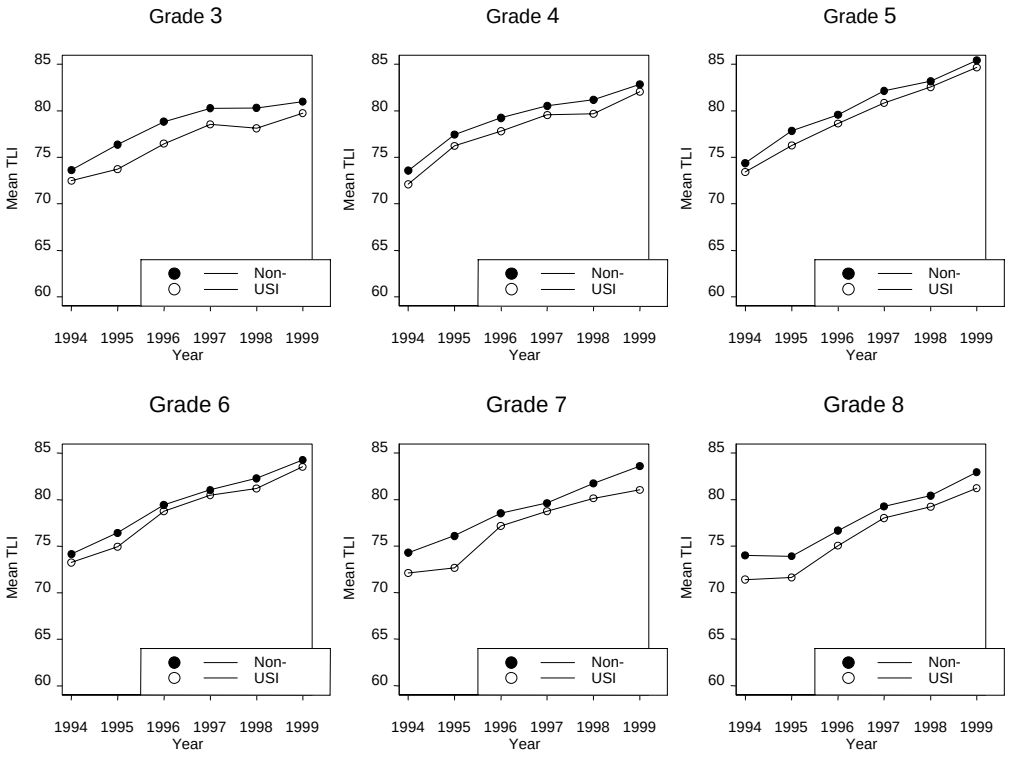


Figure 6.5. Plot of mean TLI scores comparing USI and non-USI districts, grades 3 through 8 Hispanic Students.

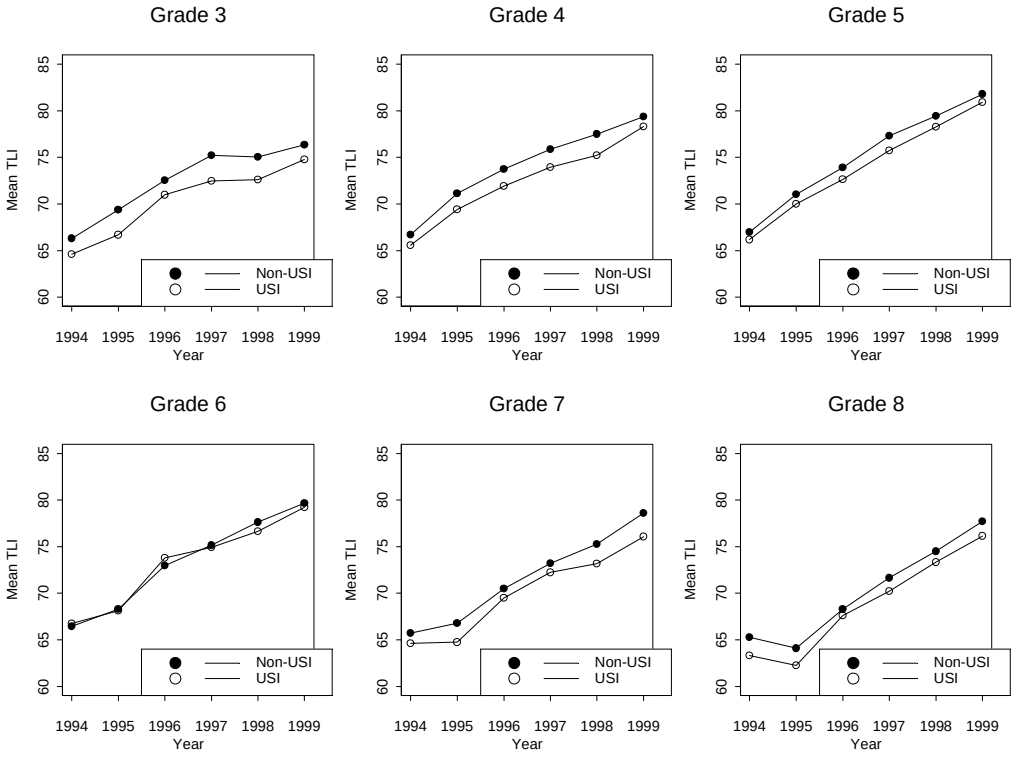
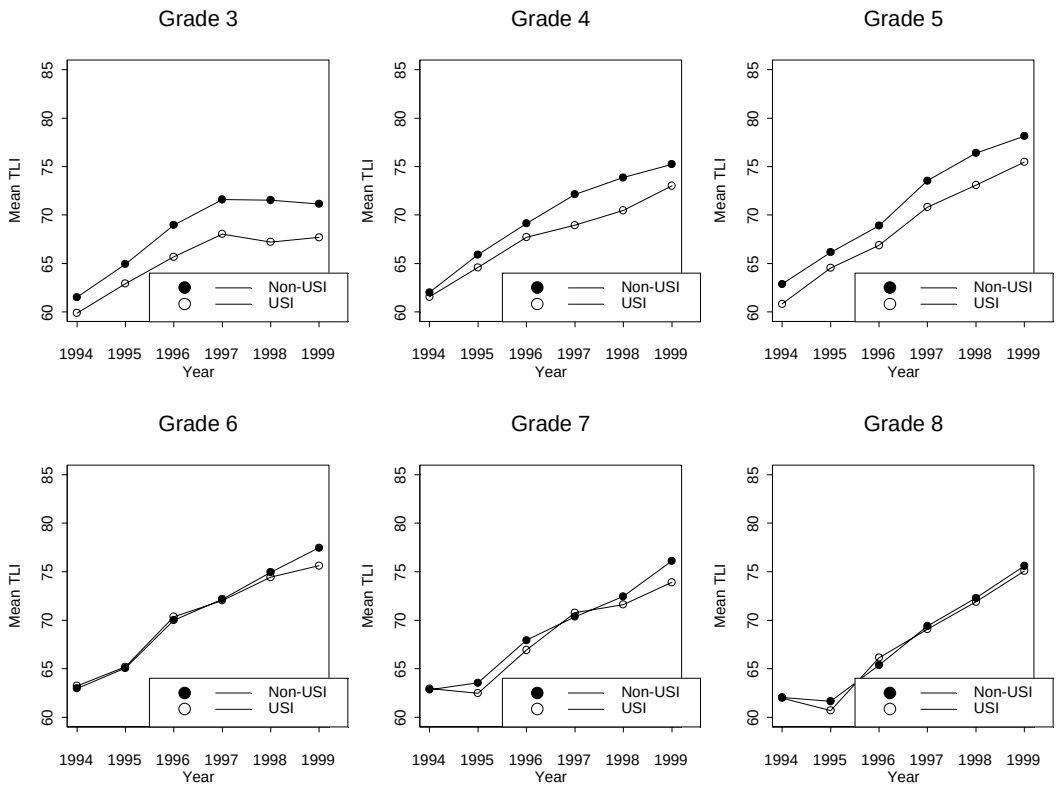


Figure 6.6. Plot of mean TLI scores comparing USI and non-USI districts, grades 3 through 8 Black students.



USI and Non-USI districts appear to be nearly identical. A more interesting result occurs for Blacks, where larger differences exist between USI and non-USI districts in grades 3 through 5, but then disappear in grades 6 through 8. In grades 3 and 4, the differences between USI and non-USI districts appear to be increasing over time.

Modeling USI/Non-USI School and District-Level Growth Based on the TLI (B)

Although the data themselves can speak to changes in TLI scores observed at the district level, a hierarchical model was also used to study growth.

The following hierarchical growth model was used to examine trends in TLI change within schools and districts at a fixed grade level from 1994-1999. For a given grade, assume $\overline{X}_{ijk}$ represents the mean TLI score in year i (0 = 1994, 1 = 1995, . . . , 5 = 1999) for school j of district k. A quadratic growth model is fit for each grade within each school:

$$\overline{X}_{ijk} = \alpha_{jk} + \beta_{1jk} i + \beta_{2jk} i^2 + e_{ijk}$$

In this model,

$$e_{ijk} \sim N(0, \sigma_{ijk}),$$

where

$$\sigma_{ijk} = s_{ijk}/\sqrt{n_{ijk}}$$

and s_{ijk} is the standard deviation of TLI scores in year *i* for school *j* of district *k*.

At level 2, school effects are modeled as:

$$\alpha_{jk} = \alpha_{.k} + u_{0jk}$$

$$\beta_{1jk} = \beta_{1.k} + u_{1jk}$$

$$\beta_{2jk} = \beta_{2.k} + u_{2jk}$$

where

$$u_{0jk} \sim N(0, \tau_{0k})$$

$$u_{1jk} \sim N(0, \tau_{1k})$$

$$u_{2jk} \sim N(0, \tau_{2k})$$

Lastly, districts are distinguished according to the presence or absence of USI participation. USI districts are regarded as a sample from one population of districts, districts not implementing USI (non-USI districts) as a sample from a distinct population of districts. Districts from the USI population are modeled as:

$$\alpha_{.k} = \alpha_{.USI} + r_{0k}$$

$$\beta_{1.k} = \beta_{1.USI} + r_{1k}$$

$$\beta_{2.k} = \beta_{2.USI} + r_{2k}$$

where

$$r_{0k} \sim N(0,\ \psi_{0,USI})$$

$$r_{1k} \sim N(0,\ \psi_{1,USI})$$

$$r_{2k} \sim N(0,\ \psi_{2,USI})$$

Districts from the non-USI population are modeled as:

$$\alpha_{.k} = \alpha_{.NonUSI} + r_{ok}$$

$$\beta_{1.k} = \beta_{1.NonUSI} + r_{1k}$$

$$\beta_{2.k} = \beta_{2.NonUSI} + r_{2k}$$

where

$$r_{0k} \sim N(0, \psi_{0,NonUSI})$$

$$r_{1k} \sim N(0, \psi_{1,NonUSI})$$

$$r_{2k} \sim N(0, \psi_{2,NonUSI})$$

This analysis is comparable to a V-known analysis in hierarchical linear modeling (Bryk & Raudenbush, 1992), in which the variances of the school means are assumed to be known (in this case, a function of the sample size and sample variance of TLI scores within each school). However, the HLM software package does not allow for implementation of V-known analysis in growth models with more than two levels. Thus, to fit the above model, the program WINBUGS (Spiegelhalter et al., 2000) was used. WINBUGS uses a Gibbs sampling procedure to estimate model parameters. In this model, there are several different types of parameters that can be estimated:

(1) school intercepts, school linear coefficients, school quadratic coefficients
(2) district intercepts, district linear coefficients, district quadratic coefficients
(3) district intercept variances, district linear coefficient variances, district quadratic coefficient variances
(4) a mean USI intercept, USI linear slope, USI quadratic slope
(5) a mean non-USI intercept, non-USI linear slope, non-USI quadratic slope
(6) USI/non-USI district intercept variances, linear coefficient variances, and quadratic coefficient variances.

The model is constructed so that the intercepts can be interpreted as the expected TLI score in 1994; the linear coefficients correspond to the increase in TLI scores per year, starting in 1994; the quadratic coefficients indicate the rate of change of this linear increase over years.

To estimate these parameters, prior distributions (weak) are initially specified for each of the parameters of the model, and initial values for each parameter are generated. Gibbs's sampling simulates a Markov chain in which values are sampled from the posterior distributions of each parameter given these priors, the data, and all other parameters in the model. After a suitable number of "burn-in" iterations, the subsequent values in the Markov chain can be regarded as a sample of values from the posterior distribution of that parameter. The estimate of the parameter can be computed as the mean of these sampled values.

The result of this analysis permits estimation of average growth curves at the school, district, and district type (USI/non-USI) level, as well as quantification of the variance of growth curves within USI and non-USI districts. No additional school or district-level predictors were included in this model. However, the model could be easily adapted to incorporate them.

Results

Figure 6.7 illustrates the estimated average quadratic growth curves for the non-USI and USI districts at each grade level. At each grade level, the difference between intercept parameter estimates for the USI and non-USI districts was statistically significant ($p < .05$), but differences in the linear and quadratic coefficients were not. (Appendix B reports the number of USI and non-USI districts and schools included in the analysis at each grade level.) The failure to find statistically significant differences can be attributed to the relatively small number of USI districts (13) included in the analysis and the allowance for district-level effects in the model. Figure 6.8 illustrates the estimated average growth curves for the individual USI districts at each grade level.

Comparison of USI and Matched Non-USI Schools Using the TLI (C)

A limitation of the above analysis can be seen from the different intercepts for the USI/non-USI districts. In general, non-USI districts contain schools that were performing much better than USI districts in 1994. This makes it difficult to compare the amounts of growth occurring across districts, since there is no guarantee that changes in TLI units represent comparable amounts of growth when they occur in different regions of the TLI scale. In addition, TAAS (and thus TLI) scores are in all likelihood subject to a ceiling effect, with high achieving schools being unable to demonstrate sustained growth once they reach a high level of TAAS achievement.

Figure 6.7. HLM-estimated growth curves for mean TLI scores comparing USI and non-USI districts, grades 3 through 8.

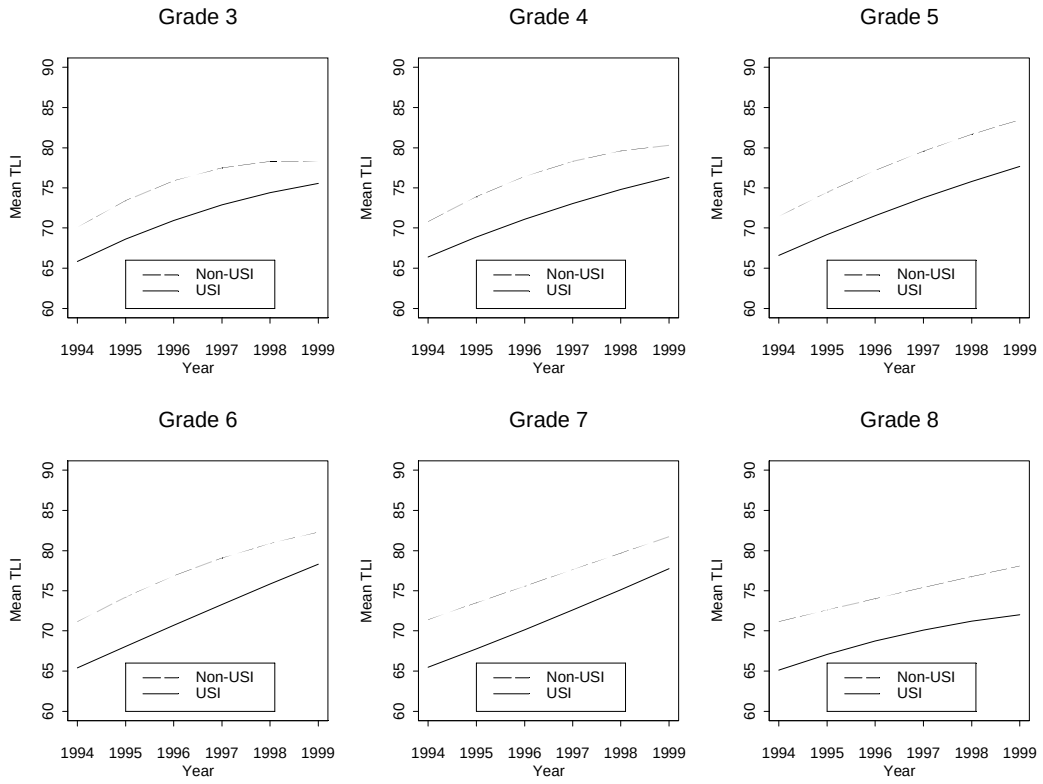
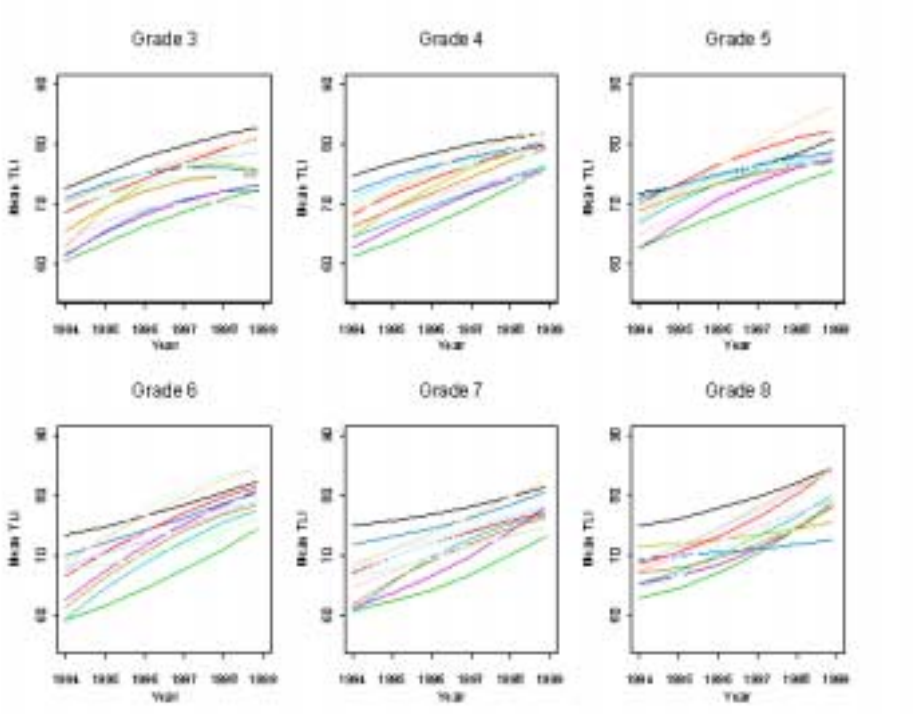


Figure 6.8. HLM-estimated growth curves for mean TLI scores for USI districts, grades 3 through 8.



In an attempt to remedy this problem, a subset of the schools from non-USI districts were randomly selected that were more comparable in terms of average TLI scores (the average being across all years 1994-1999) to the schools from USI districts. This sampling procedure was performed as follows. For each grade, a mean TLI score was computed across years (1994 to 1999) within each school. The distribution of these mean TLI scores was then computed across all USI schools. Each non-USI school was then assigned a probability of being selected for the matched sample based on its mean TLI score. Specifically, the probability of selection was based on the empirical proportion of USI schools having approximately that mean TLI score. The result was a sample of non-USI schools having a distribution of TLI scores that, on average, more closely resembled the scores of USI schools. (Appendix B reports the number of non-USI districts and schools included in the matched samples selected for each grade.) All of the USI districts and schools were included in this analysis.

Results

Figure 6.9 illustrates results for the USI, matched non-USI, and overall non-USI results for each grade. For the most part, the USI districts appear to behave similarly to the matched non-USI districts. Grade 7 appears to be slightly different. In this case, the USI districts show somewhat greater gains relative to the non-USI matched districts. Whether this is an artifact of the particular random sample of non-USI districts selected or demonstrates an actual improvement for the USI districts relative to non-USI districts may require further investigation. Unfortunately, similar results were not detected in any of the other grades. It should be noted that in this analysis, none of the estimates of the growth curves were found to be statistically different when comparing the USI and matched non-USI districts, although this might again be attributed to the relatively small number of USI districts included in the analysis.

Comparison of USI/Non-USI by Race (D)

A hierarchical model like that considered in section (B) can also be fit for distinct races. In this analysis, we distinguish three racial groups: (1) Whites, (2) Hispanics, and (3) Blacks. Assume now that $\overline{X}_{ijk}$ represents the mean TLI score in year i for school j of district k for a given grade and race. The model fit is the same as in (B), with s_{ijk} and n_{ijk} now representing the standard deviations and sample sizes for a specific race in a specific grade. As in (B), this analysis is concerned in part with comparing USI/non-USI districts with mean gains for students of a particular race. At the same time, we examine changes in the difference between races within USI and non-USI districts.

Results

Table 6.1 provides the estimated quadratic growth curves for the non-USI and USI districts with respect to each ethnic category. For each grade, the table indicates the average (mean) estimated intercept, linear coefficient, and quadratic coefficient for the USI and non-USI districts according to race. Essentially, these represent the fit of quadratic functions to the data displayed in Figures 6.4, 6.5, and 6.6 using a hierarchical

Figure 6.9. HLM-estimated growth curves for mean TLI scores comparing USI, non-USI districts and matched non-USI districts, grades 3 through 8.

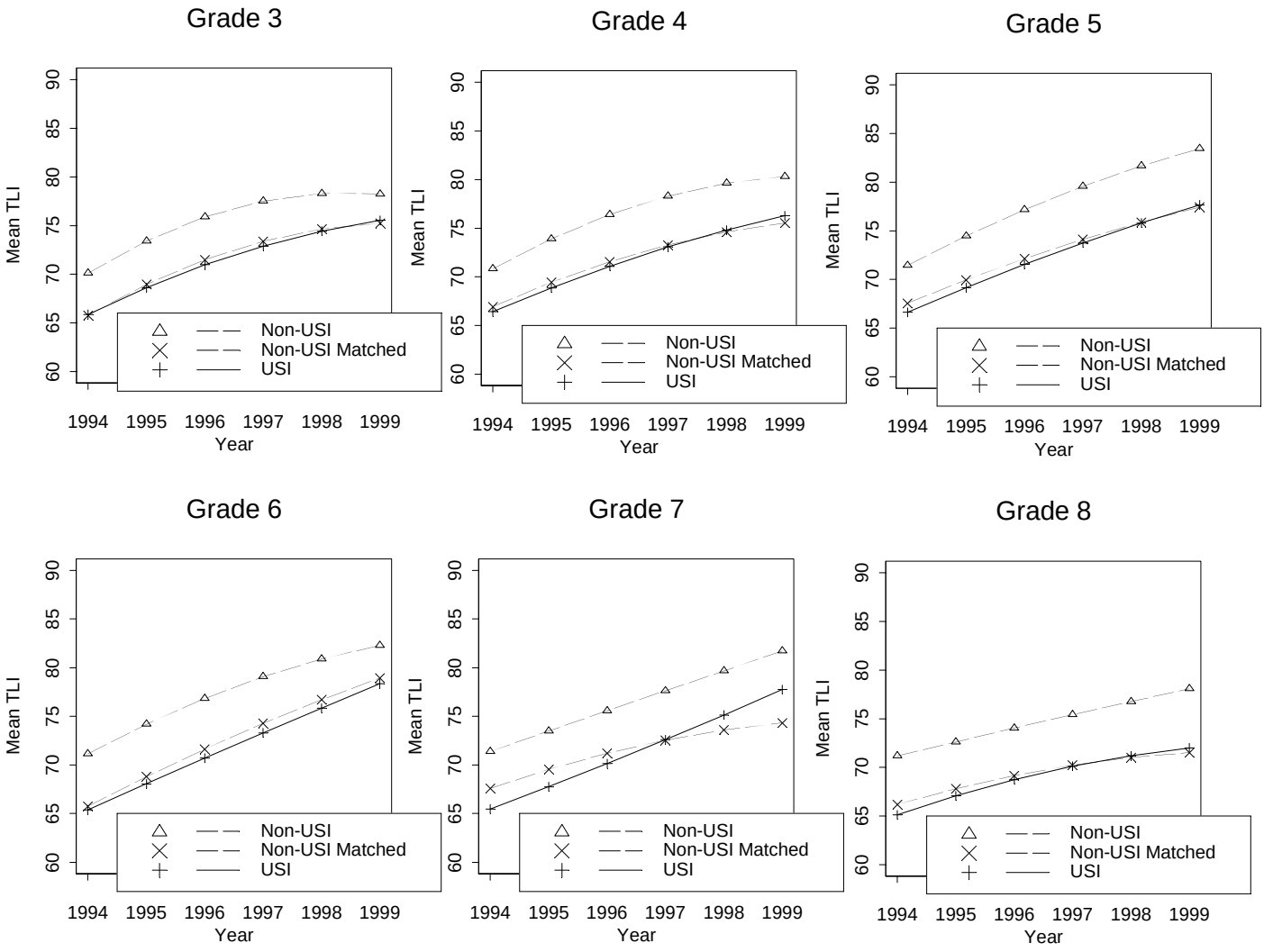


Table 6.1

Average Quadratic Growth Curves Within USI and Non-USI Districts, Estimated Separately for Whites, Hispanics, and Blacks, 1994-1999

	White				Hispanic				Black			
	Mean		Variance		Mean		Variance		Mean		Variance	
	USI	Non-USI	USI	Non-USI	USI	Non-USI	USI	Non-USI	USI	Non-USI	USI	Non-USI
						Grade 3						
Int	71.56	73.00	13.76	10.50	64.73	65.72	15.74	23.06	59.87	61.45	23.10	13.50
Lin	2.63	3.62	3.75	4.10	3.07	4.03	1.74	2.50	3.29	5.10	4.08	5.03
Quad	-.14	-.46	.30	.45	-.24	-.43	.31	.27	-.44	-.50	.37	.50
						Grade 4						
Int	72.49	73.51	14.55	11.57	65.93	66.85	15.00	19.65	60.80	61.22	19.88	14.02
Lin	2.43	3.60	3.55	4.28	2.91	3.09	1.97	3.03	2.99	2.57	3.02	1.96
Quad	-.09	-.40	.21	.55	-.09	-.11	.09	.27	-.16	-.03	.05	.15
						Grade 5						
Int	73.40	74.19	15.45	10.95	66.38	66.52	16.10	20.05	60.51	62.78	19.89	14.25
Lin	2.32	3.31	2.55	3.41	3.51	3.47	4.90	2.08	2.37	2.41	1.77	2.72
Quad	-.07	-.25	.11	.31	-.04	-.08	.04	.10	.00	-.02	.04	.03
						Grade 6						
Int	71.92	73.84	12.57	14.11	66.38	66.44	19.20	17.10	63.32	63.25	14.32	14.66
Lin	2.21	3.46	1.44	2.34	3.52	3.61	2.66	2.33	2.30	2.40	1.98	2.04
Quad	-.04	-.33	.09	.21	-.18	-.17	.13	.22	-.04	-.03	.04	.09
						Grade 7						
Int	71.87	74.16	11.31	13.55	64.89	65.44	16.58	16.90	62.79	62.81	15.73	14.50
Lin	1.41	2.48	1.55	1.75	2.74	2.90	2.38	2.11	2.05	2.16	1.58	2.09
Quad	.24	-.16	.20	.09	-.02	-.01	.03	.05	-.10	.03	.07	.02
						Grade 8						
Int	69.97	73.70	10.84	13.70	63.03	64.20	12.41	15.88	62.39	62.38	16.10	14.57
Lin	2.13	1.60	2.40	1.65	1.78	1.88	1.80	1.57	2.23	2.28	3.19	1.50
Quad	-.01	.03	.11	.05	.11	.07	.11	.12	.04	.06	.03	.06

growth model. In Table 6.1, “Int” corresponds to the average intercept, which can be interpreted as the expected TLI score in 1994; “Lin” corresponds to the linear increase in TLI scores per year starting in 1994; “Quad” to the average rate of change of this linear increase over years.

Also reported in Table 6.1 under the “Variance” columns are the estimated variances of each of these quadratic growth components within each district. These provide some idea of the amount of variation in growth that occurs among USI and non-USI districts, respectively. In other words, they indicate the degree to which growth curves of individual districts depart from the average displayed in Figures 6.4, 6.5, and 6.6. Although the mean differences observed across district type are quite consistent across grades, years, and races (with non-USI districts scoring slightly higher), the differences between the intercept, linear, and quadratic coefficients across USI and non-USI districts were not found to be statistically significant. This can again be attributed to the use of only 13 USI districts and the model's allowance for district effects.

Measuring Growth at the TAAS Domain/Objective Level Using Item Response Theory (IRT)

Due to the availability of student-level item response data in this data set, it becomes possible to trace growth at an even finer level than overall test scores.

Dimensionality Analysis of TAAS Test Forms (A)

To assess whether it might be possible to trace growth with respect to specific item types, a dimensionality analysis of the test forms at each grade level was conducted. A random sample of 1,000 examinees at each grade (3-8) was selected in 1994. The item- response vectors for these examinees were then analyzed using full information item-factor analysis (Bock, Gibbons, & Muraki, 1988), as implemented in the software package TESTFACT. Table 6.2 reports the eigenvalues for the 1994 form in each grade for the first eight factors:

Table 6.2
Eigenvalues for Each Grade for Eight Factors

Factor	Grade 3 Eigenvalue	Grade 4 Eigenvalue	Grade 5 Eigenvalue	Grade 6 Eigenvalue	Grade 7 Eigenvalue	Grade 8 Eigenvalue
1	14.58	16.00	16.80	19.18	19.35	20.34
2	2.04	2.41	3.19	2.52	2.35	2.74
3	1.64	1.86	2.36	2.35	2.06	2.36
4	1.42	1.75	2.04	1.77	1.92	1.72
5	1.37	1.48	1.64	1.57	1.53	1.54
6	1.28	1.42	1.41	1.45	1.36	1.35
7	1.20	1.33	1.38	1.31	1.28	1.27
8	1.17	1.25	1.25	1.30	1.21	1.23

These eigenvalues represent the variances of primary principal components underlying each data set. Typically, factors having eigenvalues greater than 1 are of primary importance, as they represent factors accounting for greater variance than exists in a single item. Although for each test there appears to be a single dominant dimension (i.e., “mathematics ability”), as implied by the relatively large first eigenvalue, the magnitudes of the second and higher eigenvalues may be suggestive of some multidimensionality. Thus, it is conceivable that meaningful information may reside in TAAS subscores as well as in the overall TAAS score.

Test specifications for the forms in each grade organize items according to three domains: (1) Concepts, (2) Operations, and (3) Problem Solving. Within each domain, the items are more specifically classified according to four or five objectives. In the Concepts domain, the objectives include:

1. understanding of number concepts
2. understanding of mathematical relations, functions, and other algebraic concepts
3. an understanding of geometric properties and relationships

4. understanding of measurement concepts using metric and customary units
5. an understanding of probability and statistics

The Operations domain consists of four objectives:

6. using the operation of addition to solve problems
7. using the operation of subtraction to solve problems
8. using the operation of multiplication to solve problems
9. using the operation of division to solve problems

The Problem-Solving domain consists of four objectives:

10. estimating solutions to a problem situation
11. determining solution strategies and analyzing or solving problems
12. expressing or solving problems using mathematical representation
13. evaluating the reasonableness of a solution to a problem situation

Figures 6.10(a) and (b) provide examples of items from each domain. Example items measuring two distinct objectives within each domain are displayed.

One goal of the dimensionality analysis was to attempt to align these domains and objectives with factors that could be statistically detected in the data. Table 6.3 illustrates item factor analysis results for the grade 4, 1994 form, which consisted of 50 items. For each item, Table 6.2 reports the Domain and Objective to be measured by the item, as well as the pattern loadings obtained for a three-factor maximum likelihood solution, followed by a Promax rotation. The correlations between factors were .66, .69, and .66, respectively, for factor pairs 1 and 2, 1 and 3, and 2 and 3. Thus the factors, although highly correlated, do appear to be somewhat distinct. As seems apparent from the pattern loadings in Table 6.3, Factor 1 is most related to Domain 3 (Problem-solving), Factor 2 to Domain 2 (Procedures), and Factor 3 to Domain 1 (Concepts).

Comparable results were obtained for forms in other grades and seemed to suggest potential merit in tracing progress with respect to the three domains separately.

Description of the IRT Equating Procedure (B)

To monitor changes at the levels of individual domains or objectives in the TAAS, it is necessary to perform an equating of TAAS scores at the Domain or Objective level. As described in the TAAS manual, Item Response Theory is used to perform an equating of total TAAS scores. Specifically, the Rasch model is fit to each test, and IRT

Figure 6.10(a). Example TAAS items for two objectives within each of three measurement domains.

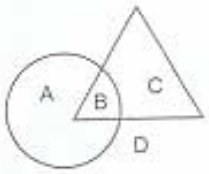
Domain 1, Objective 1

1 The highest mountain in Texas is eight thousand, seven hundred fifty-one feet high. How is this number written? Mark your answer.

- 8,751
- 87,051
- 800,751
- 8,700,051

Domain 1, Objective 3

3 Which letter is inside the triangle and outside the circle? Mark your answer.



- A
- B
- C
- D

Domain 2, Objective 6

38 Kayla had 45¢. Lisa gave her 15¢. How much money did Kayla have then? Mark your answer.

- 30¢
- 50¢
- 60¢
- 70¢

Figure 6.10(b). Example TAAS items for two objectives within each of three measurement domains.

Domain 2, Objective 7

34 Orville and Wilbur Wright first flew an airplane in 1903. In 1971 the first spacecraft flew around the planet Mars. How many years after the Wright brothers' invention did this first Mars flight take place? Mark your answer.

- ⬭ 62 years
- ⬭ 68 years
- ⬭ 72 years
- ⬭ 74 years

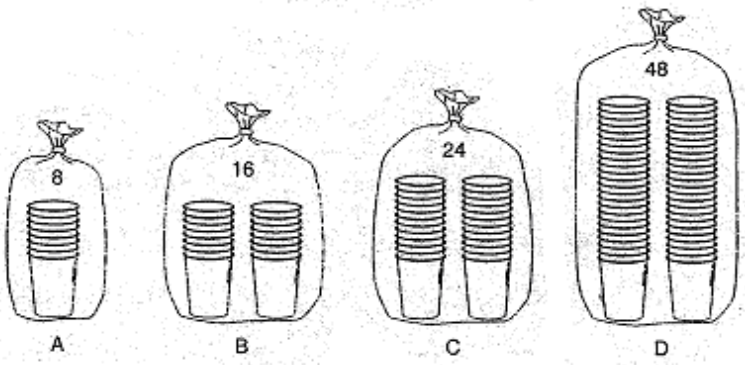
Domain 3, Objective 11

22 In the human body there are 16 bones in the wrists, 28 bones in the fingers, and 10 bones in the hands. Which shows how many bones are in the fingers and hands? Mark your answer.

- ⬭ 16 + 10
- ⬭ 28 + 16
- ⬭ 16 + 18
- ⬭ 28 + 10

Domain 3, Objective 13

30 Mrs. Elliot needs 55 cups for a party. Which 2 packages should she buy to have the least number of cups left over? Mark your answer.



- ⬭ A and D
- ⬭ B and C
- ⬭ B and D
- ⬭ C and D

Table 6.3

Pattern Matrix for Three-Factor Solution (following Promax Rotation) for the Grade 4, 1994 TAAS form (From a Random Sample of 5000 Students)

			Pattern Loading		
Item	Domain	Objective	Factor 1	Factor 2	Factor 3
1	1	3	0.217	0.415	-0.013
2	1	2	0.035	0.784	-0.202
3	1	1	0.028	0.577	0.057
4	1	2	0.038	0.362	0.161
5	1	1	0.106	0.629	-0.143
6	1	4	0.140	0.452	-0.142
7	1	3	0.211	0.132	-0.047
8	1	1	-0.014	0.478	0.219
9	1	2	0.039	0.219	0.347
10	1	5	0.637	0.039	-0.041
11	1	3	0.407	0.251	-0.090
12	1	4	0.476	0.242	-0.085
13	1	5	0.232	0.303	0.023
14	1	5	0.110	0.315	0.017
15	1	4	0.375	0.262	0.011
16	1	1	0.163	0.327	0.038
17	1	4	0.594	0.030	-0.088
18	1	3	0.508	0.206	-0.162
19	1	2	0.528	0.091	0.046
20	1	5	0.336	0.445	-0.108
21	3	10	0.098	0.265	0.001
22	3	12	0.459	0.192	0.021
23	3	10	0.507	0.055	0.162
24	3	11	0.632	-0.087	0.023
25	3	11	0.603	-0.046	0.050
26	3	11	0.454	0.168	0.045
27	3	11	0.653	-0.074	0.051
28	3	10	0.639	-0.016	0.084
29	3	10	0.656	-0.116	0.038
30	3	12	0.358	0.021	0.009
31	3	10	0.612	-0.188	0.097
32	3	10	0.611	-0.126	0.172
33	3	12	0.696	-0.112	0.010
34	3	12	0.486	0.115	0.017
35	2	8	-0.234	0.580	0.205
36	2	6	0.214	0.343	0.153
37	2	9	0.092	-0.069	0.601
38	2	7	-0.191	0.475	0.369
39	2	8	-0.105	0.530	0.165
40	2	7	0.266	-0.025	0.599
41	2	7	0.381	0.165	0.230
42	2	9	0.084	0.082	0.658
43	2	8	-0.071	0.070	0.778
44	2	9	0.071	-0.073	0.623
45	2	8	-0.060	0.052	0.724
46	2	6	-0.068	0.450	0.045
47	2	6	0.322	0.056	0.500
48	2	7	-0.279	0.513	0.276
49	2	9	0.401	0.102	0.299
50	2	6	0.068	-0.047	0.655

true-score equating is used to determine matching raw scores across different forms of the TAAS across years. The Rasch model characterizes the probability of correct response to an item as a function of a latent examinee ability, denoted θ, and the difficulty of the item, denoted b. Specifically, if $P(X=1)$ denotes the probability of a correct response, the Rasch model can be written as:

$$P(X=1|\theta) = \frac{\exp(\theta - b_i)}{1+\exp(\theta - b_i)}$$

IRT models are attractive for test equating purposes because of their invariance properties. Specifically, assuming the Rasch model fits, the latent ability metrics defined by any two distinct calibrations of the model (say to two different forms of the TAAS) will be linearly related. This linear relationship can be determined, provided there exists a common subset of items on both forms, or a common subset of examinees given both forms. Either strategy (common items or examinees) is said to provide a *link* between the two latent ability metrics. A frequent approach to linking two test forms administered in subsequent years is to administer to a small number of the examinees in year 1 a certain subset of the items to be administered in year 2. In this way, a pre-equating of the test forms can be performed; this appears to be the linking strategy used each year with the TAAS.

Once the ability metrics are linked, test characteristic curves (TCCs) are defined for each form according to the item difficulty parameter estimates of the items that comprise each test. The TCCs are functions of ability (like the item-response function) that indicate the expected test score conditional upon ability level. The TCCs are defined as a simple sum across items of the probabilities of correct response (according to the Rasch model) at each ability level θ. IRT true-score equating then matches raw scores across forms corresponding to the same θ levels.

As a final step in score reporting, equated TAAS scores are converted to TLI scores, which are scaled such that a score of 70 is taken to represent the mean proficiency for that grade level in 1994. Theoretically, a similar type of equating can be performed for subtests of items, in addition to the overall test. Unfortunately, the current analysis was limited by the lack of pre-equating results that had been used to perform the original TAAS equating. Thus, the critical information (overlapping items or examinees) needed to link the ability metrics was absent.

However, both student-level item responses and the reported student-level TLI scores (already equated) were available, thus making it possible to perform a type of "pseudo-linking" of the ability metrics for any two forms (from different years) within the same grade. This procedure is outlined next (in actual practice, where pre-equating data are available, this procedure would not be used). First, the Rasch model was fit to each test form (1994-1999) administered in a common grade. Based on this Rasch calibration, an ability estimate corresponding to each number-correct score can be estimated. (A unique feature of the Rasch model is that the number-correct score is a sufficient statistic to denote ability; thus, all respondents achieving the same number correct receive the same ability estimate, regardless of the specific items answered correct/incorrect.) Because each number correct score is also aligned with a specific TLI score, a one-to-one correspondence also exists between the TLI scores

(already equated using the pre-equating data) and ability estimates on each form. Note that each TLI score will be associated with a different ability estimate for each form because the ability metrics have not been linked. Because TLI scores are assumed to have the same meaning across forms, however, they can be used to determine how one ability metric from one calibrated form should be linearly transformed so as to be equivalent to the latent ability metric from a second calibrated form. This linking of ability metrics is achieved by identifying the best fitting line predicting the θ estimates of one form from the θ estimates of the second form having the same TLI score.

Once the ability metrics are linked, ability estimates based on any subset of items (e.g., items based on a common domain or objective) can be computed across years to evaluate growth. For the current analysis, the ability metrics for all forms from 1995-1999 were linked back to the 1994 ability metric in each grade. In the subsequent analyses, growth is assessed with respect to changes in these ability estimates as opposed to changes in raw test or subtest scores.

Contrasting the TLI and IRT Latent Ability Metrics for Analyzing Growth (C)

As currently constructed, the TAAS is designed to distinguish students at lower proficiency levels (i.e., to determine who has achieved a minimal level of mathematics proficiency). As a result, the typical item is answered correctly by most examinees; thus, many students achieve perfect or near-perfect scores on the test. For purposes of assessing growth, this is limiting in that progress among high-achieving students cannot be assessed due to a ceiling effect. A potential advantage of the IRT latent ability metric in this respect is that it is continuous; therefore, theoretically, it is not subject to a ceiling effect such as that which exists for a test score metric.

The apparent removal of ceiling effects when using the IRT ability metric does come at a price, however. Because most of the test items are relatively easy, there is much less precision in ability estimates at the high end of the latent ability metric. Thus, although the ability estimates are unbounded, there is greater error in the ability estimates of students that are of high ability.

Coincidence of Growth Across Objectives/Domains Within Grade (D)

We considered the degree to which growth with respect to each domain/objective coincides with growth in other domains/objectives within each grade. To the degree that there are differences, we may be encouraged to assess effects of the Systemic Initiative at the domain or objective score level as opposed to the total test-score level. It is conceivable that the effects of the Systemic Initiative may be apparent in improving scores for certain types of items more than others.

Although not reported in detail here, additional analyses were performed to examine whether there were meaningful differences in growth patterns for distinct objectives and domains. (Recall that the dimensionality analysis for a single test examined in isolation from the others indicated some distinction between domains, but not objectives within domains.) These analyses largely involved inspecting correlations between growth (the difference between scores in successive years) for different objectives and domains. These correlations related to growth

within a constant grade level across years (i.e., involving different cohorts of students). Separate models were fit for each grade. The distinction between the USI/non-USI status of the districts to which schools belonged was not made in this analysis, since the primary goal was to determine the level of specificity at which change should be monitored. These results, which were consistent with the dimensionality analyses, suggested some differentiation between item domains, but not the different objectives within domains.

Graphical Inspection of the Data (E)

Figures 6.11(a) and (b) illustrate changes with respect to each domain in each grade across the years 1994-1999 using the IRT latent-ability metric. Because of incomplete information regarding the composition of the grade 3 TAAS forms, results are reported only for grades 4-8. Each point thus corresponds to an average domain score for a single domain for a given grade and year in either the USI or non-USI districts. Relative to what had been observed with TLI scores, the differences between USI and non-USI over time appear to be even closer to parallel. This may be attributed in part to the use of the latent IRT metric in tracking change (which is less subject to ceiling effects). There are some awkward jumps and dips in some of these graphs (e.g., grade 6, 1998; grade 7, 1995), which may be due to equating problems, especially since the "pseudo-linking" procedure used was based on various approximations. (The problems with equating should affect both the USI and non-USI growth curves, since a common equating was performed across all schools.) These equating problems would not be expected to occur if the equating performed used the linking procedure that was the basis for determining TLI scores.

Modeling USI/Non-USI Differences

A multivariate multigroup latent-growth model was used to examine differences across USI and non-USI schools with respect to growth in each of the three domains. In this model, each school (whether from a USI or non-USI district) is assumed to follow a quadratic growth curve for each domain. However, growth is monitored simultaneously with respect to each domain, so that growth patterns across domains can also be correlated.

For this analysis, a latent-growth curve model was fit to the data for each grade. Figure 6.12 provides an illustration of a path diagram describing this model. Each of the boxes represents an observable variable—in this case, a school mean with respect to a particular domain (d1, d2, d3 for Domains 1, 2 and 3, respectively) in a particular year

Figure 6.11(a). Plot of mean IRT ability estimates comparing USI and non-USI districts over three measurement domains, grades 4 through 9.

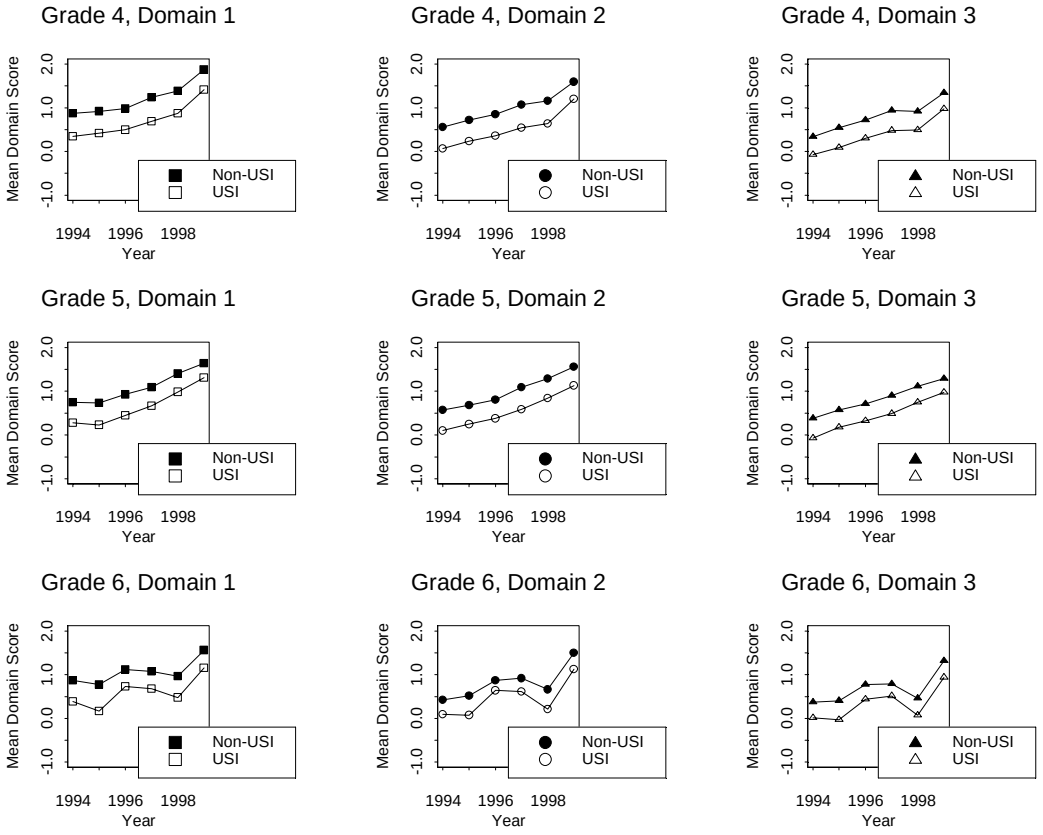


Figure 6.11(b). Plot of mean IRT ability estimates comparing USI and non-USI districts over three measurement domains, grades 4 through 9.

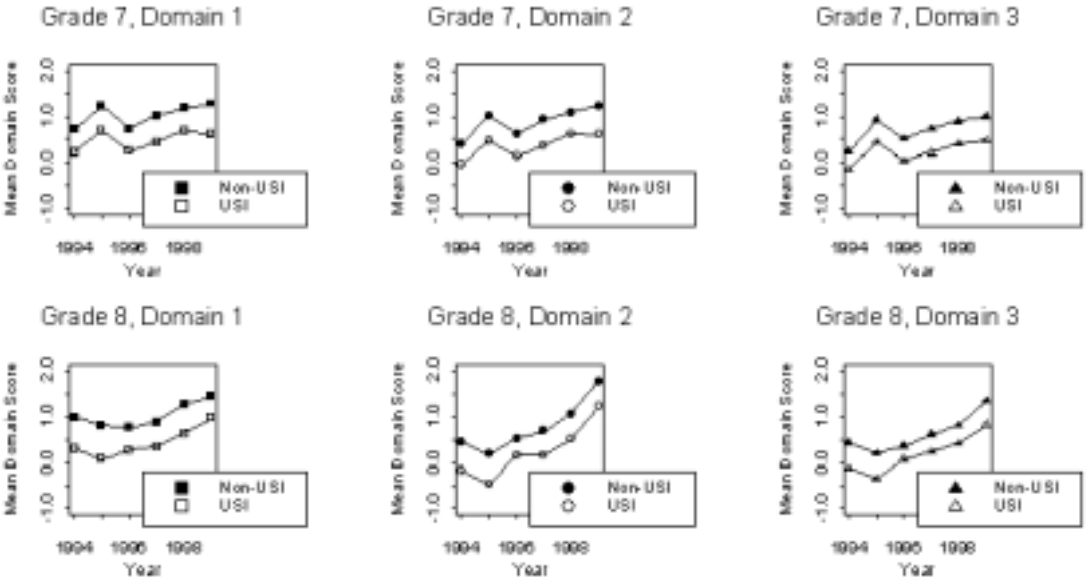
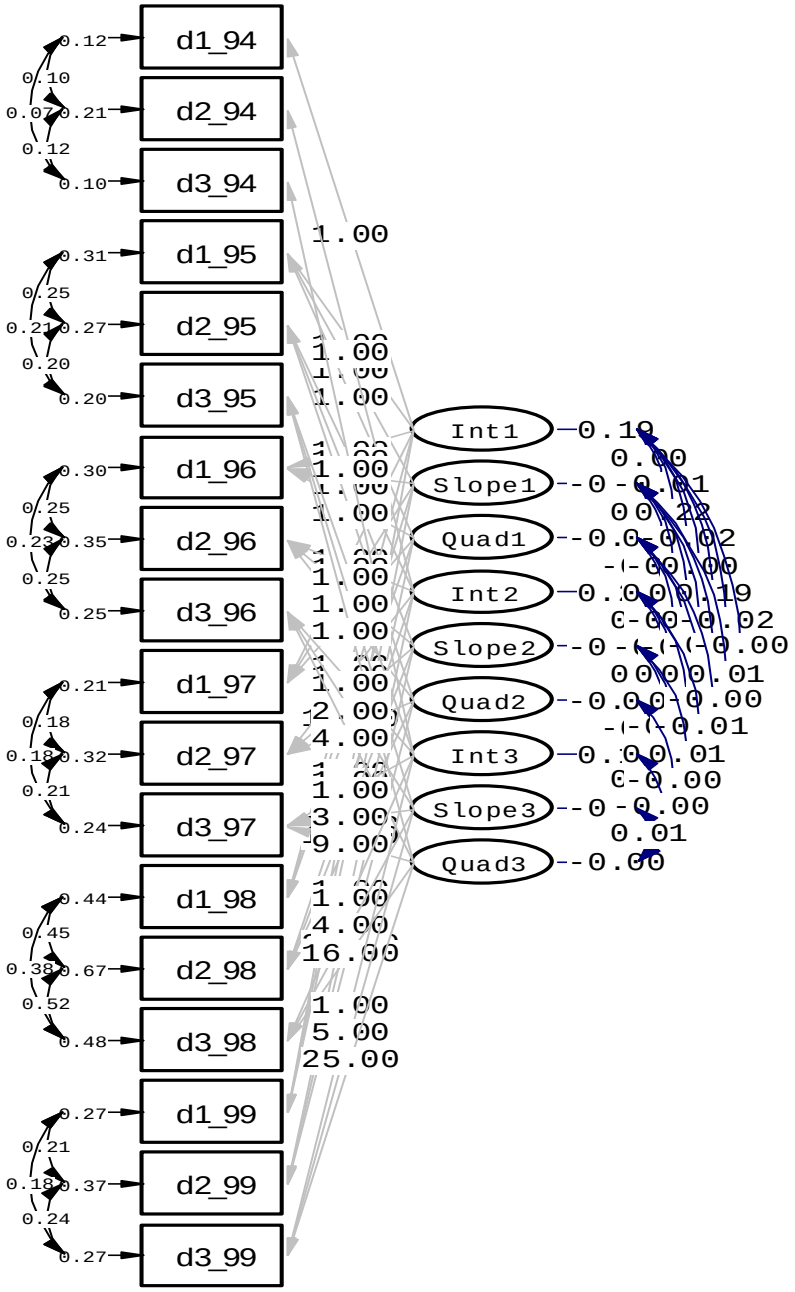


Figure 6.12. Conceptual path diagram illustrating multivariate growth model fit using the LISREL software.



(1994-99). Each of the ovals represents a different coefficient of the quadratic growth curve for each domain (Int1 = the intercept with respect to domain 1; Slope1 = the linear coefficient for Domain 1; Quad1 = the quadratic coefficient for Domain 1, etc.). In a random coefficients model, each of these coefficients has a mean and a variance. The mean represents the average of the coefficients across schools, the variance is the variance of these coefficients across schools. Since the coefficients are random, there also exists a correlation between each pair of coefficients that is estimated in the model. The numbers between the ovals and boxes represent constrained factor loadings needed to define the ovals as intercepts, linear slopes, or quadratic slopes, and carry no additional meaning. Although the diagram includes much information, the parts that are relevant to this analysis are summarized in Table 6.4.

In this analysis, schools are distinguished according to whether they come from USI or non-USI districts. Thus, separate means and variances of the coefficients are estimated for the USI and non-USI districts. A separate multigroup model was fit for each grade.

Results in Table 6.4 report the mean and variance of each of the coefficients in USI and non-USI districts with respect to each domain. The interpretation of coefficients is the same as had existed in the previous quadratic growth-curve model. Across all grades, a statistically significant difference between USI and non-USI districts is found in the intercept parameter, implying different levels of proficiency across USI and non-USI districts in 1994. As before, verifying the statistical significance of differences in other aspects of growth across district type is limited by the fact that only 13 USI districts could be included in this analysis. Table 6.5 provides an example of the correlations computed between coefficients, in this case for the grade 4 analysis. Correlations between common coefficients involving different domains (e.g., intercepts for Domains 1, 2, and 3) tend to be very highly correlated, underscoring the strong similarities of growth patterns across the different domains. Some differences in correlation patterns appear to exist between USI and non-USI districts, and might be attributed to slightly more variation in the growth patterns among schools in USI districts compared to non-USI districts.

One of the unique aspects of this type of modeling is that we can also directly assess model fit. Figure 6.13 provides an illustration of how district-level differences with respect to each domain are being fit by the quadratic growth model (the dashed lines indicate the model based estimates for the USI and non-USI districts). There would appear to be slight departures from quadratic growth (e.g., the 1998 scores consistently dip below their predicted values). Although this may reflect real departures from quadratic growth, it could also be due to the limitations of the pseudo-linking procedure applied to these data.

Table 6.4

Estimated Average Growth Curves in USI and Non-USI Districts with Respect to the Three TAAS Item Domains, 1994-1999

Grade 4

	USI				Non-USI			
	Mean	SE	Variance	SE	Mean	SE	Variance	SE
Intercepts								
Domain 1	.37*	.03	.35	.04	.89*	.01	.42	.02
Domain 2	.11*	.03	.39	.04	.59*	.01	.44	.02
Domain 3	-.04*	.03	.31	.03	.36*	.01	.30	.01
Linear Coefficients								
Domain 1	-.05	.02	.03	.01	-.03	.01	.08	.01
Domain 2	.05	.02	.04	.02	.08	.01	.08	.01
Domain 3	.13	.02	.03	.01	.17	.01	.05	.01
Quadratic Coefficients								
Domain 1	.05	.00	.00	.00	.04	.00	.00	.00
Domain 2	.03	.00	.00	.00	.02	.00	.00	.00
Domain 3	.01	.00	.00	.00	.00	.00	.00	.00

p<.05, Difference between USI and Non-USI

Grade 5

	USI				Non-USI			
	Mean	SE	Variance	SE	Mean	SE	Variance	SE
Intercepts								
Domain 1	.39*	.02	.18	.02	.75*	.01	.34	.01
Domain 2	.27*	.03	.24	.03	.60*	.01	.44	.02
Domain 3	.09*	.03	.18	.02	.42*	.01	.32	.01
Linear Coefficients								
Domain 1	-.03	.02	.05	.01	.01	.01	.09	.01
Domain 2	.05	.02	.05	.01	.08	.01	.10	.01
Domain 3	.15	.01	.02	.01	.14	.01	.07	.01
Quadratic Coefficients								
Domain 1	.05	.00	.00	.00	.04	.00	.00	.00
Domain 2	.03	.00	.00	.00	.02	.00	.00	.00
Domain 3	.01	.00	.00	.00	.01	.00	.00	.00

p<.05, Difference between USI and Non-USI

Grade 6

	USI				Non-USI			
	Mean	SE	Variance	SE	Mean	SE	Variance	SE
Intercepts								
Domain 1	.33*	.03	.23	.04	.86*	.02	.47	.02
Domain 2	.13*	.04	.26	.04	.45*	.02	.47	.03
Domain 3	.03*	.03	.18	.03	.38*	.01	.34	.02
Linear Coefficients								
Domain 1	-.08	.02	.03	.02	-.03	.01	.13	.01
Domain 2	.00	.02	.00	.02	.08	.01	.12	.01
Domain 3	-.05	.02	-.01	.01	.06	.01	.08	.01
Quadratic Coefficients								
Domain 1	.05	.00	.00	.00	.03	.00	.00	.00
Domain 2	.05	.00	.00	.00	.02	.00	.00	.00
Domain 3	.05	.00	.00	.00	.02	.00	.00	.00

p<.05, Difference between USI and Non-USI

Grade 7

	USI				Non-USI			
	Mean	SE	Variance	SE	Mean	SE	Variance	SE
Intercepts								
Domain 1	.23*	.05	.34	.05	.72*	.02	.45	.03
Domain 2	-.12*	.06	.39	.07	.39*	.02	.65	.04
Domain 3	-.14*	.05	.26	.04	.31*	.02	.30	.03
Linear Coefficients								
Domain 1	.08	.03	.05	.02	.06	.01	.07	.01
Domain 2	.19	.03	.03	.02	.18	.01	.09	.02
Domain 3	.18	.02	.01	.02	.14	.01	.02	.01
Quadratic Coefficients								
Domain 1	.00	.00	.00	.00	.01	.00	.00	.00
Domain 2	-.01	.01	.00	.00	-.01	.00	.00	.00
Domain 3	-.01	.00	.00	.00	-.01	.00	.00	.00

p<.05, Difference between USI and Non-USI

Grade 8

	USI				Non-USI			
	Mean	SE	Variance	SE	Mean	SE	Variance	SE
Intercepts								
Domain 1	.26*	.05	.36	.06	.91*	.02	.49	.03
Domain 2	-.23*	.05	.36	.06	.32*	.02	.49	.03
Domain 3	-.16*	.05	.33	.05	.33*	.02	.36	.02
Linear Coefficients								
Domain 1	-.18	.02	.04	.02	-.24	.01	.12	.01
Domain 2	-.18	.03	.07	.02	-.20	.01	.10	.01
Domain 3	-.09	.02	.05	.02	-.17	.01	.07	.01
Quadratic Coefficients								
Domain 1	.07	.00	.00	.00	.07	.00	.00	.00
Domain 2	.09	.01	.00	.00	.09	.00	.00	.00
Domain 3	.06	.00	.00	.00	.07	.00	.00	.00

p<.05, Difference between USI and Non-USI

Table 6.5

Estimated Correlations Between Growth Curve Coefficients in USI and Non-USI Districts Across the Three TAAS Item Domains, Grade 4, 1994-1999

Correlation Matrices Among Random Coefficients

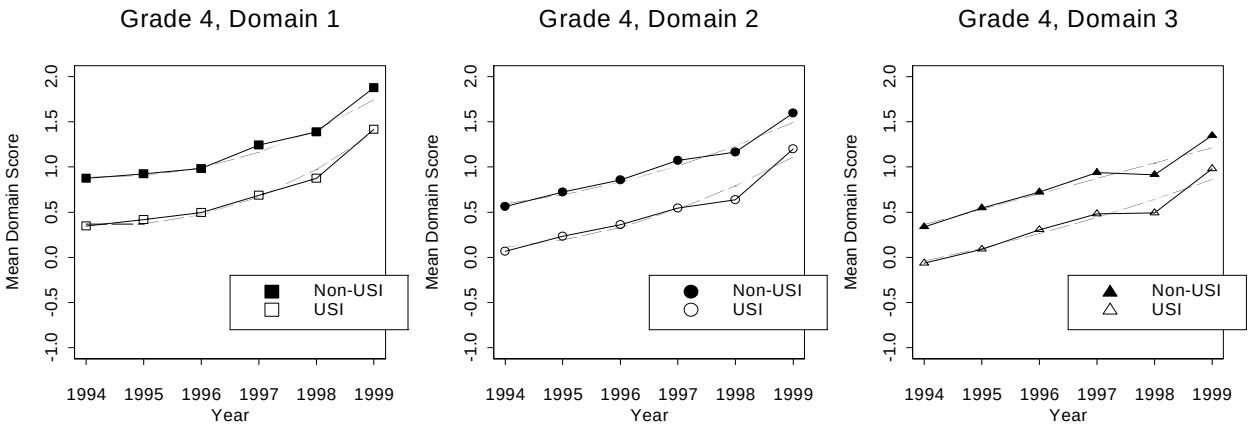
Grade 4, Non-USI

	Int,D1	Lin,D1	Quad,D1	Int,D2	Lin,D2	Quad,D2	Int,D3	Lin,D3	Quad,D3
Int,D1	1.00								
Lin,D1	-0.51	1.00							
Quad,D1	0.35	-0.92	1.00						
Int,D2	0.92	-0.43	0.28	1.00					
Lin,D2	-0.42	0.92	-0.92	-0.44	1.00				
Quad,D2	0.23	0.81	0.97	0.22	-0.89	1.00			
Int,D3	0.95	-0.43	0.27	0.95	-0.39	0.18	1.00		
Lin,D3	-0.41	0.96	-0.92	-0.39	0.96	-0.86	-0.37	1.00	
Quad,D3	0.24	-0.87	0.98	0.22	-0.91	0.98	0.19	-0.92	1.00

Grade 4, USI

	Int,D1	Lin,D1	Quad,D1	Int,D2	Lin,D2	Quad,D2	Int,D3	Lin,D3	Quad,D3
Int,D1	1.00								
Lin,D1	-0.47	1.00							
Quad,D1	0.25	-0.82	1.00						
Int,D2	0.94	-0.40	0.17	1.00					
Lin,D2	-0.38	0.80	-0.59	-0.45	1.00				
Quad,D2	0.14	-0.54	0.68	0.16	-0.77	1.00			
Int,D3	0.97	-0.50	0.26	0.95	-0.44	0.16	1.00		
Lin,D3	-0.35	0.81	-0.55	-0.34	0.80	-0.55	-0.47	1.00	
Quad,D3	0.12	-0.62	0.71	0.09	-0.66	0.81	0.22	-0.81	1.00

Figure 6.13. LISREL-based growth curve estimates comparing USI and non-USI districts over three measurement domains, grades 4 through 9.



References

Bock, R.D., Gibbons, R.D., & Muraki, E. (1988). Full information item factor analysis. *Applied Psychological Measurement, 12*, 261-280.

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

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

Appendices

A. USI District Codes, Number of Schools Included in Analyses

B. Number of USI/Non-USI Districts and Schools in Multilevel and IRT Analyses

Appendix A

USI District Codes, Number of Schools Included in Analyses

Code	District Name	Number of Schools					
		Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
1	Alamo Heights	2	2	2	1	1	1
2	Harlandale	9	9	9	4	4	5
3	Edgewood	15	15	15	13	4	6
4	San Antonio	64	64	65	17	17	21
5	South San Antonio	9	9	9	3	3	4
6	Northeast	30	30	30	8	8	13
7	East Central	4	4		2		5
8	Northside	32	32	32	12	12	19
9	Judson	11	11	11		2	6
10	Dallas	110	103	103	104	23	38
11	El Paso	47	47	47	13	13	21
12	Ysleta	32	32	32	30	11	17
13	Socorro	11	11	11	4	4	7

Appendix B

Number of USI/Non-USI Districts and Schools in Multilevel and IRT Analyses

	Full Sample				Non-USI Matched Sample	
Grade	Non-USI Dist	Non-USI Sch	USI Dist	USI Sch	Non-USI Dist	Non-USI Sch
3	938	2565	13	376	312	979
4	938	2492	13	369	404	1144
5	875	2224	13	366	338	1082
6	834	1397	13	211	246	650
7	906	1252	13	102	191	417
8	1052	1967	13	163	213	439