8. AN EVALUATION OF THE URBAN SYSTEMIC INITIATIVE AND OTHER ACADEMIC REFORMS IN TEXAS: STATISTICAL MODELS FOR ANALYZING LARGE-SCALE DATA SETS

Robert H. Meyer

Executive Summary

A multidisciplinary team worked over a year to achieve the *Systemic Initiatives: Student Achievement Analysis Study's* main goal of developing an analytic framework for studying the degree to which systemic reform contributes to improved student achievement and other outcomes. In pursuing this goal, we sought to provide information to the National Science Foundation that will enable NSF, its education constituencies, and the education research community address the following questions:

1. How can the data submitted to NSF by systemic initiatives (SIs) be used to evaluate systemic reform?
2. How does the precision of analysis depend upon the qualities of student assessment data?
3. What statistical models best fit the data linking systemic initiatives to student achievement?
4. What are the lessons learned about the kind of databases and analyses that are most effective for evaluating and understanding systemic reform?

To demonstrate our approach to developing analytic frameworks, we analyzed data from the Texas Assessment of Academic Skills (TAAS) for grades 3 through 8 from 1994 through 2000 and compared Urban Systemic Initiative (USI) districts with other districts in the state.

We identified a number of desirable features of data and databases that are needed to study the impact of the systemic initiatives and other large-scale reforms:

1. Data that describe for the USI, or other initiative, the participation level of teachers by school;
2. An identifiable control group;
3. Testing of students in consecutive years;
4. Vertically scaled scores over grades;
5. An assessment that measures the full range of student knowledge without ceiling effects;
6. Assessment data linked with student, school, and district demographic and program data;
7. Alignment of assessments with district standards and USI goals; and,
8. A means of determining student attrition rates in the population and the selective exclusion of students from testing.

We identified three general areas that can influence the precision of analyses of student achievement data in studying the impact of systemic reform that need to be considered: (1) the extent to which teachers, schools, and districts participated in the systemic initiative over time; (2) what students were excluded from the testing and analyses; and, (3) the standard error of measurement in the assessment instruments.

There is no one best model for analyzing the link between systemic initiatives and student achievement. Each model is based on specific assumptions made necessary by the incompleteness of available data or other constraints. We developed three approaches, each by a different researcher. All three models provide information about students' growth over time and compare the performance by students in USI districts with those in other districts. In this respect, the results from one model serve as a replication of those from the other models. However, each of the three researchers made different assumptions about growth that influenced which groups of students were tracked over time. Bolt examined changes in school means on 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 (e.g., 4^{th} grade students in 1995 compared with 4^{th} grade students in 1996) is due to program effects, rather than to other factors.

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 students with four data points in order to estimate the three parameters *intercept*, *slope* (linear term), and *changed* (quadratic term). Students with fewer than four data points contribute to estimating the lower-order terms. Thus, Gamoran's growth model included students who left 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 were deleted from the database used in the analysis.

Meyer's value-added analysis examined students' performance in a grade by taking into consideration achievement from the year before. His analyses included only students who 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 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 between two years.

Based on the three analytic models, we drew the following conclusions:

1. Texas Assessment of Academic Skills (TAAS) scores improved from 1994 to 2000 for all groups. Annual gain scores by Blacks and Hispanic students relative to White students

improved over time. Annual gain scores for economically disadvantaged remained constant over time.

2. USI districts began lower, but raised scores faster than non-USI districts:

   A. There is some evidence that USI districts improved faster than non-USI districts, but this may be due to the scoring metric used (e.g., TLI rather than Rasch). Meyer found positive USI effects over time for most grade levels. Bolt found a positive USI effect at grade 7.
   B. Texas USI scores started below those of non-USI districts and produced smaller annual gains than non-USIs in 1994.
   C. There is no evidence that USI districts lost ground compared to non-USI districts from 1994-2000.

3. Achievement gaps between minority and White students and between advantaged and disadvantaged students narrowed statewide on the TLI. However, this finding was highly dependent on the scaling metric used. There was much less narrowing of gaps when the Rasch scale, which is more sensitive to gains at the extreme ends of achievement, was used rather than the TLI. There was some evidence from one model that the gap between White students and Black students narrowed more in USI districts than in the contrast districts.

4. There is no difference in the rate at which achievement gaps are narrowing in USI and non-USI districts.

5. An increasing proportion of students were tested over time.

   A. TAAS attrition rates in students not being tested went down over time.
   B. There was no difference in the TAAS attrition rate between USI districts and large urban districts that are not USIs.

6. There are large differences in TAAS attrition by demographic group.

Demonstrating the impact of large-scale reform is immensely complex. This project has revealed the reasons for much of this complexity and has demonstrated specific analytic techniques that can be used to study the growth in student learning over time, given this complexity. Being restricted to only using existing data, we were unable to over come major design flaws and the lack of data on the independent variables. Even with these deficiencies, the models produced some evidence of the improvement in student learning by districts with USIs compared to other districts. The analytic models used in this study have wide applicability in studying large-scale reform. The most important implication of this study is to inform the design of the evaluations of large-scale reform efforts so that the necessary data will in future be available to more effectively measure the impact of such interventions on student learning.

Introduction

This paper draws on a large-scale data set, consisting of all students tested as part of the Texas Assessment of Academic Skills (TAAS), as a basis for evaluating the effectiveness of the Urban Systemic Initiative (USI) and academic reforms as a whole in Texas over the period 1994 to 2000. The paper focuses on developing statistical methods that are suitable for analyzing programs using large-scale data sets consisting of many students assessed in multiple years. The Texas data set examined here includes information on approximately two million students each year in grades 3 to 8. The study employs a model of student achievement in mathematics that features two levels, a student level (the "micro," or level-one equation) and a district level (the "macro," or level-two equation). It builds on the conventional postachievement-on-preachievement model, but includes several novel features:

- The model uses two alternative scales for measuring mathematics achievement. One is the Texas Learning Index (TLI), the scale used by the Texas Educational Agency (TEA). The other is a Rasch scale developed as a part of this research project. The paper investigates the extent to which major conclusions are sensitive to the choice of achievement scale.
- The model controls for measurement error in prior mathematics achievement. See Meyer (1992, 1999) for achievement models that address the problem of measurement error.
- Estimates of measurement error are derived for the Rasch scale using two approaches: the conventional approach based on an asymptotic (maximum likelihood) formula for the variance of the achievement estimate and an approach based on a finite sample formula. The latter provides substantially different and more accurate results than the conventional approach.

In the next part, we consider alternative scales for measuring mathematics achievement. Subsequent sections present the evaluation methods used in the study, finite sample methods for estimating the measurement error characteristics of the Rasch scale and Texas Learning Index, the empirical results of the study, and the conclusions of the study.

Alternative Scales for Measuring Mathematics Achievement

Achievement scores produced by the Texas Assessment of Academic Skills (TAAS) are typically reported on a scale known as the Texas Learning Index (TLI). The TLI scale is derived from a standard psychometric model of test scores, the Rasch model, but it differs substantially from the scale typically produced by the Rasch model. One of the objectives of this study is to compare results based on the TLI and the Rasch scale. TLI scores (and the Rasch scale scores that underlie this index) are designed to be comparable over time at a given grade level (horizontally equated), but are not comparable across grades (not vertically equated). As a result, it is not possible to explicitly measure student achievement growth using the Texas achievement data. This is not a major limitation, however, because the "post on pre" evaluation models used in this study do not require that posttests and pretests be measured on the same scale.[1]

[1] Many growth curve models require test scores to be measured on the same scale.

The Texas Learning Index is quite simple. In the base year (1994), TLI was a linear transformation of the raw (number correct) score *R*. In subsequent years, it was (and continues to be) a linear transformation of the equated (predicted) raw score. The transformation from raw score to TLI is given by:

$$\text{TLI} \equiv \text{T(R)} = \left[\left(\frac{R - \bar{R}_{94}}{s_{94}} \right) - z_0 \right] * 15 + 70 \qquad (1)$$

where $\bar{R}_{94}$ and s_{94} represent the mean and standard deviation of raw test scores in the base year and z_0 is the passing standard selected by the state (defined in the units of a standardized *z* score). The index is normed so that it had a standard deviation in the base year equal to 15 and a passing cut point equal to 70. It appears that in most grades the state selected a passing standard fairly close to the mean score (that is, a z score equal to zero). As a result, the TLI is essentially normed to have a mean of 70 in the base year.

Since raw scores are generally not comparable across different test forms (due to differences in the difficulty of test items), the TLI in years other than the base year is derived from an underlying Rasch score (θ) that is designed to be comparable across years (at a given grade level). Given an estimate of θ, a Rasch model is used to predict the raw score that a student would have received on the original 1994 TAAS ($\hat{R}_{94}$). Finally, this score is converted to a TLI, using formula (1). (See below for the appropriate formulas.)

The TLI and the Rasch score are both legitimate measures of student achievement. However, as indicated below, they have very different characteristics. In this study we investigate the extent to which major empirical findings are sensitive to the choice of achievement scale. To make it easier to compare the results using both scores, we normed the Rasch scale so that it has the same mean and standard deviation as the TLI in the base year (at each grade level). A separate appendix presents the technical details for how we computed the Rasch scale.

Formulas for Estimating the Rasch Scale and the Texas Learning Index (TLI)

The maximum likelihood estimate of the Rasch ability parameter θ is given by[2]:

$$\hat{\theta} = C_t^{-1}(R_t) \qquad (2)$$

where $C_t^{-1}(R)$ is the inverse function of the test characteristic curve (TCC) for the test administered in year *t.* The test characteristic curve is equal to the expected raw score:

[2] Since there is no closed-form solution to the inverse of the test characteristic curve, estimates of Rasch ability parameters are typically obtained using numerical methods. Although this is not particularly difficult or time consuming, Wright (1977) presents formulas that allow direct computation of approximate estimates of Rasch ability. These formulas were used by TEA to compute their Rasch estimates (which were used, in turn, to compute TLI). The Rasch estimates used in this paper were obtained by solving equation (2).

$$E(R_t) = C_t(\theta) = \sum_{i=1}^{n} P_i(\theta; \underset{\sim}{d}_t) \tag{3}$$

that is, the sum over all test items of the probability that an individual with ability θ gets item i correct. In the Rasch model (and other item response theory (IRT) models), ability estimates for students with extreme scores (zero correct (*R=0)* and zero incorrect (*R=n*)) are not defined. We follow the common practice of assigning Rasch scores for these students as if their actual number correct scores were $R = 0.5$ and $R = n - 0.5$, respectively. As indicated above, the Rasch ability parameter estimates are linearly transformed to obtain a Rasch scale that has the same mean and standard deviation as the TLI in the base year (at each grade level):

$$\text{Rasch Scale } = a_g + b_g\hat{\theta} \tag{4}$$

where a_g and b_g are the transformation parameters for grade *g*.

In order to obtain the TLI score in years other than the base year, the test characteristic curve for 1994 is used to predict the raw score $\hat{R}_{94}$, given $\hat{\theta}$ (computed in any year):[3]

$$\hat{R}_{94} = C_{94}(\hat{\theta}). \tag{5}$$

The TLI is then computed using (1).

Given the assumptions of the Rasch model, the item probabilities in (3) are given by the logit function:

$$P_i(\theta; \underset{\sim}{d}_t) = \left\{1 + \exp\left[-D\left(\theta - d_{it}\right)\right]\right\}^{-1} \tag{6}$$

where d_{it} is the difficulty of item i in year *t*, $\underset{\sim}{d}_t$ represents the vector of item difficulties, and $D = 1.7$.[4] In the next section, we will look at some graphs that depict the highly nonlinear relationship (implied by formula (3)) between the Rasch scale and the TLI and other features of the two scales.

A Comparison of the Texas Learning Index with the Rasch Scale

Figure 8.1 presents two graphs that compare the TLI with the Rasch scale for grade 3 in 1994. The graphs for other grades and years are very similar. The top graph plots the official TLI, a number that is rounded to two digits, against the Rasch scale. The bottom graph plots the TLI, without rounding, against the Rasch scale. Due to the lack of rounding, the bottom curve

[3] For the purpose of computing predicted raw scores (and ultimately the TLI), there is no need to "patch up" the θ estimates for the zero-correct and perfect raw scores. If a student received a perfect (zero-correct) score in any year, the predicted raw score on the 1994 test is also a perfect (zero-correct) score.

[4] The constant $D = 1.7$ scales θ so that the logit function closely approximates the standard normal cumulative distribution function (also the probit function).

exhibits greater smoothness than the top curve. Both graphs make the same point: relative to the Rasch scale, the TLI significantly rolls off achievement at the high end.[5] In fact, the curve asymptotically approaches a maximum TLI score of 93. This means that the TLI is simply not equipped to register (possibly future) increases in achievement beyond a maximum level. This implies that achievement growth near the high end of the distribution will appear more modest when measured with the TLI than with the Rasch scale. Furthermore, the distribution of the TLI is likely to be substantially skewed to the left, relative to the Rasch scale.

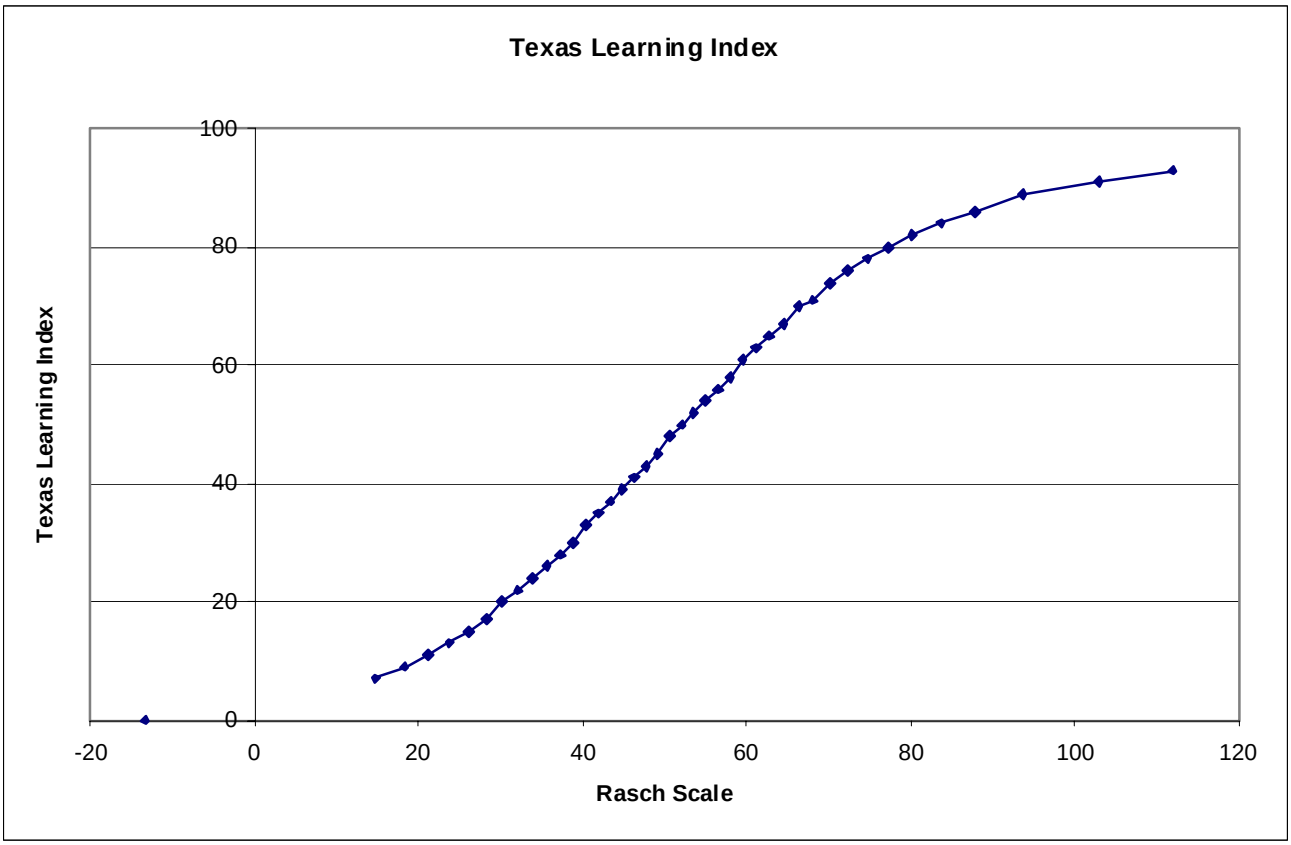
The latter prediction is evident in Figures 8.2 to 8.5. These figures present the distribution of the Rasch scale and TLI in grades 3 and 8 in two different years, 1994, the base year, and 2000, the final year of data. As indicated in Figure 8.2 (top graph), the distribution of the Rasch scale in 3rd grade in 1994 looks approximately bell-shaped. The distribution of the TLI (bottom graph), on the other hand, is heavily skewed to the left and piled up on the right side. Figure 8.3 indicates that the distribution of 3rd grade achievement shifted significantly to the right from 1994 to 2000. The average Rasch score increased from 69.78 to 80.65. The average TLI increased somewhat less, from 69.78 to 78.48. As in 1994, the distribution of the Rasch score in 2000 is reasonably symmetric with a modest concentration of data at the high end. The standard deviation of Rasch scale scores increased from 15.32 in 1994 to 16.69 in 2000. In contrast, the standard deviation of the TLI decreased from 15.32 in 1994 to 12.69 in 2000 and the test distribution became even more skewed. Figures 8.4 and 8.5 tell a similar story with respect to 8th grade test scores, although the differences between the Rasch scale and TLI are even more pronounced. Finally, Tables 8.1 to 8.6 indicate that at all grade levels the Rasch scale and TLI exhibit completely different patterns with respect to the spread of test scores. The Rasch scale paints a picture of rising average test scores and expanding spread (as measured by the standard deviation). The TLI tells a different story—rising average test scores and declining spread. These simple statistics highlight the fact that conclusions about the efficacy of programs and educational reforms may be quite sensitive to the choice of achievement scale. As a result, we will present results later in the study that make use of both the Rasch scale and the TLI.

Evaluation Methods

This section explains the statistical models and methods we use in our analyses. The first part presents a conventional post-achievement on pre-achievement model with two levels, a student level (the "micro," or level-one equation) and a district level (the "macro," or level-two equation). The second part investigates the validity of the model in light of the fact that the available set of explanatory variables is thin (as is the case with almost all administrative data bases). The third presents methods for correcting for measurement error in achievement tests. The final section presents an estimation strategy that is optimized for very large data sets (in our

[5] Achievement is also theoretically rolled off at the low end but there are very few students with test scores in this region.

Figure 8.1. Scale transformation: Texas Learning Index (TLI) and the Rasch scale in grade 3, 1994.



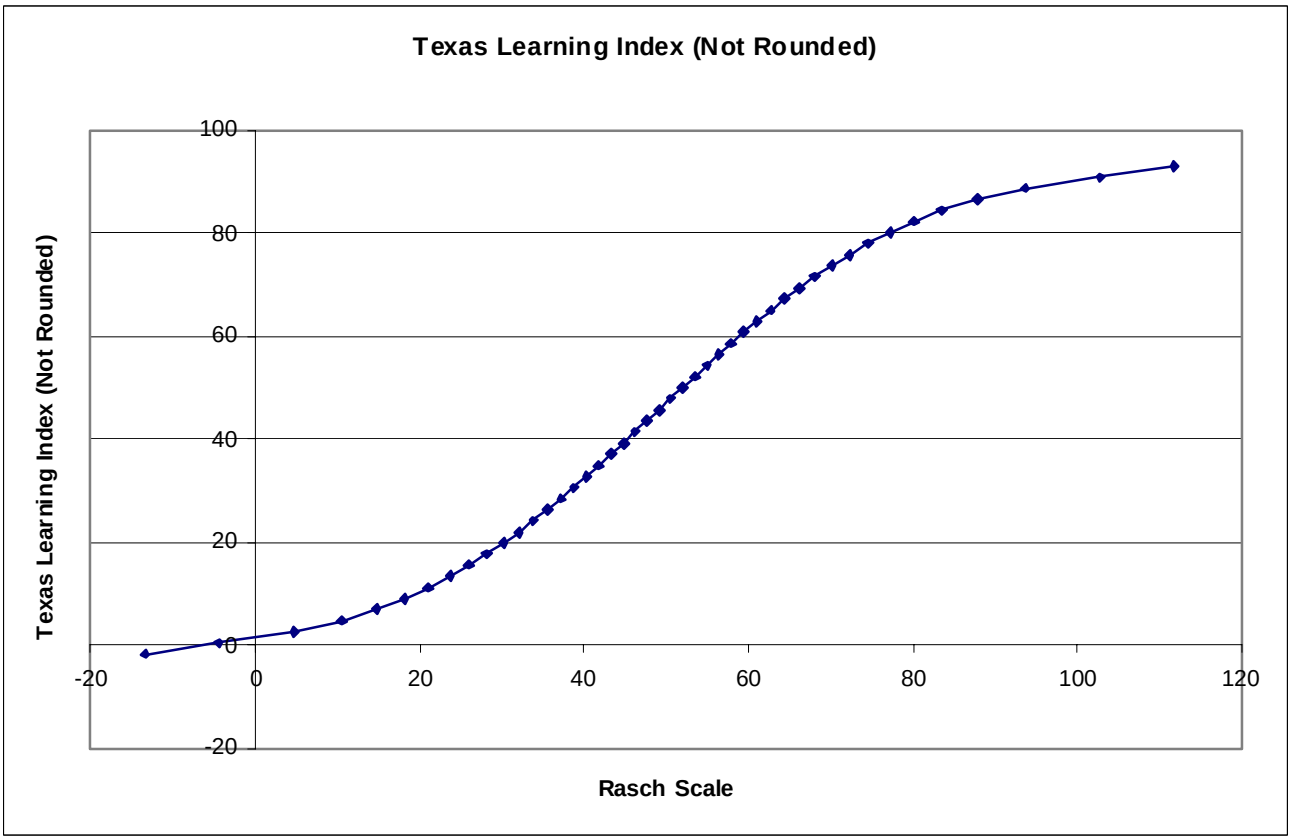
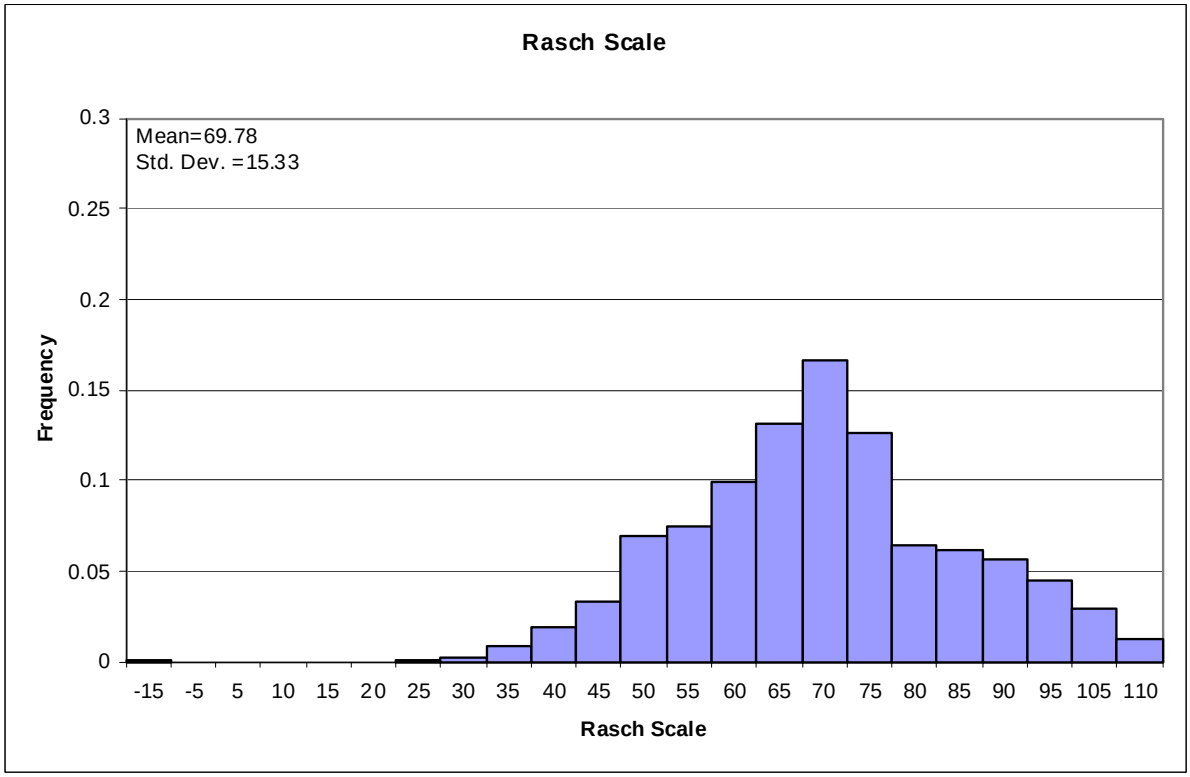


Figure 8.2. Distribution of the Rasch scale and Texas Learning Index in grade 3, 1994.



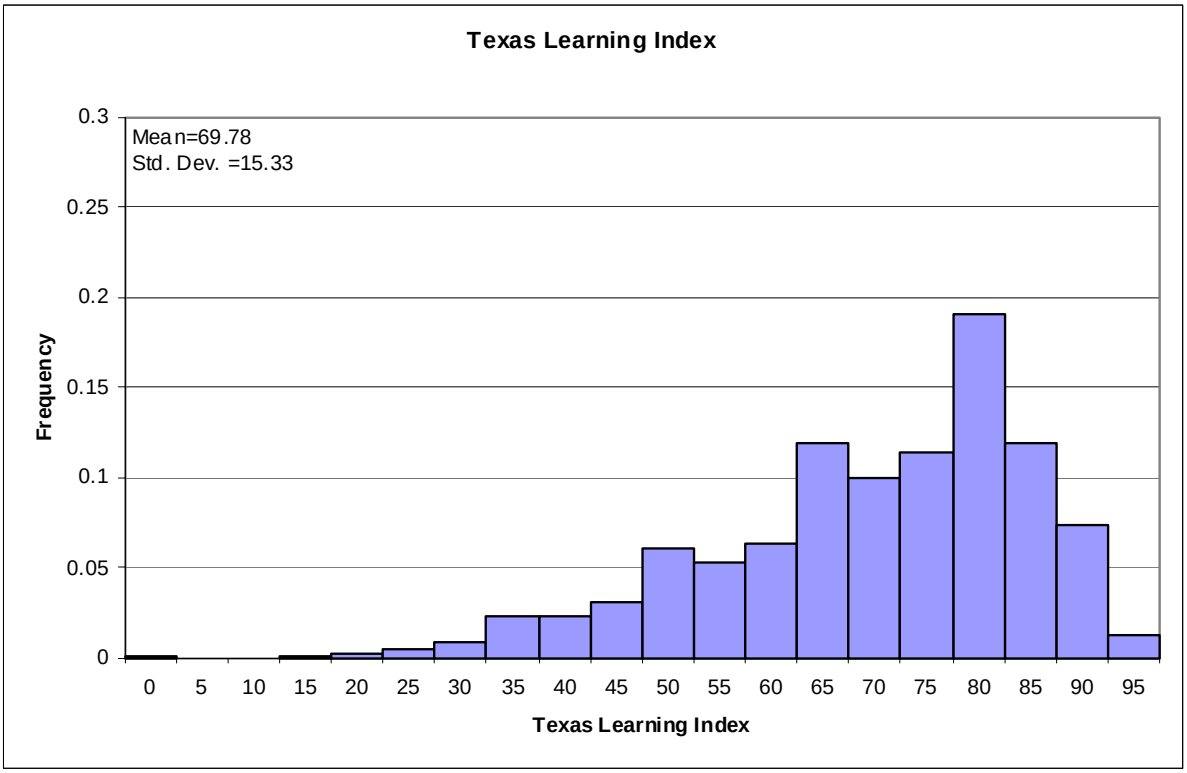


Figure 8.3. Distribution of the Rasch scale and Texas Learning index in grade 3, 2000.

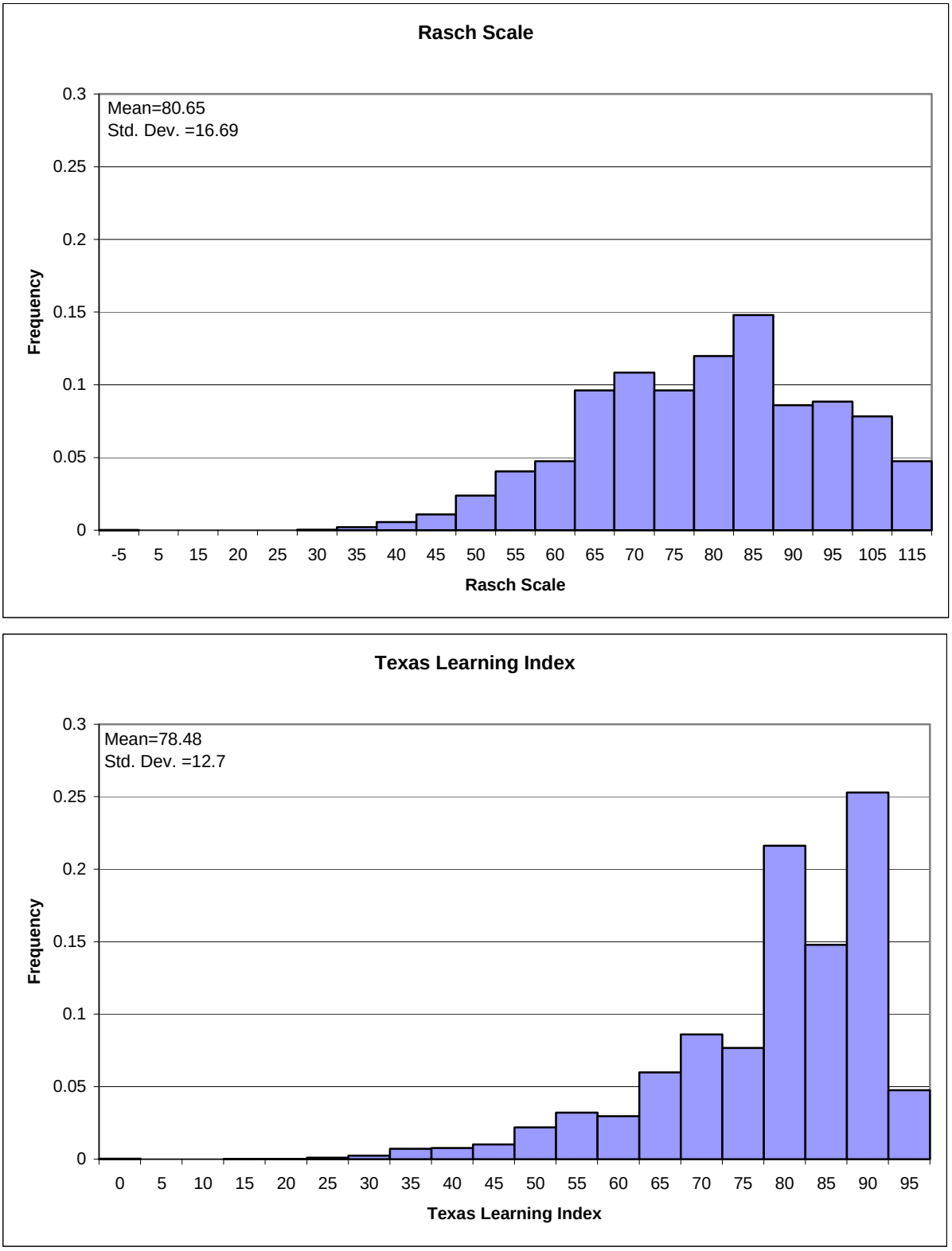
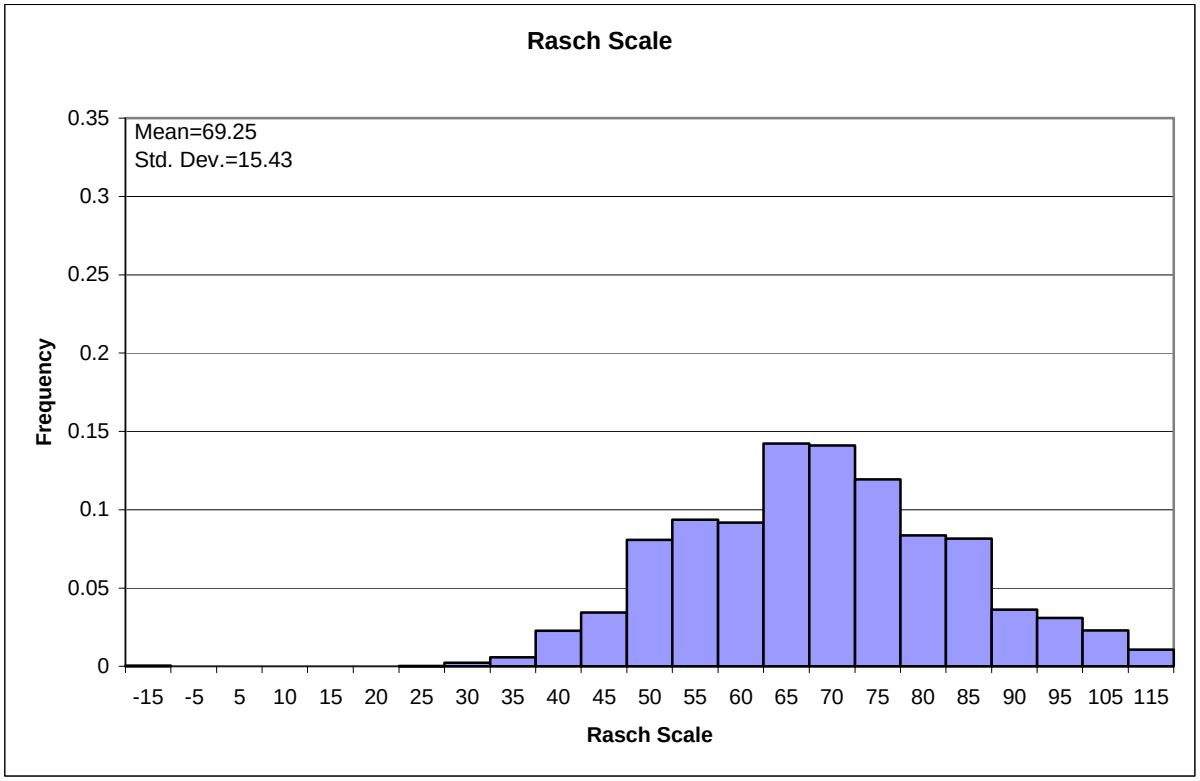


Figure 8.4. Distribution of the Rasch scale and Texas Learning index in grade 8, 1994.



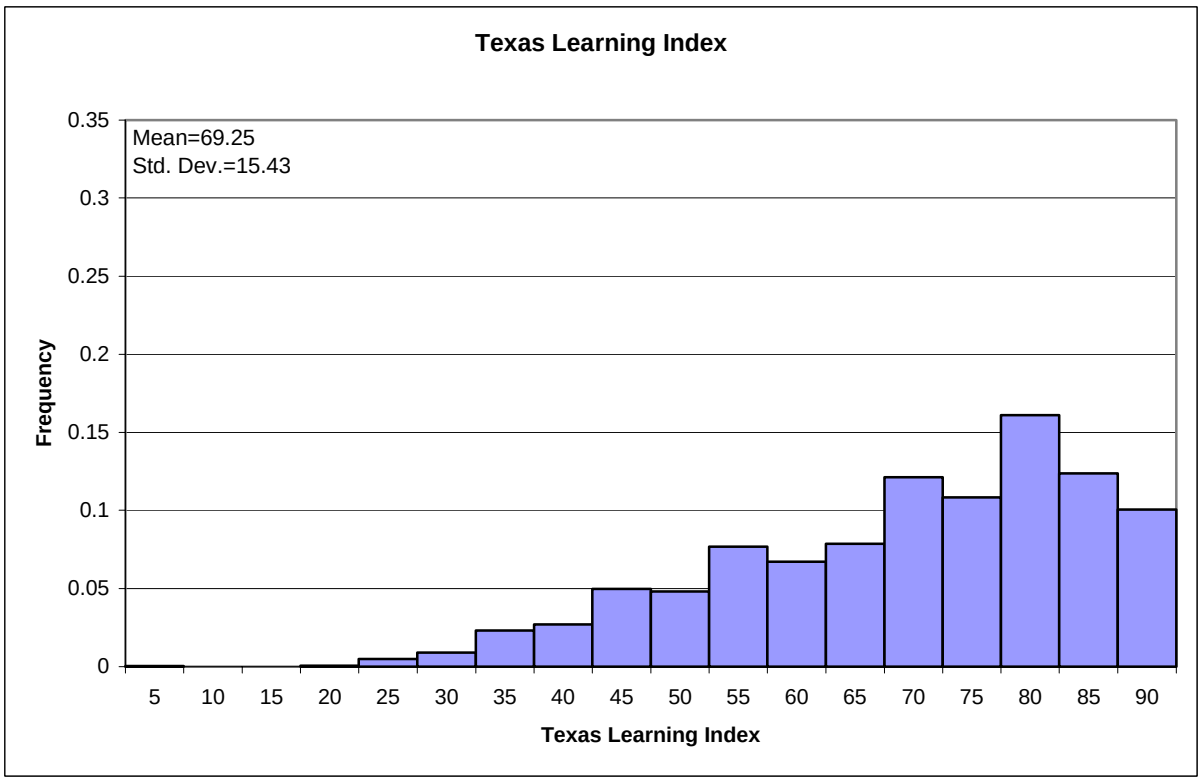
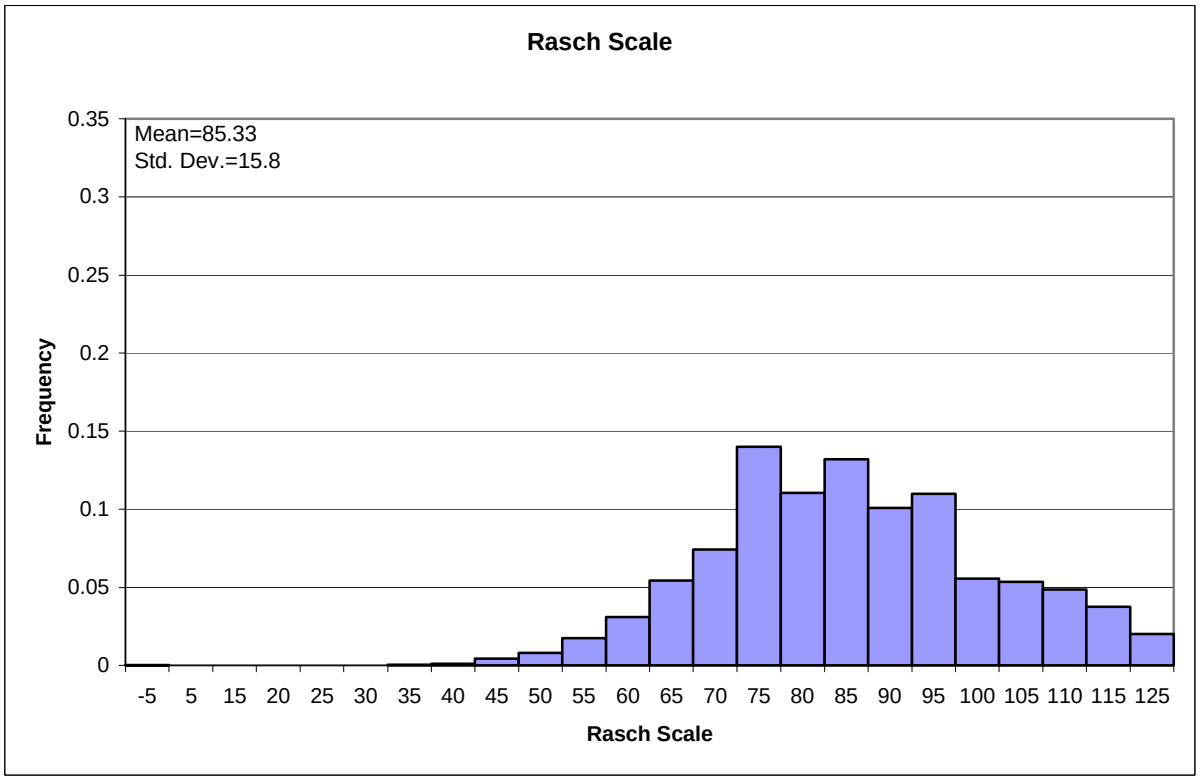


Figure 8.5. Distribution of the Rasch scale and Texas Learning Index in grade 8, 2000.



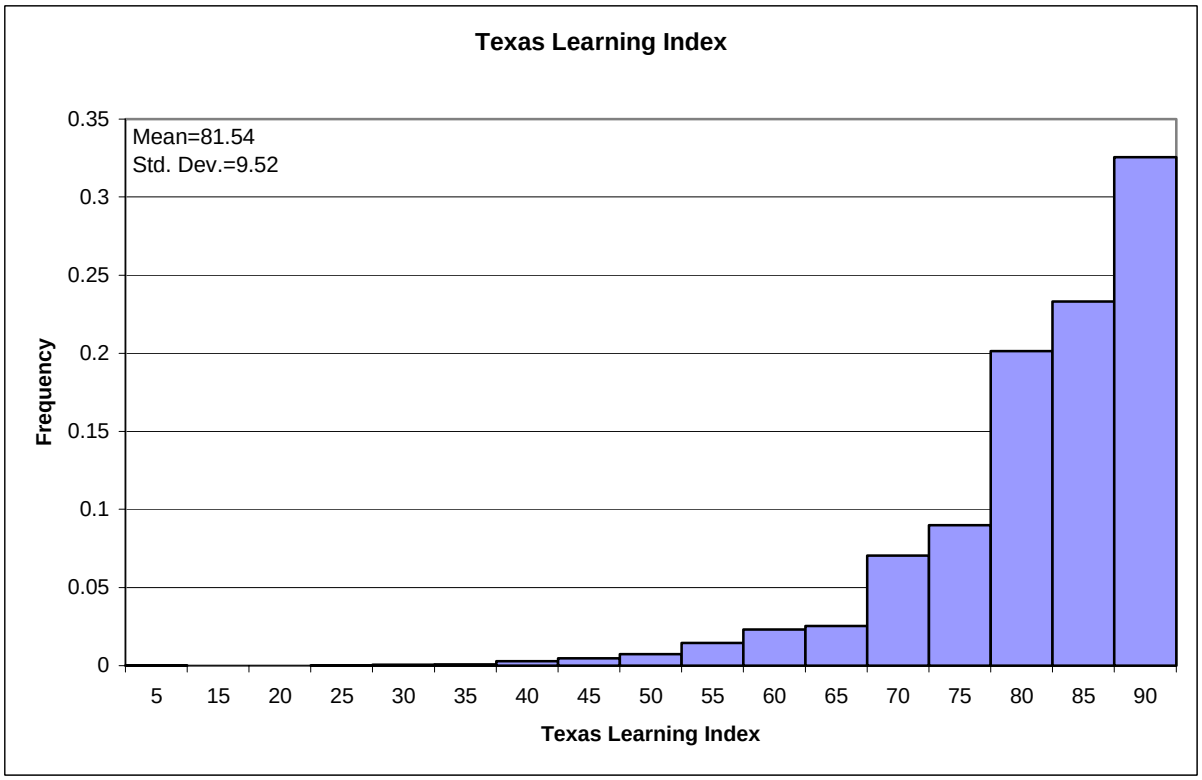


Table 8.1

Average Test Scores and Demographic Characteristics by Year, Grade 3

			1994	1995	1996	1997	1998	1999	2000
			Statistics Computed by WCER						
	TLI	Mean	69.78	72.63	75.57	77.56	77.21	78.07	78.48
		Std. Dev.	15.33	15.25	14.89	13.47	13.07	12.38	12.70
	Rasch Scale	Mean	69.78	73.25	77.18	79.40	78.45	79.60	80.65
		Std. Dev.	15.33	16.19	17.07	16.62	15.92	15.74	16.69
Ethnicity	**Black**	Mean	0.141	0.143	0.146	0.145	0.144	0.142	0.137
	Hispanic	Mean	0.286	0.288	0.290	0.297	0.306	0.322	0.359
	White	Mean	0.541	0.536	0.532	0.530	0.527	0.518	0.490
	Other	Mean	0.015	0.015	0.015	0.016	0.014	0.012	0.013
	Mixed Report	Mean	0.018	0.018	0.017	0.013	0.009	0.006	0.001
Gender		Mean	0.501	0.500	0.504	0.503	0.501	0.498	0.502
Disadvantaged		Mean	0.420	0.427	0.440	0.455	0.469	0.473	0.504
Sample			227076	218998	219919	223059	224648	223562	221098
			Statistics Reported on Website of Texas Education Agency						
	TLI	Mean	69.7	72.7	75.4	77.3	77.0	77.9	78.3
		Sample	240420	235238	238002	243208	249463	253022	263481

Table 8.2

Average Test Scores and Demographic Characteristics by Year, Grade 4

			1994	1995	1996	1997	1998	1999	2000
			Statistics Computed by WCER						
	TLI	Mean	69.87	73.92	76.16	77.82	78.87	80.54	81.03
		Std. Dev.	15.16	14.00	13.69	12.71	11.85	10.25	10.90
	Rasch Scale	Mean	69.87	74.35	77.54	79.32	81.19	83.14	84.95
		Std. Dev.	15.16	15.83	16.53	15.78	15.97	15.11	16.27
Ethnicity	**Black**	Mean	0.137	0.136	0.138	0.140	0.139	0.138	0.130
	Hispanic	Mean	0.297	0.296	0.301	0.310	0.315	0.328	0.360
	White	Mean	0.536	0.534	0.527	0.517	0.516	0.509	0.491
	Other	Mean	0.016	0.016	0.016	0.017	0.018	0.015	0.013
	Mixed Report	Mean	0.013	0.018	0.018	0.016	0.013	0.009	0.006
Gender		Mean	0.498	0.498	0.501	0.501	0.502	0.497	0.496
Disadvantaged		Mean	0.416	0.422	0.433	0.447	0.461	0.476	0.492
Sample			223607	225205	227058	230670	233130	228620	233228
			Statistics Reported on Website of Texas Education Agency						
	TLI	Mean	69.8	73.8	76.1	77.6	78.7	80.5	80.9
		N	236303	240071	235762	247002	250832	251193	264865

Table 8.3

Average Test Scores and Demographic Characteristics by Year, Grade 5

			1994	1995	1996	1997	1998	1999	2000
			Statistics Computed by WCER						
	TLI	Mean	70.23	73.78	76.22	79.33	80.83	83.04	83.99
		Std. Dev.	15.13	14.87	13.71	12.29	11.59	10.13	9.55
	Rasch Scale	Mean	70.23	74.24	77.02	81.11	83.35	87.37	89.33
		Std. Dev.	15.13	16.09	16.06	16.10	16.09	16.21	16.20
Ethnicity	**Black**	Mean	0.135	0.135	0.132	0.133	0.137	0.135	0.130
	Hispanic	Mean	0.309	0.310	0.309	0.319	0.326	0.338	0.360
	White	Mean	0.528	0.525	0.525	0.513	0.503	0.496	0.485
	Other	Mean	0.016	0.017	0.017	0.018	0.018	0.019	0.016
	Mixed Report	Mean	0.011	0.013	0.017	0.018	0.016	0.012	0.009
Gender		Mean	0.499	0.498	0.501	0.501	0.502	0.498	0.496
Disadvantaged		Mean	0.407	0.424	0.431	0.442	0.455	0.470	0.493
Sample			228383	226018	236624	239639	242323	239675	238753
			Statistics Reported on Website of Texas Education Agency						
	TLI	Mean	70.2	73.8	76.2	79.2	80.7	83.0	83.9
		N	241963	240577	252219	254528	256008	254344	263231

Table 8.4

Average Test Scores and Demographic Characteristics by Year, Grade 6

			1994	1995	1996	1997	1998	1999	2000
			Statistics Computed by WCER						
	TLI	Mean	69.81	71.69	75.65	77.62	79.28	81.27	81.95
		Std. Dev.	15.23	14.26	13.22	12.91	11.67	10.86	10.10
	Rasch Scale	Mean	69.81	71.50	76.15	79.02	81.50	84.95	85.41
		Std. Dev.	15.23	15.55	15.54	16.18	15.70	16.65	15.62
Ethnicity	**Black**	Mean	0.135	0.134	0.132	0.129	0.130	0.134	0.131
	Hispanic	Mean	0.319	0.325	0.330	0.333	0.345	0.355	0.366
	White	Mean	0.520	0.512	0.507	0.502	0.488	0.475	0.471
	Other	Mean	0.017	0.017	0.018	0.019	0.019	0.020	0.020
	Mixed Report	Mean	0.008	0.012	0.013	0.017	0.017	0.016	0.012
Gender		Mean	0.500	0.500	0.502	0.503	0.504	0.500	0.500
Disadvantaged		Mean	0.397	0.417	0.439	0.444	0.457	0.469	0.487
Sample			236264	235222	240656	254166	256868	253774	252164
			Statistics Reported on Website of Texas Education Agency						
	TLI	Mean	69.7	71.7	75.6	77.5	79.2	81.2	81.9
		N	248142	249185	255797	267428	268503	263847	266829

Table 8.5.
Average Test Scores and Demographic Characteristics by Year, Grade 7

		1994	1995	1996	1997	1998	1999	2000
		Statistics Computed by WCER						
TLI	Mean	69.68	70.92	74.30	76.34	78.22	80.39	81.55
	Std. Dev.	15.39	15.23	14.36	13.10	12.57	12.06	10.48
Rasch Scale	Mean	69.68	71.50	75.35	76.91	79.44	86.34	86.54
	Std. Dev.	15.39	16.42	16.91	15.41	16.04	18.73	17.21
Black	Mean	0.133	0.132	0.132	0.129	0.128	0.132	0.133
Hispanic	Mean	0.319	0.329	0.332	0.336	0.339	0.349	0.362
White	Mean	0.525	0.513	0.506	0.502	0.495	0.482	0.468
Other	Mean	0.017	0.018	0.018	0.020	0.020	0.021	0.021
Mixed Report	Mean	0.005	0.008	0.011	0.013	0.017	0.017	0.016
	Mean	0.500	0.500	0.504	0.504	0.504	0.500	0.501
	Mean	0.367	0.397	0.420	0.434	0.443	0.452	0.469
		232149	235539	243272	250259	260077	257337	257212
		Statistics Reported on Website of Texas Education Agency						
TLI	Mean	69.6	70.9	74.3	76.2	78.1	80.4	81.5
	N	245376	249928	258200	263350	271295	266437	267249

Table 8.6
Average Test Scores and Demographic Characteristics by Year, Grade 8

			1994	1995	1996	1997	1998	1999	2000
			Statistics Computed by WCER						
	TLI	Mean	69.25	68.79	72.55	75.48	77.39	80.11	81.54
		Std. Dev.	15.43	14.90	14.91	14.04	12.65	11.04	9.52
	Rasch Scale	Mean	69.25	68.30	72.97	76.67	78.10	81.88	85.33
		Std. Dev.	15.43	14.98	16.27	16.73	15.69	15.38	15.80
Ethnicity	**Black**	Mean	0.132	0.130	0.130	0.128	0.128	0.128	0.129
	Hispanic	Mean	0.314	0.323	0.331	0.332	0.339	0.339	0.352
	White	Mean	0.535	0.524	0.513	0.509	0.499	0.495	0.480
	Other	Mean	0.017	0.018	0.019	0.020	0.021	0.021	0.022
	Mixed Report	Mean	0.001	0.005	0.008	0.011	0.013	0.017	0.017
Gender		Mean	0.497	0.496	0.501	0.503	0.503	0.500	0.498
Disadvantaged		Mean	0.339	0.363	0.394	0.409	0.430	0.431	0.446
Sample			218012	225479	236061	244756	250303	253494	254410
			Statistics Reported on Website of Texas Education Agency						
	TLI	Mean	69.1	68.8	72.5	75.3	77.3	80.1	81.5
		N	236016	241880	252425	259065	262324	263165	263858

case, well over 10 million observations). These methods produce consistent parameter estimates using SAS software with computer run times of less than 10 minutes on a personal computer.

A Multilevel Model of Student Achievement

The model used here takes the form of a conventional multilevel "post on pre" model (postachievement on preachievement) where, at this point, we assume that all variables are measured without error. The two levels of the model are given by:

$$Y_{igt} = \gamma_g Y_{ig-1,t-1} + \beta'_g X_{igt} + \sum_{j=1}^{j=J_{gt}} \alpha_{jgt} S_{ijgt} + \varepsilon_{igt} \quad (7)$$

$$\alpha_{jgt} = \delta W_{jgt} + r_{jgt} \quad (8)$$

where i indexes students, j indexes districts, g indexes grades, t indexes years, and J_{gt} = number of districts in grade g and year t.

Equation (7) captures the student-level determinants of growth in student achievement. Y_{igt} and $Y_{ig\text{-}1,t\text{-}1}$ represent student mathematics achievement in years t and (t-1), respectively (for student i in grade g); X_{igt} represents student characteristics (such as gender, race/ethnicity, and income status); S_{ijgt} is a zero/one indicator that is equal to one if student i attends school in district j, zero otherwise; γ_g, β_g, and α_{jgt} are parameters; and ε_{igt} is a student-level error term. The parameter α_{jgt} measures district productivity, a value-added measure of the contribution of district j to growth in mathematics achievement in a single grade (g) and year (t). We follow the convention of "norming" the α parameters so that their average is zero in the base year (1994-1995) in each grade. Note that all of the parameters are allowed to differ by grade (and hence are subscripted by g). It might be reasonable to impose the restriction that γ_g and β_g are identical across grades if the achievement scores at different grade levels were measured on a common (vertically equated) scale. Since this is not the case with the Texas achievement data, we do not impose this restriction. We do, however, consider the possibility that the parameters may change over time in response to changes in policy. Equation (8) captures the district-level determinants of district productivity. W_{jgt} represents district program variables—in particular, whether the district participated in the Urban Systemic Initiative (USI) or not; δ is a parameter vector; and r_{jgt} is a district-level error term. One of the major strengths of this model specification is that it is straightforward to measure school district productivity by grade and year. This is essential if the determinants of productivity (that is; programs and policies) also vary across grades and over time, as they do in this study.

Although the above model has been used extensively to evaluate educational programs and policies, we address several potential threats to the validity of the model, in particular, weak control variables and measurement error in prior achievement. These issues are considered below.

Interpretation of a Model With Weak Control Variables

One potential problem with conducting evaluation research using administrative data is that student and family information is typically rather thin. Variables such as parental education, family status (two-parent or one-parent family), parental income, parents' occupation, and parents' attitudes toward education are known from previous research to influence growth in student achievement, but are rarely available in administrative data. This is the case in the present study. As a result, it is important to understand how this affects the validity of the analysis.

To simplify the notation, consider the student-level model for 2nd grade in a given year:

$$Y_{2i} = \gamma_2 Y_{1i} + \beta_2' X_i + \alpha_2' S_{2i} + \varepsilon_{2i} \quad (9)$$

where Y_{2i} and Y_{1i} are the post-test and pre-test, respectively, X_i is a vector with a weak set of student-level control variables, and S_{2i} is a vector of school district indicators. (The year subscripts are suppressed for convenience.) Given that the control variables are weak, it is inappropriate to adopt the conventional assumption that the student-level error term is random and uncorrelated with the right-hand side variables. Instead, we assume that the error term consists of two orthogonal components: a random component e_{2i} and a component u_{2i} that may be correlated with the regressors; that is:

$$\varepsilon_{2i} = u_{2i} + e_{2i}.$$

To allow for the possibility that that u_{2i} may be correlated with the regressors in equation (9), we define the following linear predictor equation:

$$u_{2i} = \lambda_o Y_{1i} + \lambda_1' X_i + \lambda_2' S_{2i} + w_{3i} \quad (10)$$

where w_{3i} is by definition uncorrelated with the regressors. Substituting this equation into (3) yields:

$$Y_{2i} = (\gamma_2 + \lambda_2) Y_{1i} + (\beta_2 + \lambda_1)' X_i + (\alpha_2 + \lambda_2)' S_{2i} + w_{2i} + e_{2i.} \quad (11)$$

As indicated, the λ parameters capture the bias in the parameters due to the weak regressors.

Given the possibility that the model parameters could be biased if the control variables are weak, it is important to consider whether there is some other, perhaps more statistically advanced, method for obtaining consistent parameter estimates. One possibility is the dynamic fixed effects model studied by Anderson and Hsiao (1981, 1982). (Also see Hsiao, 1986). To implement this model, we add an equation for 3rd grade achievement (for the same cohort of students as in 2nd grade). This yields the following pair of equations:

$$\begin{aligned} Y_{2i} &= \gamma_2 Y_{1i} + \beta_2' X_i + \alpha_2' S_{2i} + u_{2i} + e_{2i} \\ Y_{3i} &= \gamma_3 Y_{2i} + \beta_3' X_i + \alpha_3' S_{3i} + u_{3i} + e_{3i} \end{aligned} . \tag{12}$$

Assume that the following conditions hold:

- The parameters γ and β are identical in the models for grades two and three. (The critical assumption is that γ is identical in both equations.)
- Unobserved differences in student characteristics are captured by a time-invariant component u_i; that is, $u_i = u_{2i} = u_{3i}$.
- The random components e_{2i} and e_{3i} are not (serially) correlated.

Given these assumptions the two equations can be differenced to eliminate the time-invariant component u_i, yielding:

$$Y_{3i} - Y_{2i} = \gamma(Y_{2i} - Y_{1i}) + \alpha_3' S_{3i} - \alpha_2' S_{2i} + e_{3i} - e_{2i} . \tag{13}$$

It is straightforward to estimate this equation using the method of instrumental variables (Anderson & Hsiao, 1981).[6] At first glance, it appears that this approach yields estimates of school district effects in both second and third grade. Unfortunately, this is correct only if $S_{2i} \neq S_{3i}$; that is, only if at least some students move between districts between second and third grade. In fact, this approach will yield precise estimates of α_2 and α_3 only if there is substantial mobility between districts. This condition is unlikely to be met.

In the absence of substantial district mobility, one option is to use (13) to estimate γ and then estimate the other parameters of the model given the consistent estimate of γ (call this $\hat{\gamma}$). This is equivalent to estimating the following equation for Y_{2i} (and similarly for Y_{3i}):[7]

$$Y_{2i} - \hat{\gamma} Y_{1i} = \beta' X_i + \alpha_2' S_{2i} + u_i + e_{2i} . \tag{14}$$

Unfortunately, this method does not eliminate bias due to the possible correlation between u_i and X_i and S_{2i}. Indeed, as is suggested below, this bias may be larger than the bias obtained from the standard model (see equation (11)). To see this, define the following linear predictor equations:

$$u_i = \phi_1' X_i + \phi_2' S_{2i} + w_{1i} \tag{15}$$

$$Y_{1i} = \eta_1' X_i + \eta_2' S_{2i} + w_{2i} \tag{16}$$

where by definition w_{1i} and w_{2i} are uncorrelated with the regressors. Substituting (15) into (14) yields:

[6] Given that $(Y_2 - Y_1)$ is correlated with the error term $(e_3 - e_2)$, the acceptable instrumental variables include Y_1 and any prior lagged values of Y, S_3, S_2, and X.

[7] This method is suggested by Hsiao (1986). Note that some efficiency gains could be obtained by using the general method of moments (GMM) to estimate equations (13) and (14) jointly.

$$Y_{2i} - \hat{\gamma} Y_{1i} = (\beta + \phi_1)' X_i + (\alpha_2 + \phi_2)' S_{2i} + w_{1i} + e_{2i}. \quad (17)$$

As indicated, the ϕ parameters capture the bias in the parameters due to the weak regressors.

Now, let's compare the biases obtained from the two different approaches for the school district effects. It can be shown that the biases from the two approaches are given by:

Conventional Approach: Bias = $\lambda_2 = \phi_2 - \lambda_0 \eta_2$ (18)

Difference Equation Approach: Bias = ϕ_2 (19)

Notice that the bias from the conventional approach is the sum of two terms. If the vectors ϕ_2 and η_2 have the same sign,[8] then the two terms in (18) may be partially offsetting. As a result, the conventional approach (which yields a biased estimate of the pretest parameter if the control variables are weak) may yield estimates of school district effects that are actually less biased than the approach (based on the difference equation) that yields a consistent estimate of the pretest parameter: that is, if the control variables are weak, the best strategy may be to allow the pretest variable Y_{1i} to serve as a proxy for the omitted variables. The conventional method does this. This yields a (presumably) upward biased estimate of the pretest parameter $(\lambda_0 > 0)$, but an estimate of the school district effect that may be less biased than other estimates.

The bottom line is perhaps somewhat surprising: the conventional approach may yield estimates of school district effects that are reasonable, even if the control variables are weak.[9] In any case, the difference equation approach discussed above is not an available option in the present study due to the fact that the achievement data used in this study were not vertically equated across grades.[10] As a result, it is unreasonable to assume that pretest parameters from different grade-level models are identical, a requirement of the difference equation approach.

In the remainder of this paper, we build on the conventional "post on pre" value-added model. To keep the notation simple, we drop the bias parameters (λ's) and, in effect, redefine the level-one parameters (γ, β, and α) to incorporate the bias due to thin control variables. In the next section, we extend this model to control for measurement error in achievement scores.

Controlling for Measurement Error in Achievement Scores

As demonstrated in Meyer (1992, 1999), measurement error in prior achievement, if uncorrected, induces a downward bias in the pretest parameter γ_g and causes bias in all of the other parameters in the model. Meyer presented two methods for correcting for measurement

[8] Consider the following thought experiment. Suppose that the component u_i is a linear combination of the following omitted variables: parental education and parental income. It seems likely that school districts (or schools) that have high average achievement scores (Y_{1i}) also have high average parental education and parental income (u_i). This implies that ϕ_2 and η_2 have the same sign, or equivalently, that the parameter vectors are positively correlated across districts.

[9] A topic for future research is to develop estimation methods that yield, given reasonable assumptions, consistent estimates of all parameters.

[10] The test scores were horizontally equated across years at each grade level.

error, the errors in variables (EV) approach and the instrumental variables (IV) approach. (See Fuller, 1987) for a comprehensive presentation of these methods.) The EV approach uses externally provided information on the variance of measurement error to correct the sample variance-covariance matrix. The IV method corrects for measurement error by using instrumental variables to obtain error-free predictors of variables measured with error. The validity of the IV approach depends on whether the selected instrumental variables satisfy the required conditions—namely, that the variables are correlated with regressors in the model but uncorrelated with in-equation error and all errors in measurement. Below, we discuss how the EV and IV methods were implemented in the present study.

The key to using the EV approach is to obtain external information on the variance of measurement for all variables measured with error. In the previous studies by Meyer (1992, 1999), achievement was measured using raw (number right) scores. In the case of raw scores, it is customary to assume that the variance of test error is constant for all individuals and to compute this variance using the formulas for Cronbach's coefficient alpha (or, equivalently, Kuder-Richardson formula KR-20) (Allen & Yen, 1979). Most test developers routinely provide this information for the tests that they publish.

In the present study, we need to correct for measurement error in mathematics achievement as measured by the Rasch scale and the TLI. This poses some new technical wrinkles. With respect to the Rasch scale (or any scale developed using an alternative item response theory (IRT) model), it is customary to assume that the variance of test error differs as a function of true achievement (that is, achievement measured without error). This is, in fact, implied by item response theory (Lord, 1980). We next consider two alternative methods for estimating this variance:

- Approach 1: an approach based on a finite sample formula for the error variance.
- Approach 2: the conventional approach based on an asymptotic (maximum likelihood) formula for the error variance.

A major finding of the study is that the finite sample approach provides substantially different and more accurate results than the conventional approach. This approach deserves to be used more widely where information on the accuracy of student test scores is needed. Both estimates of measurement error are used to implement the errors in variable estimators. They are referred to as the EV1 and EV2 methods.

The standard EV approach corrects for measurement error, but it does not explicitly treat the fact that the measurement error variances of the pretest and posttest variables are not constant (homoscedastic). As a result, we experimented with a weighted EV estimator that corrects for measurement error and heteroscedasticity in the error term. Since this produced estimates (and standard errors) that were nearly identical to those obtained using the standard EV approach, we do not report the estimates in this study.

The key to implementing the IV approach is to find a credible set of instrumental variables. As indicated above, an acceptable instrumental variable must be correlated with regressors in the model but uncorrelated with the in-equation error and all errors in measurement.

Possible instrumental variables include twice-lagged achievement $Y_{ig\text{-}2,t\text{-}2}$ and the other regressors X_{igt} and S_{ijgt} (the latter assumed to be measured without error). Twice-lagged achievement satisfies two of the three conditions for an instrumental variable: it is certainly correlated with prior achievement $Y_{ig\text{-}1,t\text{-}1}$ and it is reasonable to assume that it is uncorrelated with the measurement error component of prior achievement. But, is it uncorrelated with the in-equation error ε_{igt}? In a model with rich control variables this may be a very reasonable assumption. In a model with weak control variables—our situation—it is likely that twice-lagged achievement is correlated with the nonrandom (persistent) component of the in-equation error (u_{igt}). Suppose, for example, that the nonrandom error component includes parental education and income. These variables are surely correlated with prior achievement and twice-lagged achievement and, more generally, achievement measured at any grade level.

Despite the fact that twice-lagged achievement does not satisfy one of the conditions for an instrumental variable, IV estimates based on it as an instrumental variable may be useful as a check on the EV estimates. As a result, we report EV and IV estimates later in the study. For reference, we also report ordinary least squares (OLS) estimates that do not correct for measurement error. Finally, we report IV estimates that also control for heteroscedasticity in the error term (referred to as the IV-H method). The latter estimates are nearly identical to the IV estimates.

Estimation Methods for Large-Scale Datasets

Our objective is to develop computationally efficient methods of estimating multilevel errors in variables (EV) and instrumental variables (IV) models where the number of students and districts in the sample is large—over a million student observations at a single grade level. This is important because most software packages designed for multilevel models do not offer the option to correct for measurement error using either the EV or IV approaches (for a sample of any size) and most are not well equipped to handle large datasets.

They key to efficient estimation of multilevel models with large data sets is to estimate each level equation in stages. This approach is a bit difficult to implement with small to modestly sized data sets, so many software packages (such as HLM) estimate all level equations jointly, an approach that is also difficult to implement. It turns out, perhaps surprisingly, that most of the difficulties posed by the multi-stage estimation strategy disappear with very large data sets. In contrast, the joint estimation strategy becomes harder to implement as the sample size increases.

Let's begin by developing the multistage approach, given the assumption that there is no measurement error. This assumption will subsequently be relaxed.

Level-one model with no measurement error. To simplify the presentation, we focus on the level-one (student-level) equation for a single grade (2nd grade) and year. This equation, with bias parameters added to capture bias due to weak regressors, is listed above—equation (11). To simplify the analysis, it is rewritten below in a different form:

$$Y_{2ij} = \gamma Y_{1i} + \beta' X_{ij} + \alpha_j + \varepsilon_{ij} \qquad (20)$$

where the multilevel structure of the data is explicitly recognized by including a district subscript j and a student-within-district subscript i and the grade and year subscripts are dropped for convenience. In this equation, the school district parameter α_j can be treated as a fixed effect and hence differenced out of the equation (Greene, 2000; Hsiao, 1986). To implement this, take the school district mean of (20):

$$\bar{Y}_{2.j} = \gamma \bar{Y}_{1.j} + \beta' \bar{X}_{.j} + \alpha_j + \bar{\varepsilon}_{.j} \tag{21}$$

where the bar over a variable designates the variable as a district mean and the dot in place of the i subscript indicates that the mean is computed over students in each district j. Subtracting (21) from (20) sweeps away the fixed effect α_j:

$$(Y_{2ij} - \bar{Y}_{2.j}) = \gamma(Y_{1ij} - \bar{Y}_{1.j}) + \beta'(X_{ij} - \bar{X}_{.j}) + (\varepsilon_{ij} - \bar{\varepsilon}_{.j}). \tag{22}$$

This equation can be estimated quite easily even if the number of districts is large. The variables in parentheses are constructed by computing district means and then subtracting them from the student-level regressors. This method is referred to as the "deviations from group means" approach.[11]

Given estimates of the slope parameters $\hat{\gamma}$ and $\hat{\beta}$, estimates of the district fixed effects and the precision and covariance of these effects can be computed as:

$$\begin{aligned}
\hat{\alpha}_j &= \bar{Y}_{2.j} - (\hat{\gamma}\bar{Y}_{1.j} + \hat{\beta}'\bar{X}_{.j}) \\
\omega_{jj} &\equiv \text{Var}(\hat{\alpha}_j \mid \alpha_j) = \frac{\hat{\sigma}^2}{n_j} + \bar{Z}'_{.j} \sum_N \bar{Z}'_{.j} \\
\omega_{jk} &\equiv \text{Cov}(\hat{\alpha}_j, \hat{\alpha}_k \mid \alpha_j, \alpha_k) = \bar{Z}'_{.j} \sum_N \bar{Z}'_{.k} \\
\hat{\sigma}^2 &= \frac{\text{(Sum of Squared Errors)}}{(N - K - J)}
\end{aligned} \tag{23}$$

where $\hat{\sigma}^2$ is an estimate of the variance of the student-level error, n_j is the number of students in district j, $\sum_N$ is the variance-covariance matrix of the estimated vector of slope coefficients $\hat{\mu}' \equiv [\hat{\alpha}, \hat{\beta}']$, and $Z'_{.j} = [\bar{Y}_{1.j}, X'_{.j}]$.

Note that the precision matrix $\sum_N$ depends on N. Indeed, as N increases, the matrix converges to a matrix of zeros. Hence, with large data sets, the precision and covariance of the district effects are approximately equal to:

[11] The alternative approach to estimating (20) is to include in the model indicator variables for all districts. If the number of districts us much larger than 100, this method is either very slow or simply infeasible. The Texas data includes about 1000 districts.

$$\omega_{jj} \equiv \text{Var}(\hat{\alpha}_j \mid \alpha_j) \approx \frac{\hat{\sigma}^2}{n_j}$$
$$\omega_{jk} \equiv \text{Cov}(\hat{\alpha}_j, \hat{\alpha}_k \mid \alpha_j, \alpha_k) \approx 0. \tag{24}$$

It turns out that this fact greatly simplifies estimation of level-two equations, as is discussed below.

Level-two model for a given grade. Given estimates of school district effects, the parameters of the level-two model—equation (8)—can be estimated using weighted least squares (WLS). Weighting is required for efficient estimation because, as is demonstrated below, the variance of the error in the stage-two equation is not constant. To obtain a level-two equation defined in terms of the estimated district effect (rather than the true (unknown) district effect), it is necessary to explicitly allow for the error in estimating this effect:

$$\hat{\alpha}_{jgt} = \alpha_{jgt} + \nu_{jgt} \tag{25}$$

where ν_{jgt} is the error in estimation. The stage-two equation for a given grade is then given by:

$$\begin{aligned} \hat{\alpha}_{jt} &= \delta W_{jt} + r_{jt} + \nu_{jt} \\ &= \delta W_{jt} + f_{jt} \end{aligned} \tag{26}$$

where the error term f_{jt} is composed of two parts: the in-equation error r_{jt}, assumed to have constant variance (homoscedastic), and the error in estimation ν_{jt}. As indicated in (23), the error component ν_{jt} is not, in general, independent and identically distributed (IID).

Despite the absence of IID errors, unbiased estimates of the parameters of this equation could be obtained using ordinary least squares (OLS). In small and medium-sized data sets these estimates would be inefficient and the reported standard errors could be incorrect. This problem could, in principal, be addressed by estimating the model using generalized least squares (GLS), but with large data sets this approach would be difficult if not impossible to implement. As indicated in (24), however, the dependence between estimation errors vanishes in large data sets. Thus, in large data sets it is only necessary to address the fact that the variance of the error in estimating $\hat{\alpha}_{jt}$ (ν_{jt}) is not constant.[12] This can be done quite easily, using weighted least squares (WLS).

To implement WLS, we first estimate (26) using OLS. Second, we compute the variance of f_{jt} using the estimated residuals from this regression. Third, we estimate a weighted regression where the weighted is equal to the inverse of the estimated variance of the residual f_{jt}. In large samples, the formula for this variance is equal to:

[12] In fact, in data sets where the number of students in each district is large (n_j is large), the variance of ν_{jt} approaches zero, thereby eliminating the heteroscedasticity problem. In Texas, districts vary greatly in size, so we address the problem of heteroscedasticity.

$$s^2_{fjt} \equiv \text{Var}(f_{jt}) = \sigma^2_r + \omega_{jt,jt} \tag{27}$$

where σ^2_r is the variance of the level-two in-equation error, assumed to have constant variance, and the second term is the variance of the error in estimating $\hat{\alpha}_{jt}$ (which can easily be computed using (23)). In order to compute the variance for each level-two observation (with subscript *jt*), it is necessary to estimate σ^2_r. If the number of observations in the level-two model is large (so that the slope coefficients δ are precisely estimated), then an estimate of σ^2_r is given by:

$$\hat{\sigma}^2_r = \sum_j \sum_t (\hat{f}^2_{jt} - \omega_{jt,jt}) / (N_2 - K_2) \tag{28}$$

where $\hat{f}_{jt}$ is the residual from OLS estimates of (26) and N_2 is the number of observations and K_2 is the number of regressors in the level-two model. Hanushek (1974) provides an estimate of σ^2_r for the case where the number of observations in the level-two model is not large. In this study (with about 6,000 level-two observations in 1,000 districts over six years), formula (28) was quite accurate.

Level-two model for multiple grades. One new issue arises with a level-two model that encompasses multiple years and grades, namely, the fact that the errors in estimating school district effects could be correlated for pairs of effects that are based on the same cohort of students, for example, 4th grade students in 1995 and 5th grade students in 1996. However, as indicated above, OLS estimates (and WLS estimates) of level-two parameters are unbiased even if some or all of the errors are dependent across observations. The problem with dependent errors is twofold: the WLS estimates could be inefficient (in the sense of failing to estimate parameters with the highest precision) and the reported standard errors could be incorrect. We expect that these effects are likely to be small or nonexistent in our situation because the variance and covariance of the estimated school district effects are generally quite small because the number of students in most districts is quite large (obviously much larger than the number of students per school). Furthermore, in the multiple-grade level-two model, most of the pairs of effects are based on different cohorts and thus are uncorrelated. This could be a greater problem if the data for a study were based on a single cohort, but that is not the case here.

Level-one model with measurement error. In this section, we discuss how to control for measurement error if the level-one model is estimated using the "deviations from group means" approach advocated above. As indicated in equation (22), the pretest variable, expressed as a "deviations" variable, is equal to $(Y_{1ijt} - \overline{Y}_{1.jt})$. The average variance of the measurement error component of this variable is therefore equal to:

$$\overline{\sigma}^2_v = \frac{\sum_t \left[\sum_i \sigma^2_{vit} \left(1 - \frac{1}{n_{jt}} \right) \right]}{N} \tag{29}$$

where σ^2_{vijt} is the variance of measurement error for student i in year t (in a given grade). This average variance can be computed quite simply given externally provided information on the variance of test measurement error for each student. The average variance is then used as an input to the EV method to correct for measurement error. Note that centering the pretest variable on the district mean reduces the average measurement error variance only slightly.

In the next section, we consider two alternative approaches for measuring variance of test measurement error for each student.

Exact Finite Sample and Asymptotic Methods for Estimating the Measurement Error Characteristics of the Rasch Scale and the Texas Learning Index

We considered two alternative methods for estimating the measurement error characteristics of the Rasch scale and the Texas Learning Index. Our ultimate objective was to compute the average variance of measurement error as discussed in the previous section. It is important to obtain accurate measures of measurement error variances because incorrect (biased) values of these variances will yield biased EV estimates of school district effects (and all other parameters). The first section reviews the most commonly used method for computing measurement error variances, the maximum likelihood approach. There are two potential problems with this approach. First, since maximum likelihood estimates of Rasch ability parameters are technically undefined for scores at the absolute extremes (perfect and zero-correct scores), the theory provides no suitable estimates of precision for the Rasch values that are assigned to these scores. Second, the maximum likelihood method yields estimates of ability and their precision that are consistent but not unbiased; that is, the estimates converge to the true values as the number of test items approaches infinity. Given that the number of test items on most, if not all, tests is relatively small (typically no more than 50 items), these estimates may exhibit significant bias, particularly at the extremes of the distribution. In the second section, we present an alternative method of computing error variances based on exact finite sample methods. This method yields values of error variances that are unbiased even when applied to tests with very few test items. We find that the finite sample approach produces much better estimates of error variances, particularly at the extremes. Finally, we show how to use the finite sample approach to compute error variances for the Texas Learning Index.

Maximum Likelihood Estimation of Achievement and Its Precision in the Rasch Model

Given estimates of the item difficulties for an assessment, maximum likelihood (ML) estimates of ability can be computed using equation (2). The asymptotic sampling variance (precision) of the maximum likelihood estimator is given by the inverse of the information matrix; that is

$$\sigma^2_{ML}(\theta) \equiv \text{Var}(\hat{\theta} \mid \theta) = \left\{ D^2 \sum_{i=1}^{i=n} P_i\left(\theta; \underset{\sim}{d}_t\right)\left[1 - P_i\left(\theta; \underset{\sim}{d}_t\right)\right] \right\}^{-1} \tag{30}$$

where the item probabilities are defined in (6) (Lord, 1980).[13] Although the variance is defined in terms of the true θ, this value is unknown. Thus, it is customary to evaluate (30) at the estimated value of θ.

One of the problems with the ML estimator is that it is undefined for perfect and zero-correct scores. (Technically the ML estimates associated with perfect and zero-correct raw scores are positive and negative infinity, respectively.) As a result, it is customary to assign arbitrary upper and lower bound values to the extreme scores. One commonly used ad hoc method of generating bounds is to compute the ML estimates associated with raw scores of (n – ½) and ½, respectively. The variance of these estimates is then computed using (30). One apparent problem with this approach is that it produces estimates of error variances that are enormous (see below for specific estimates). A second (possible) problem with the ML estimator is that it yields estimates of achievement and precision that are consistent but not unbiased. Given that the number of test items on most, if not all, tests is relatively small (typically no more than 50 items), these estimates may exhibit significant bias, particularly at the extremes of the distribution. In the next section we present an alternative method of computing error variances based on exact finite sample methods.

Exact Finite Sample Estimation of Precision in the Rasch Model

In our analysis, we developed formulas for computing the exact finite sample formulas of the sampling variance for the ML estimator of Rasch achievement. As is the case with the asymptotic sampling variance, the formulas developed below are a function of the true (unknown) achievement parameter θ. The key to this approach is that the formulas are defined as the summation over the raw (number right) score, a sufficient statistic in the Rasch model (as opposed to the summation over items—as in (3)). As indicated in the next section, this approach also works for computing the sampling variance of the TLI, since this scale is also a function of the raw score—see equations (1) and (5).

We are interested in the formulas for the conditional mean and variance (given θ) of a scaling function $S(r)$ of the raw score r. In our application, the scaling functions yield the Rasch scale and TLI, respectively. These formulas (Kolen, Zeng, & Hanson (1996) and Lee, Brennan, and Kolen (2000) are given by:

$$\text{Conditional Mean: } \mu(\theta) \equiv ES(\theta) = \sum_{r=0}^{r=n} \pi_{rt}(\theta) S(r) \qquad (31)$$

[13] The precision formulas considered in this section all neglect possible uncertainty due to imprecision in the item parameters. This is a reasonable assumption if the item parameters are estimated from a large sample of individuals (as is the case in this study).

$$\text{Conditional Variance: } \sigma_v^2(\theta) \equiv E[S(r) - ES(r)]^2$$
$$= \sum_{r=0}^{r=n} \pi_{rt}(\theta)[S(r) - ES(r)]^2 \qquad (32)$$
$$= \sum_{r=0}^{r=n} \pi_{rt}(\theta)S(r)^2 - [ES(r)]^2$$

where $\pi_{rt}(\theta)$ equals the probability of obtaining a raw score of *r* in year *t* (on the test form administered in that year). Lord and Wingersky (1984) and Kolen, Zeng, and Hanson (1996) provide a clever recursion formula for computing these probabilities (given the item probabilities defined in (6)).[14]

To obtain the desired conditional mean and variance of the Rasch ability parameter, the above formulas are evaluated using the scaling function that maps the raw score to the Rasch ability parameter—equation (2). (The results below are reported using the Rasch scale, a linear transformation of the underlying Rasch ability parameter—see equation (4).) Let's focus first on the conditional mean. As indicated in Figure 8.6, the mean of the Rasch scale is very close to the true value except at the extremes. At the top of the distribution, the Rasch scale is biased downward. Similarly, it is biased upward at the low end.[15] This indicates that the Rasch scale is approximately unbiased over most of the distribution of scores. Nonetheless, since the distribution of test scores is concentrated on the high side of the distribution (which means that the TAAS tests are relatively easy for the student population—see Figures 8.2 to 8.5), the Rasch scale appears to be slightly biased against high achievers. Moreover, due to the shift in the test score distributions over time, this bias is likely to increase over time.

Now, let's focus on the finite sample and asymptotic approaches to computing the magnitude of measurement error of the Rasch scale. Figure 8.7 reports the standard error of measurement (SEM) on the 3rd grade test in 1994 and 2000. (The results for other grades and years were similar and thus are not reported.) As indicated, the SEM profile computed using the asymptotic formula exhibits the typical U-shaped pattern reported by all test developers that score their tests using Item Response Theory models. In our case, the SEMs at the high and low extremes of the distribution are more than four times the minimum SEM. In contrast, the SEM profile computed using the finite sample formula exhibits an M-shaped pattern. Moreover, the level of measurement error at the extremes is only slightly higher than the minimum value. Interestingly, the asymptotic SEM is a bit too low near the center of the distribution.

The overall differences between the two methods of computing the measurement error variance of the Rasch mathematics scale are summarized in Table 8.7. The table reports the

[14] These formulas, as well as the ML formulas discussed above, assume that the assumptions of the Rasch model are valid. The assumption of local independence (Lord, 1980) is particularly important since it is equivalent to an assumption that the test measures a unidimensional achievement construct.

[15] These effects at the extremes are inevitable. Given that tests have a finite number of items, there is always a minimum and maximum test score. As a true score rises, it is increasingly likely that a test score will bump into the test ceiling. Similarly, as a true score falls, it is increasingly likely that a test score will crash into the test floor. These effects can be avoided if a test includes items that are well matched (at the bottom, in the middle, and at the top) to the abilities of the sample being tested.

Figure 8.6. The conditional mean of the estimated Rasch scale in grade 3 by year.

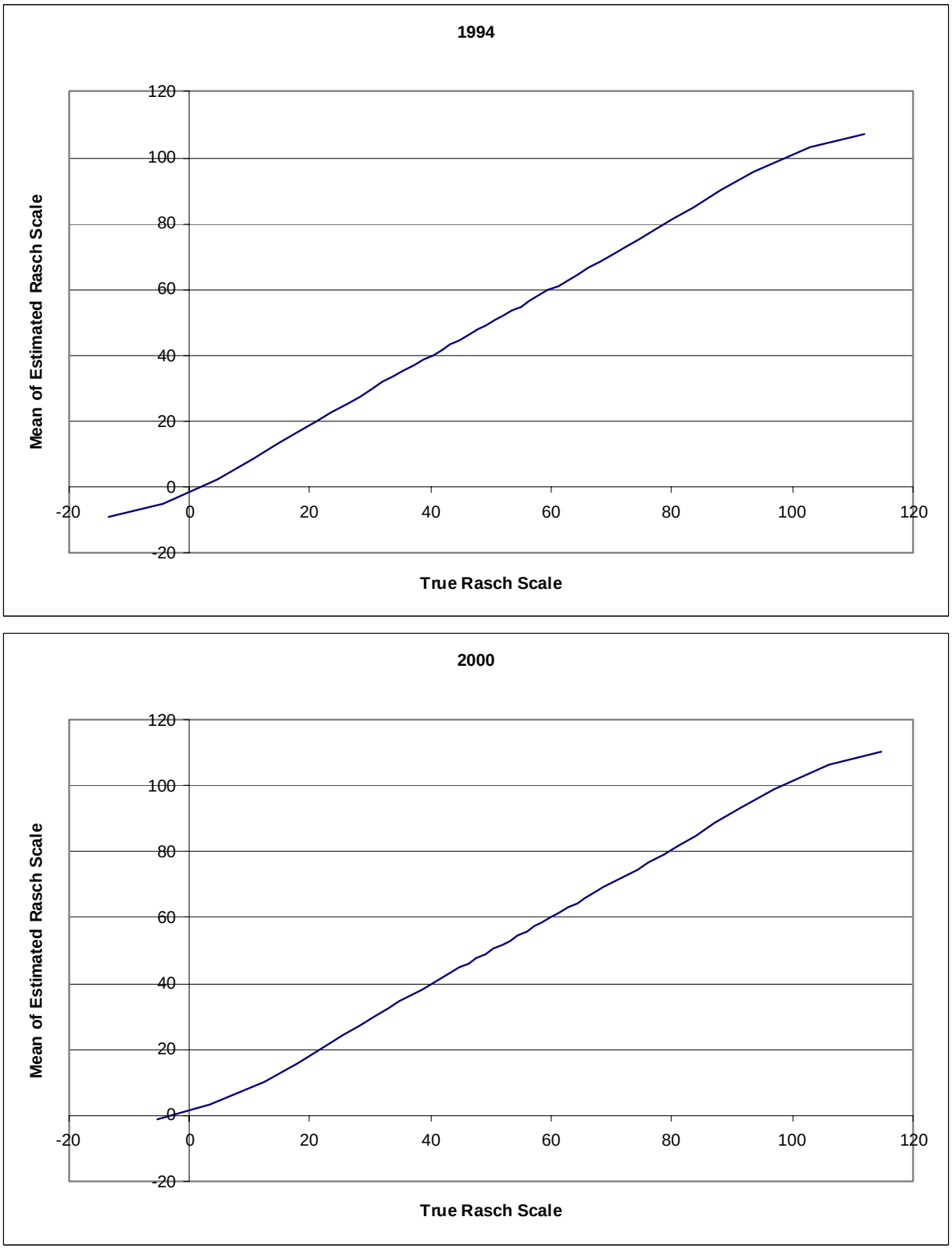
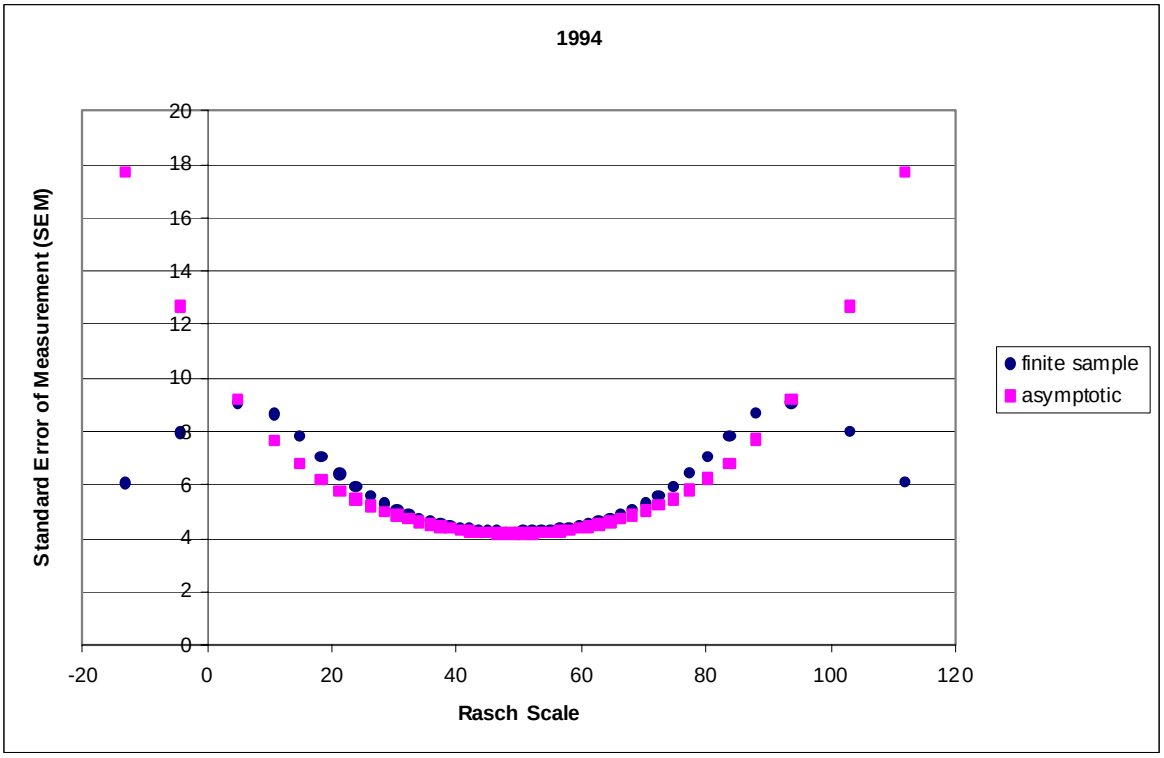
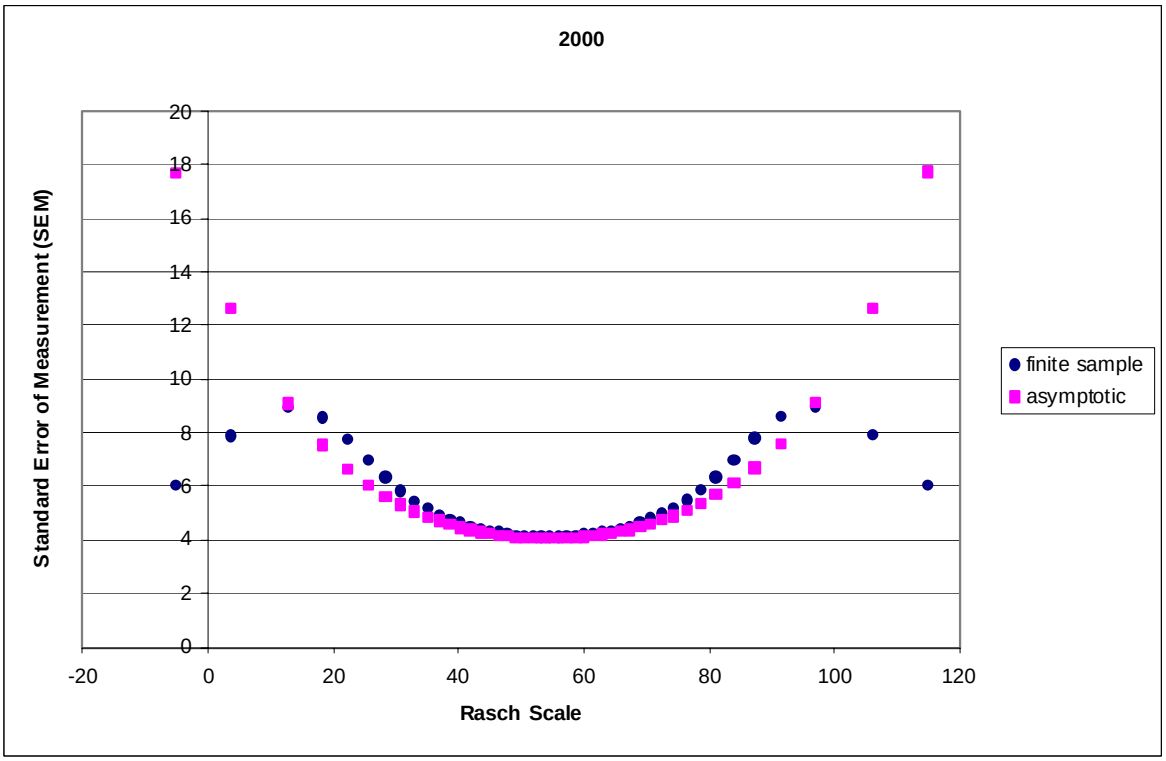


Figure 8.7. Rasch scale finite sample and asymptotic standard error of measurement in grade 3 by year.





average error variance by grade and year for the two methods (and the associated sample sizes).[16] Note that in all cases the asymptotic approach exaggerates the level of measurement, in some cases by almost 50%. Note also that the level of measurement error is highest in the early grades. The bottom line is that the traditional approach based on the ML estimator of the asymptotic sampling variance works quite poorly. We strongly recommend the finite sample approach.

Exact Finite Sample Estimation of the Precision of the TLI

It is straightforward to apply the finite sample approach—formula (32)—to compute the sampling variance of the TLI. The relevant scaling function is equal to:

$$S(r) = \text{Round}\left\{\text{T}\left[\text{C}_{94}[C_t^{-1}(r)]\right]\right\} \tag{33}$$

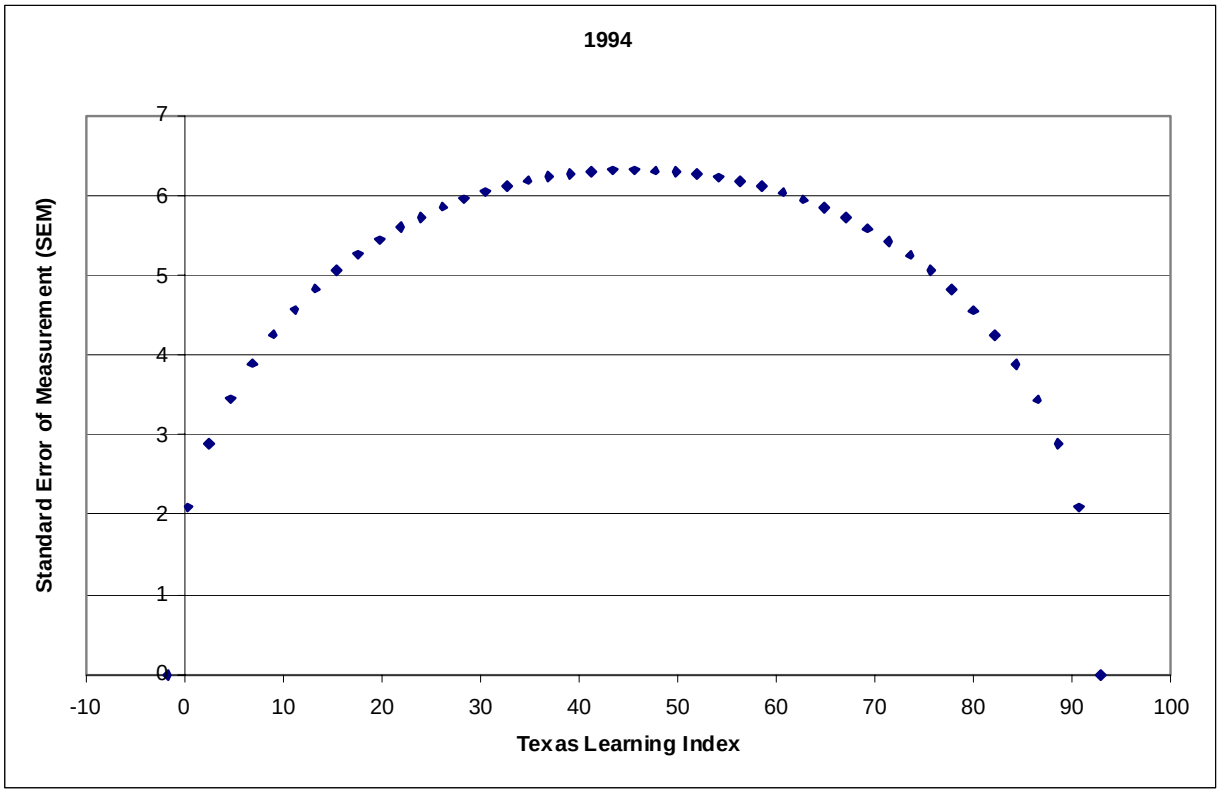
where the relevant functions are defined in equations (1), (2), and (5). Note that the scaling function explicitly allows for error due to the fact that the TLI is rounded to an integer value. This scaling function applies to both the base year (1994) and other years. In the base year, however, the interior part of the function (that is, $\text{C}_{94}[C_t^{-1}(r)]$) simplifies to the raw score *r*.

Using the finite sample approach, the SEM profile for the TLI is graphed in Figure 8.8 for 3rd grade in 1994 and 2000. In contrast to results obtained with the Rasch scale, the minimum error variance is associated with the high and low test scores (the perfect and zero-correct scores). As in the case of the Rasch model, the computed error variances for the TLI are used to obtain EV estimates of the TLI achievement model.

In the remainder of this study, we present our empirical findings. The next section briefly describes the Texas data used in the study.

[16] The average variance reported in the table is equal to the average of the computed individual-specific variances. As indicated in the text, the formulas for individual variances are theoretically defined in terms of the true score. Since this score is unknown, however, we follow the practice of evaluating the formulas at the estimated score. It is possible that the average variance computed in this way differs from the average variance based on the true individual-specific variances. If so, then the measurement error corrections based on these computed variances could be faulty. To evaluate this possibility, we conducted several Monte Carlo simulations to determine whether the computed average variance differs from the true average variance. The simulations indicated that the two average variances were very close.

Figure 8.8. The standard error of measurement of the Texas Learning Index (TLI) in grade 3 by year.



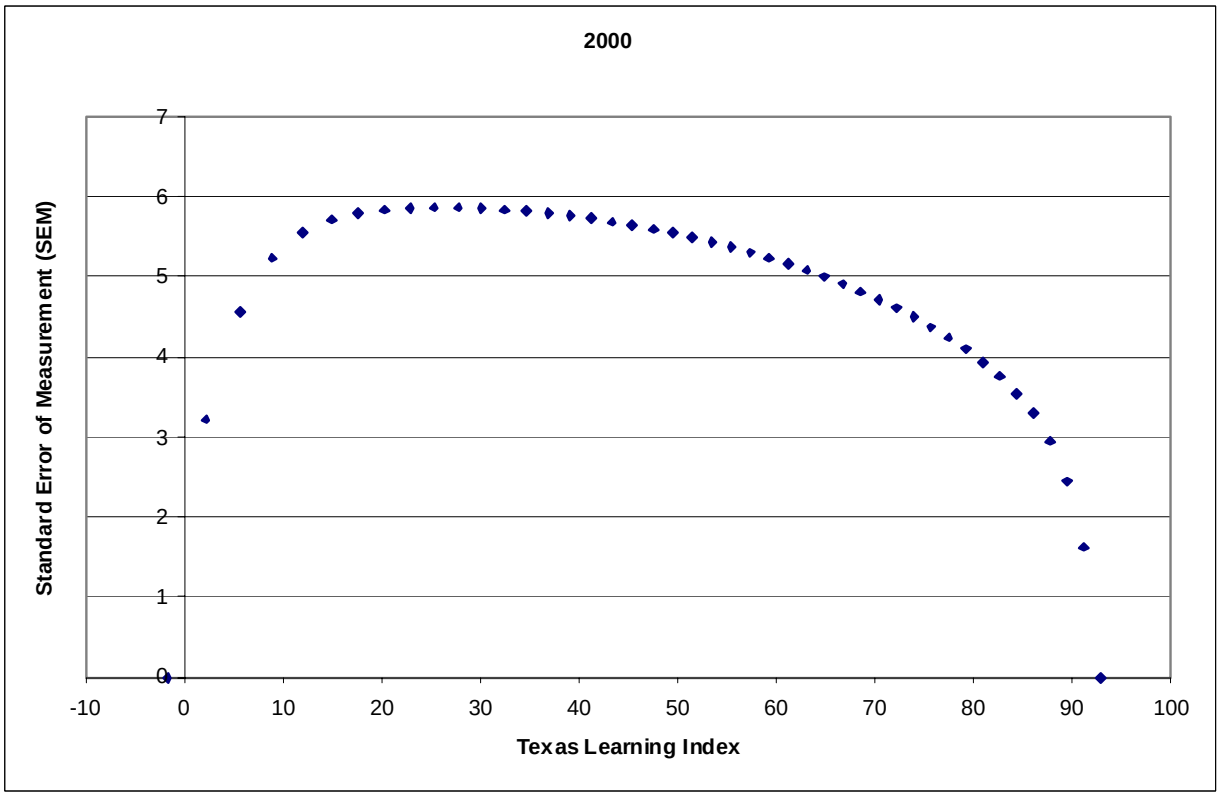


Table 8.7
Average Measurement Error Variance by Method, Grade, and Year

Average Finite Sample Measurement Error Variance by Grade and Year

	3rd	4th	5th	6th	7th
1994	36.84	28.66	27.91	25.22	25.42
1995	41.97	29.98	33.16	22.97	24.18
1996	43.74	33.50	31.59	28.77	27.05
1997	46.24	35.47	33.46	31.20	27.32
1998	46.41	38.99	37.65	33.78	28.49
1999	49.81	39.02	40.61	37.16	42.52

Average Asymptotic Measurement Error Variance by Grade and Year

	3rd	4th	5th	6th	7th
1994	39.22	31.76	29.42	26.62	26.18
1995	49.77	37.58	41.91	25.68	26.35
1996	60.09	48.18	39.59	34.69	31.78
1997	69.08	49.21	47.61	41.88	29.33
1998	56.86	55.28	53.30	41.90	32.94
1999	64.84	57.01	53.52	52.46	50.46

Sample Size (Data Used in Achievement Model)

	3rd	4th	5th	6th	7th
1994	195851	196677	200428	203956	198652
1995	195602	203434	202551	207982	206547
1996	196328	204943	212387	213514	213692
1997	202298	211026	217953	228390	222799
1998	200186	209899	218170	228349	228496
1999	201576	209523	219885	230078	230756
Total	1191841	1235502	1271374	1312269	1300942

Average Measurement Error Variance by Grade (Pooled Across Years)

	3rd	4th	5th	6th	7th
Asymptotic	56.75	46.70	44.49	37.63	33.16
Finite Sample	44.21	34.36	34.18	30.06	29.39
Ratio	1.28	1.36	1.30	1.25	1.13

Data

In this section, we briefly consider the student-level data and the district-level data used in the study. The most notable characteristic of the Texas data set is that it is huge: over 10 million student-level observations and a thousand school districts. Student data is available for seven different years, 1994-2000, and six different grades, 3-8. In each grade and year there are approximately 220,000 to 260,000 student observations. Thus, there is more than enough data to support separate analyses by grade and year, as well as analyses pooled across years at a given grade level. All public school districts are included in the database, as are all public school students who took the mathematics part of the Texas Assessment of Academic Skills (TAAS).

The TAAS is administered in the late spring of each school year. As previously indicated, TAAS results are reported on a scale developed for the Texas Education Agency (TEA), the Texas Learning Index (TLI). Extensive information on the TAAS and the TLI is provided at their website (www.tea.state.tx.us/student.assessment/). TLI data was provided to us by TEA. We constructed the Rasch scale scores that underlie the TLI in order to be able to compare the two different ways of scaling the test scores. As already indicated, information on student characteristics is rather thin. Included in the data are measures of race/ethnicity (Black, Hispanic, White, other, and mixed reporting of ethnicity over time), gender (male, female), and income status/indicator of economic disadvantage (whether a student participated in free lunch or not).

Tables 8.1 to 8.6 (see above) provide summary information on the student test scores and demographic characteristics by grade and year. The tables report mean TLI scores and associated sample sizes as computed from our database and as reported in the TEA website. In all cases the means are close, but not exactly the same. The differences are probably due to differences in the samples used to estimate the means. In general, the samples used in our analyses appear to be somewhat smaller than the samples used to compute the state statistics. This is probably due to the fact that we imposed stricter criteria for including observations in the analysis sample than the state uses.[17]

As discussed previously, Tables 8.1 to 8.6 tell a story of rising test scores over time at all grade levels—both with respect to the Texas Learning Index (TLI) and the Rasch scale. The details give quite a different picture, however. In 3rd grade, for example, the average TLI increased from 69.78 in 1994 to 78.48 in 2000, an increase of 8.70. Over that period, the spread of the TLI, as measured by the standard deviation, declined from 15.33 to 12.70. Over the same period, the average Rasch scale increased from 69.78 to 80.65, an increase of 10.87, and the standard deviation also increased, from 15.33 to 16.69. This is a striking difference. It is due to the fact that the Rasch scale is more sensitive to growth at the high end of the achievement scale than the TLI.

In addition to information on test scores, Tables 8.1 to 8.6 report the fraction of students by demographic group. Note that the fraction of Hispanic students and the fraction of students who are disadvantaged increased significantly over time at all grade levels. For example, in 3rd

[17] We eliminated student observations with missing demographic information, missing or duplicative student identifiers, and inconsistent grade enrollment data.

grade in 1994, 28.6% of the students were Hispanic. By 2000, Hispanic representation had increased to 35.9%. Similarly, the proportion of disadvantaged students in 3rd grade increased from 42.0 percent in 1994 to 50.4 percent in 2000. These trends may reflect underlying demographic changes in the Texas school-aged population, or they may indicate that the system did a better job of testing Hispanic and disadvantaged students in later years.

In order to explore these possibilities, we have plotted data on the number of students tested for the three largest racial/ethnic groups in Texas (White, Hispanic, and Black) and for disadvantaged and non-disadvantaged students by year and cohort (Figures 8.9 to 8.13). Cohorts are identified by the year in which the students attended 3rd grade. Twelve different cohorts are represented (1989 to 2000). As indicated in the graphs, the number of Hispanic students tested varied enormously over time and across cohorts. In general, the number of students tested increased substantially over time (equivalently, over grades) within each cohort. For example, the number of Hispanic students tested in the 1995 cohort increased from about 63,000 students in 1995 (3rd grade) to 90,000 students in 2000 (8th grade). In contrast, the number of Black students tested varied minimally over time and across cohorts. The number of White students tested varied somewhat, but with no strong overall trend.

Given the striking differences in the testing patterns of Hispanic and non-Hispanic and disadvantaged and non-disadvantaged students, it may be useful to examine the above data using a formal model of the incidence of testing. Define N_{gt} as the number of students tested in grade g and year t for a given demographic group; E_{gt} as the number of students enrolled in grade g and year t for a given demographic group; and F_{gt} as the fraction of enrolled students tested in grade g and year t. Then, the number of students tested is by definition equal to $N_{gt} = E_{gt} \cdot F_{gt}$. Since we do not have information on student enrollments (only those tested), we cannot estimate separate models of E_{gt} and F_{gt}. As a result, the two models must be combined to obtain an overall model of the number of students tested. Since the two components of N_{gt} interact multiplicatively, it is convenient to specify the enrollment and fraction-tested models as semi-log models.[18]

We suppose that student enrollment is determined by the size of the corresponding cohort c (where, as above, cohort is defined by the year in which a student attended 3rd grade) and possibly by changes in enrollment over time due to migration in and out of the state. A semi-log model of student enrollment is given by:

$$E_{gt} = \exp(\delta_c + \phi_t + \nu_{1gt}) \tag{34}$$

where δ_c captures the size of cohort c, ϕ_t captures possible changes in enrollment over time, and ν_{1gt} is an error component. We suppose that the fraction of students tested may vary systematically due to possible differences in testing policies over time and across grades.

[18] Since fraction tested is the outcome of individual events (tested/not tested), F_{gt} could alternatively be modeled as the average of individual probabilities (using, for example, probit or logit models of the probability that an individual is tested). This approach is less convenient than the one presented in the text because the resulting models are highly nonlinear.

Figure 8.9. Number of White students tested by cohort and year.

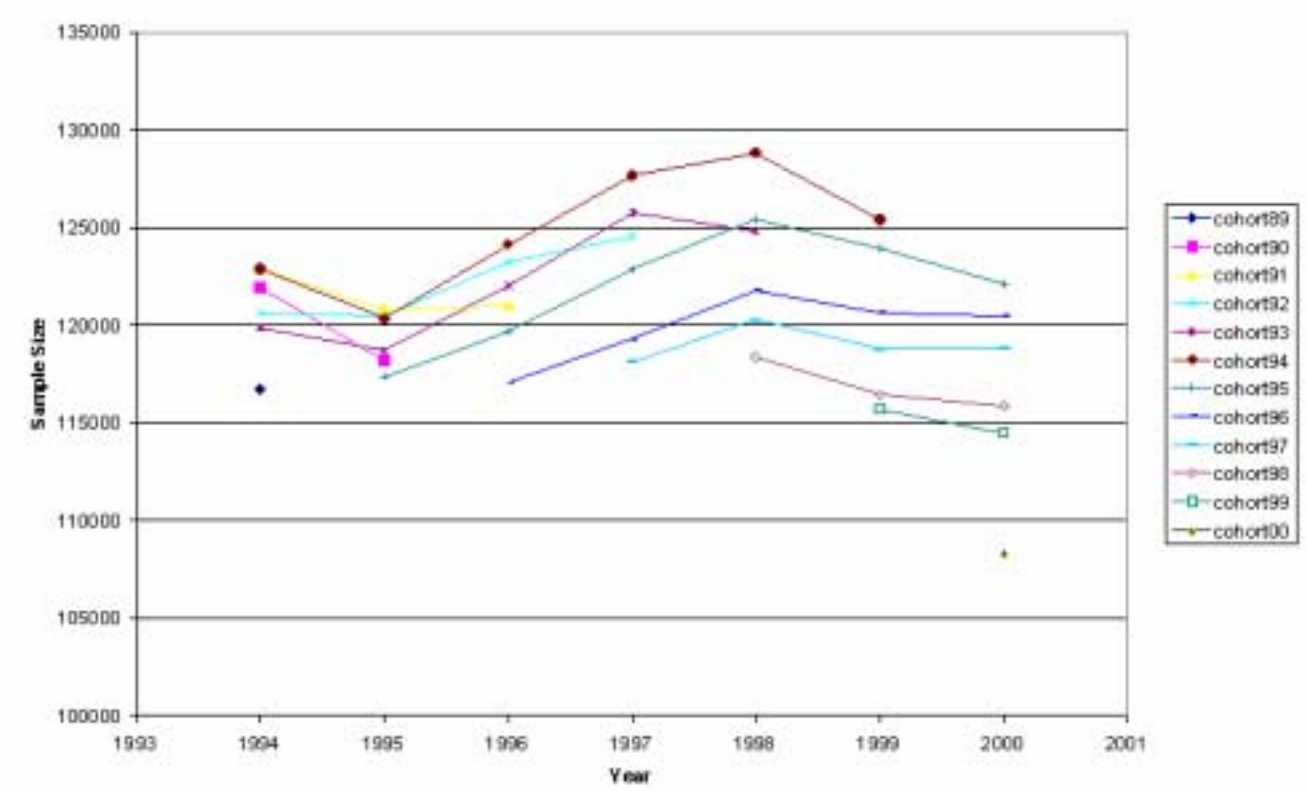


Figure 8.10. Number of Hispanic students tested by cohort and year.

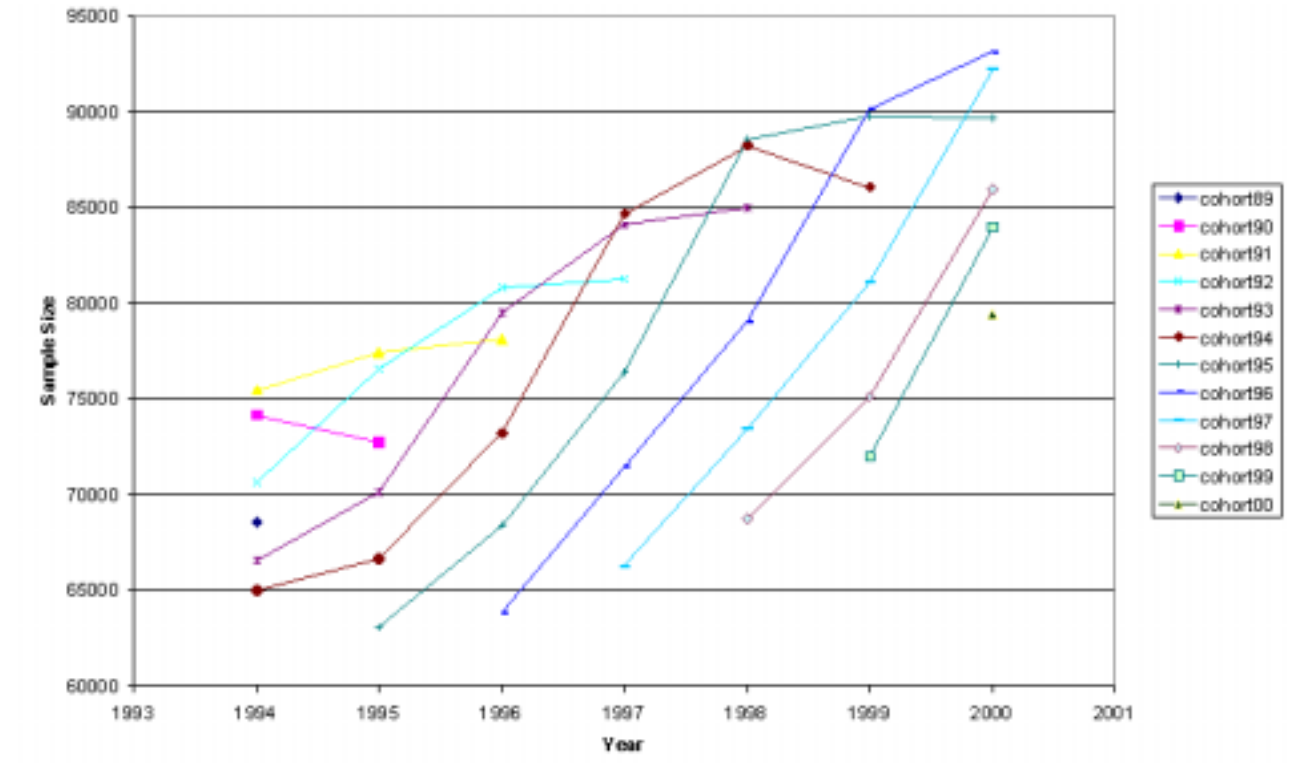


Figure 8.11. Number of Black students tested by cohort and year.

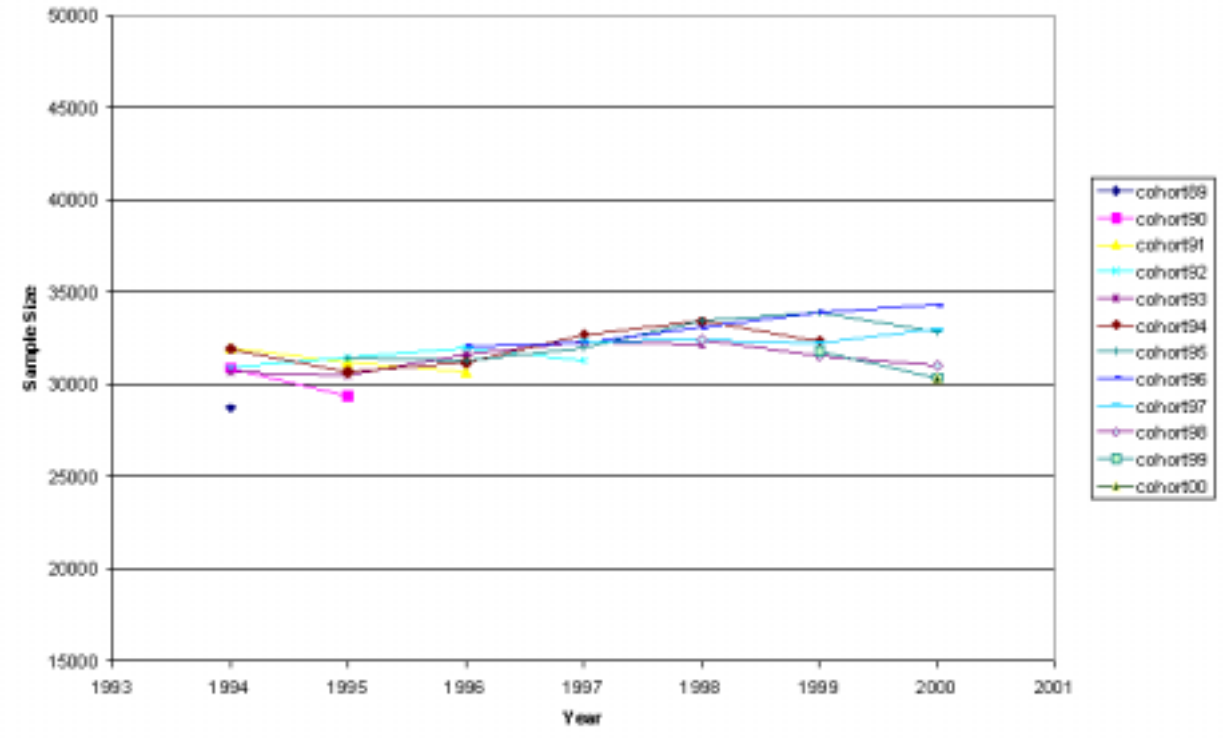


Figure 8.12. Number of nondisadvantaged students tested by cohort and year.

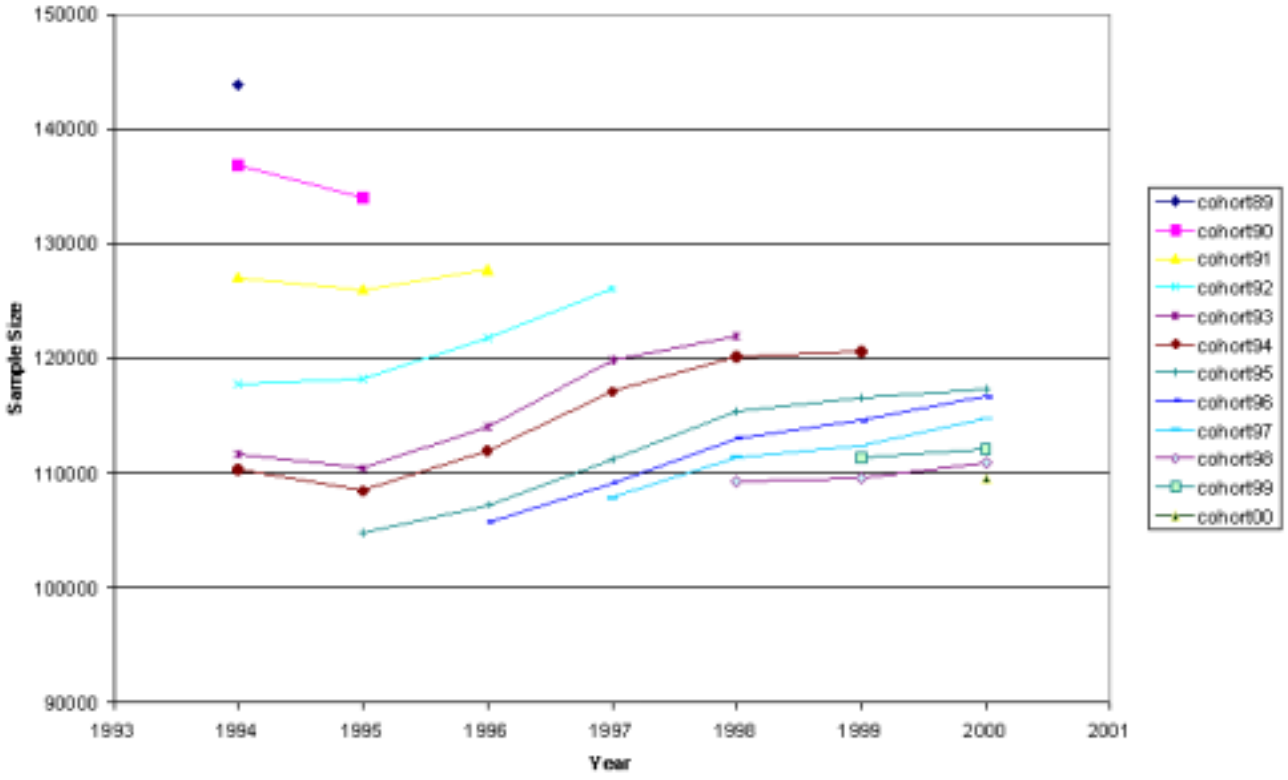
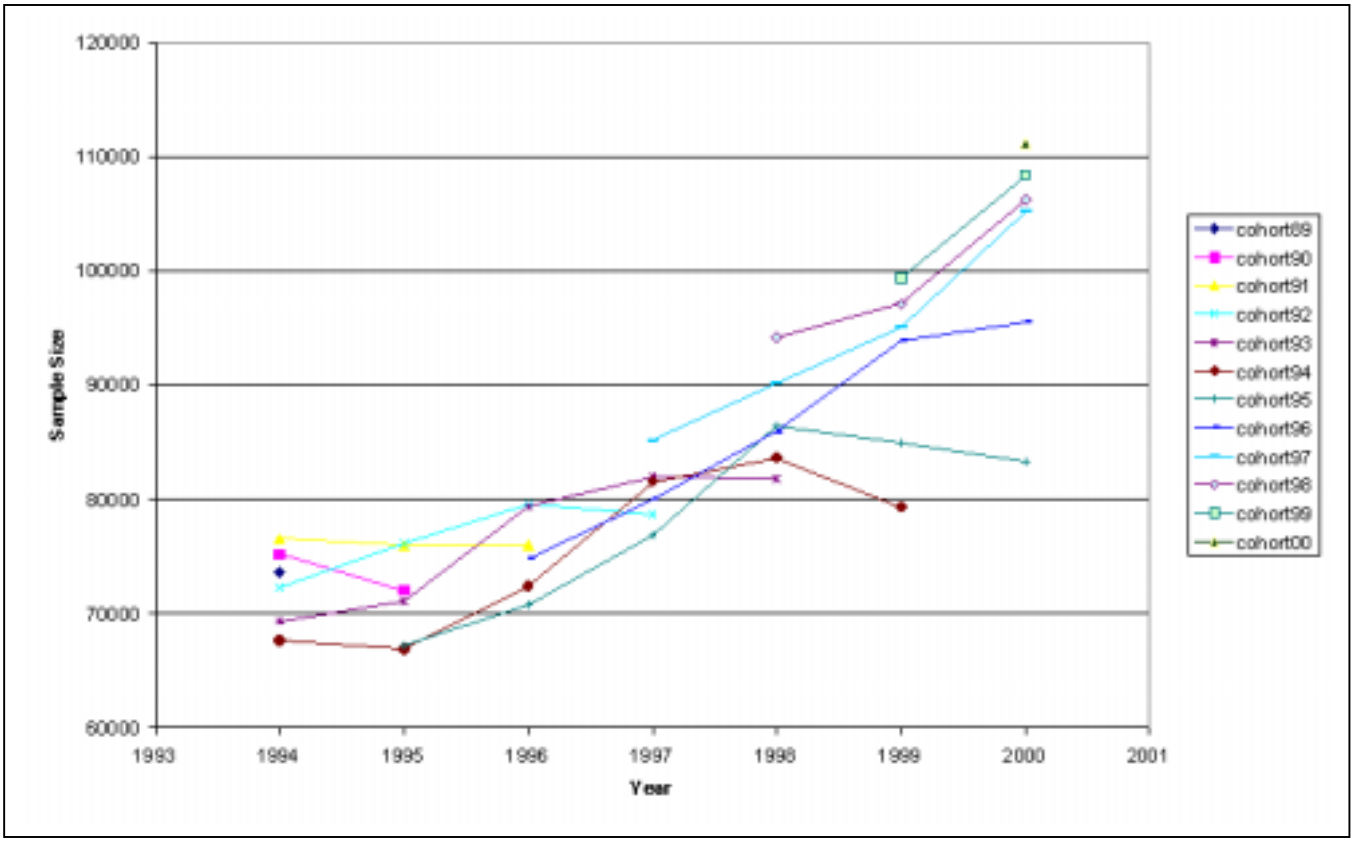


Figure 8.13. Number of disadvantaged students tested by cohort and year.



A semi-log model of this outcome is given by:

$$F_{gt} = \exp(\alpha_t + \beta_g + v_{2gt}) \tag{35}$$

where α_t is an effect parameter for year *t*, β is an effect parameter for grade *g*, and v_{2gt} is an error component. Combining the two models and taking the log of both sides of the equation yields a standard semi-log model of the number of students tested:

$$\ln N_{gt} = (\alpha_t + \phi_t) + \beta_g + \delta_c + v_{gt} \tag{36}$$

where $v_{gt} = v_{1gt} + v_{2gt}$.[19] One immediate consequence of defining the model in terms of students tested (rather than the components E_{gt} and F_{gt}) is that the year-specific parameters α_t and ϕ_t cannot be separately estimated. (Only their sum can be estimated.) If we assume that migration in and out of the state is relatively small, then the estimated year-specific parameter may largely

[19] Some of the parameter symbols used in the above model are also used elsewhere in this chapter. The symbols mean different things in the different models.

reflect α_t. In any case, to simplify our notation, we will drop ϕ_t from the model. As a result, α_t should be interpreted as including year effects with respect to enrollment and fraction tested.

Apart from the identification problem discussed above, there is a larger identification problem that flows from the fact that cohort, year, and grade are fully interrelated. In fact, a student's cohort is completely defined given his or her year and grade: $c = t - (g - 3)$. This means that the parameters in (36) are not separately identified.[20] To develop a better understanding of what exactly can be estimated, let's consider a linear model defined in terms of the regressors year (*t*-1994), grade (*g*-3), and cohort (*c*-1989):

$$\ln N_{gt} = \mu + \alpha(t - 1994) + \beta(g - 3) + \delta(c - 1989) + v_{gt} \tag{37}$$

where μ is an intercept and α, β, and δ have been redefined as slope parameters rather than as fixed effects.[21] To obtain an estimable model, it is necessary to eliminate one of the variables from the model. We eliminate *g* by substituting in the identity $g = t - c + 3$, which yields:

$$\ln N_{gt} = (\mu + 5\beta) + (\alpha + \beta)(t - 1994) + (\delta - \beta)(c - 1989) + v_{gt}. \tag{38}$$

This specification indicates that the regression parameter on time *t* captures the combined effects of year and grade (α-β). The regression parameter on cohort *c* captures the cohort effect minus the grade effect (δ-β). Estimates of this model (for different demographic groups) are discussed below.

A More General Model

The above linear model is useful in that it summarizes patterns over time and across cohorts with a limited number of parameters. In this section, in order to follow a linear trend, we consider a more general model that does not restrict changes over time, grades, and cohorts. Although we do not present estimates from the general model in this study, we describe it below since it poses some interesting analytic problems and may be useful in subsequent research. Readers interested in our empirical results may skip this section.

To obtain an estimable version of the general model (see equation (36)), we again need to eliminate a single variable from the model. Before we do this, however, it is helpful to rewrite (36) in the following (equivalent) form:

$$\begin{aligned} \ln N = \mu &+ \alpha_{95}T_{95} + \alpha_{96}T_{96} + \alpha_{97}T_{97} + \alpha_{98}T_{98} + \alpha_{99}T_{99} + \alpha_{00}T_{00} \\ &+ \beta_4 G_4 + \beta_5 G_5 + \beta_6 G_6 + \beta_7 G_7 + \beta_8 G_8 \\ &+ \delta_{90}C_{90} + \delta_{91}C_{91} + \ldots + \delta_{00}C_{00} + v \end{aligned} \tag{39}$$

[20] These parameters cannot be estimated given data on the number of students tested. They could be estimated given data on enrollments and the fraction of students tested.

[21] The regressors are expressed as the deviation from the base period (year = 1994, grade = 3, and cohort = 1989) so that the intercept can be interpreted as the predicted log of the number of students tested in the base period.

where the T_t, G_g, and C_c are indicator variables and the subscripts gt have been dropped from the model to simplify the notation.[22] To obtain an estimable model, we eliminate G_4 by substituting the following identity:

$$\begin{aligned} G_4 = 5 + 1Y_{95} + 2Y_{96} + 3Y_{97} + 4Y_{98} + 5Y_{99} + 6Y_{00} \\ -2G_5 - 3G_6 - 4G_7 - 5G_8 - 1C_{90} - 2C_{91} - 3C_{92} + ... - 10C_{99} - 11C_{00}. \end{aligned} \tag{40}$$

This yields the following model:

$$\begin{aligned} \ln N = (\mu + 5\beta_4) + (\alpha_{95} + \beta_4)T_{95} + (\alpha_{96} + 2\beta_4)T_{96} \\ +(\alpha_{97} + 3\beta_4)T_{97} + (\alpha_{98} + 4\beta_4)T_{98} + (\alpha_{99} + 5\beta_4)T_{99} + (\alpha_{00} + 6\beta_4)T_{00} \\ +(\beta_5 - 2\beta_4)G_5 + (\beta_6 - 3\beta_4)G_6 + (\beta_7 - 4\beta_4)G_7 + (\beta_8 - 5\beta_4)G_8 \\ +(\delta_{90} - \beta_4)C_{90} + (\delta_{91} - 2\beta_4)C_{91} + ... + (\delta_{99} - 10\beta_4)C_{99} + (\delta_{00} - 11\beta_4)C_{00} + v. \end{aligned} \tag{41}$$

This somewhat complicated looking equation is, in fact, very similar to the linear model considered above in that the year and cohort parameters are defined relative to a grade-level parameter. In the linear model, we add and subtract $\beta(t-1994)$ and $\delta(c-1989)$, respectively (where the slope parameter β is assumed to be the same at all grade levels). In the more general model, we add and subtract $\beta_4(t-1994)$ and $\beta_4(c-1989)$, respectively. The year and cohort parameters are measured relative to the grade parameter in grade 4. In the more general model, the included grade parameters are also measured relative to β_4.

The general model can also be written in a way that more directly highlights its link with the linear model:

$$\begin{aligned} \ln N = (\mu + 5\beta_4) + (\alpha_{95} + \beta_4)(t-1994) + (\delta_{90} - \beta_4)(c-1989) \\ +(\alpha_{96} - 2\alpha_{95})T_{96} + (\alpha_{97} - 3\alpha_{95})T_{97} + (\alpha_{98} - 4\alpha_{95})T_{98} \\ +(\alpha_{99} - 5\alpha_{95})T_{99} + (\alpha_{00} - 6\alpha_{95})T_{00} \\ +(\beta_5 - 2\beta_4)G_5 + (\beta_6 - 3\beta_4)G_6 + (\beta_7 - 4\beta_4)G_7 + (\beta_8 - 5\beta_4)G_8 \\ +(\delta_{91} - 2\delta_{90})C_{91} + (\delta_{92} - 3\delta_{90})C_{92} + ... + (\delta_{99} - 10\delta_{90})C_{99} + (\delta_{00} - 11\delta_{90})C_{00} + v. \end{aligned} \tag{42}$$

The first line of this equation is almost identical to the linear model. In (42), however, the slope coefficients are equal to effect (slope) parameters for the base periods (year = 1995, grade = 3, and cohort = 1989). All of the remaining parameters capture the extent to which the number of students tested at time *t*, grade *g*, and cohort *c* departs from the slope trajectories (with respect to time, grade, and cohort) in the base year. Model specifications (41) and (42) are fully equivalent ways of capturing the effects of year, grade, and cohort on the number of students tested.[23]

[22] The parameters α_{94}, β_3, and δ_{89} have implicitly been absorbed into the intercept. The parameters in the model should be understood to capture the effect of a given year, grade, or cohort relative to the omitted group: year = 1994, grade = 3, and cohort = 1989.

[23] In fact, given estimates of either model, the parameters of the other model can be derived directly.

Table 8.8 reports estimates of equation (38) for White, Hispanic, and Black students and disadvantaged and nondisadvantaged students. Recall that the coefficient on the time variable captures the combined effects of time and grade ($\alpha + \beta$) (for a given cohort) and the coefficient on the cohort variable captures the cohort effect minus the grade effect ($\delta - \beta$). For convenience, the table also reports the sum of these two coefficients ($\alpha + \delta$). This coefficient captures the effects of time and cohort changes on the number of students tested in a given grade. As indicated in the table, the linear models explain more than 75% of the variation (as measured by the R^2 statistic) in the number of students tested for Hispanic students and for disadvantaged and nondisadvantaged students. In contrast, the explanatory power of the model for Blacks is only 27%, but this doesn't really matter because there is almost no variation to explain—see Figure 8.9).

Table 8.8
Estimates of the Determinants of the Number of Students Tested by Racial/Ethnic Group and Disadvantaged Status

(Standard errors in parentheses.)

Parameter	Variable	White	Hispanic	Black
$\mu+5\beta$	Intercept	11.733	11.237	10.341
		(0.010)	(0.017)	(0.012)
$\alpha+\beta$	(t-1994)	0.007	0.075	0.010
		(0.003)	(0.005)	(0.004)
$\delta-\beta$	(c-1989)	-0.010	-0.038	-0.001
		(0.002)	(0.004)	(0.003)
$\alpha+\delta$	Derived	-0.003	0.037	0.009
R^2		0.324	0.831	0.270
Sample Size		42	42	42

Parameter	Variable	Non Disadvantaged	Disadvantaged
$\mu+5\beta$	Intercept	11.777	11.115
		(0.013)	(0.024)
$\alpha+\beta$	(t-1994)	0.022	0.048
		(0.004)	(0.008)
$\delta-\beta$	(c-1989)	-0.033	0.010
		(0.003)	(0.006)
$\alpha+\delta$	Derived	-0.011	0.059
R^2		0.762	0.773
Sample Size		42	42

The combined time and grade effects are large for two groups: Hispanic students and disadvantaged students. With respect to Hispanic students, the estimates indicate that for a given cohort, the number of students tested increased by 7.5% per year, or by 45% over a five-year period. The comparable estimates for disadvantaged students are 4.8% per year and 27% over a five-year period. These effects are quite large. They indicate that Texas may have had policies and practices in place from 1994 to 2000 that led to systematic undertesting of Hispanic and disadvantaged students in earlier as opposed to later grades and nearer the beginning as opposed to the end of the period. Some part of the increase in the number of students tested could be due to positive net in-migration of students over this period, but we doubt that this could account for all of the increase. One of the implications of the above findings is that it would be highly problematic to limit models of achievement growth to students with multiple (longitudinal) data points over time. A restriction of this type would systematically exclude Hispanic and disadvantaged students, since these students were disproportionately likely to be excluded from testing in the early grades and years. The models used in this study require that a student have no more than two consecutive test scores.

The coefficient in the fourth row of Table 8.8 ($\alpha + \delta$) captures the effects of time and cohort changes on the number of students tested in a given grade. (See also Tables 8.1 to 8.6.) As indicated in Table 8.8, this coefficient is large for disadvantaged students (5.9%) and Hispanic students (3.7%) and very small for all other groups. This indicates that it could be quite misleading to compare average test scores (with respect to any scale, TLI or other) over time at given grade levels. Our analysis indicates that the demographic composition of these samples changed dramatically from 1994 to 2000. As a result, the models of achievement growth used in this study explicitly account for differences in achievement growth associated with demographic factors.

We next report estimates of the determinants of achievement growth in Texas over the period 1994-2000.

Estimates of Micro Models of Student Achievement Growth

In this section, we present estimates of micro (level-one) models of student achievement. Macro-model estimates, including estimates of the effectiveness of the Urban Systemic Initiative, are presented in the next section. Table 8.9 presents micro-model results for several different models. Separate results are reported by:

- Grades: 4-8.
- Test score scale: Texas Learning Index (TLI) and Rasch scale.
- Estimation method:
    - OLS: Ordinary least squares.
    - EV1: Errors and variables method based on exact finite sample estimates of the error variance.
    - EV2: Errors and variables method based on maximum-likelihood (asymptotic) estimates of the error variance.
    - IV: Instrumental variables method with twice-lagged achievement (and the other exogenous variables) used as instrumental variables.

- IV-H: Instrumental variables (as above) with weighting to control for heteroscedasticity in the variance of measurement error.

Table 8.9
The Effects of Prior Achievement and Demographic Characteristics by Grade, Scale Score, and Model

4th Grade

	TLI		Rasch Scale				
Variable	**OLS**	**EV1**	**OLS**	**EV1**	**EV2**	**IV**	**IV - H Model**
3rd Grade Achievement	0.603	0.698	0.630	0.801	0.868	na	na
	(0.0006)	(0.0007)	(0.0007)	(0.0009)	(0.0010)		
Female	-0.077	-0.085	-0.046	-0.102	-0.124	na	na
	(0.0146)	(0.0148)	(0.0205)	(0.0210)	(0.0214)		
Disadvantaged	-1.926	-1.404	-2.991	-1.844	-1.396	na	na
	(0.0211)	(0.0214)	(0.0296)	(0.0306)	(0.0313)		
Black	-2.172	-1.479	-3.449	-2.011	-1.449	na	na
	(0.0279)	(0.0283)	(0.0391)	(0.0403)	(0.0413)		
Hispanic	-0.313	0.023	-0.882	-0.104	0.200	na	na
	(0.0242)	(0.0244)	(0.0338)	(0.0347)	(0.0355)		
Other Ethnicity	1.496	1.263	3.228	2.577	2.323	na	na
	(0.0623)	(0.0629)	(0.0872)	(0.0893)	(0.0912)		
Mixed Report	-0.590	-0.314	-1.193	-0.572	-0.329	na	na
	(0.0640)	(0.0647)	(0.0897)	(0.0918)	(0.0938)		

5th Grade

	TLI		Rasch Scale				
Variable	**OLS**	**EV1**	**OLS**	**EV1**	**EV2**	**IV**	**IV - H Model**
4th Grade Achievement	0.680	0.756	0.702	0.847	0.916	0.986	0.987
	(0.0006)	(0.0006)	(0.0007)	(0.0008)	(0.0009)	(0.0014)	(0.0013)
Female	-0.324	-0.317	-0.334	-0.344	-0.348	-0.418	-0.426
	(0.0130)	(0.0131)	(0.0192)	(0.0195)	(0.0199)	(0.0240)	(0.0237)
Disadvantaged	-1.256	-0.840	-2.225	-1.207	-0.730	-0.169	-0.150
	(0.0190)	(0.0192)	(0.0280)	(0.0287)	(0.0294)	(0.0359)	(0.0352)
Black	-1.821	-1.289	-2.842	-1.593	-1.007	-0.165	-0.169
	(0.0248)	(0.0251)	(0.0365)	(0.0374)	(0.0383)	(0.0468)	(0.0457)
Hispanic	-0.377	-0.156	-0.913	-0.333	-0.061	0.338	0.346
	(0.0213)	(0.0214)	(0.0313)	(0.0320)	(0.0326)	(0.0397)	(0.0389)
Other Ethnicity	1.021	0.804	2.572	1.853	1.515	1.141	1.132
	(0.0522)	(0.0526)	(0.0768)	(0.0783)	(0.0799)	(0.0992)	(0.1007)
Mixed Report	-0.590	-0.390	-1.109	-0.598	-0.359	0.045	0.059
	(0.0551)	(0.0555)	(0.0810)	(0.0825)	(0.0842)	(0.0997)	(0.0984)

Table 8.9 (continued)
The Effects of Prior Achievement and Demographic Characteristics by Grade, Scale Score, and Model

6th Grade

	TLI		Rasch Scale				
Variable	**OLS**	**EV1**	**OLS**	**EV1**	**EV2**	**IV**	**IV - H Model**
5th Grade Achievement	0.679	0.753	0.699	0.838	0.891	0.935	0.934
	(0.0005)	(0.0006)	(0.0006)	(0.0007)	(0.0008)	(0.0011)	(0.0010)
Female	-0.598	-0.573	-0.617	-0.596	-0.588	-0.698	-0.738
	(0.0119)	(0.0120)	(0.0173)	(0.0176)	(0.0179)	(0.0216)	(0.0212)
Disadvantaged	-1.430	-1.027	-2.398	-1.405	-1.023	-0.721	-0.703
	(0.0174)	(0.0176)	(0.0254)	(0.0261)	(0.0266)	(0.0324)	(0.0315)
Black	-1.382	-0.851	-2.388	-1.153	-0.678	-0.028	-0.014
	(0.0226)	(0.0229)	(0.0329)	(0.0338)	(0.0345)	(0.0420)	(0.0403)
Hispanic	-0.684	-0.460	-1.342	-0.748	-0.519	-0.186	-0.171
	(0.0192)	(0.0194)	(0.0280)	(0.0286)	(0.0292)	(0.0355)	(0.0343)
Other Ethnicity	1.077	0.861	2.331	1.577	1.287	1.044	1.070
	(0.0459)	(0.0463)	(0.0668)	(0.0682)	(0.0695)	(0.0847)	(0.0891)
Mixed Report	-0.650	-0.442	-1.197	-0.670	-0.467	-0.217	-0.199
	(0.0498)	(0.0502)	(0.0723)	(0.0738)	(0.0752)	(0.0881)	(0.0878)

7th Grade

	TLI		Rasch Scale				
Variable	**OLS**	**EV1**	**OLS**	**EV1**	**EV2**	**IV**	**IV - H Model**
6th Grade Achievement	0.781	0.859	0.800	0.946	0.991	1.046	1.036
	(0.0005)	(0.0006)	(0.0006)	(0.0007)	(0.0008)	(0.0011)	(0.0010)
Female	-0.249	-0.183	-0.335	-0.241	-0.212	-0.379	-0.335
	(0.0120)	(0.0121)	(0.0171)	(0.0175)	(0.0177)	(0.0216)	(0.0207)
Disadvantaged	-1.590	-1.158	-2.434	-1.374	-1.043	-0.679	-0.657
	(0.0177)	(0.0179)	(0.0252)	(0.0259)	(0.0263)	(0.0325)	(0.0305)
Black	-1.596	-1.065	-2.344	-1.091	-0.701	0.048	0.014
	(0.0229)	(0.0232)	(0.0327)	(0.0335)	(0.0341)	(0.0419)	(0.0388)
Hispanic	-0.713	-0.406	-1.211	-0.427	-0.182	0.312	0.281
	(0.0194)	(0.0196)	(0.0277)	(0.0283)	(0.0288)	(0.0353)	(0.0331)
Other Ethnicity	1.220	0.988	2.961	2.184	1.942	1.805	1.717
	(0.0453)	(0.0457)	(0.0646)	(0.0660)	(0.0670)	(0.0829)	(0.0881)
Mixed Report	-0.627	-0.399	-0.967	-0.375	-0.191	0.240	0.216
	(0.0518)	(0.0523)	(0.0738)	(0.0754)	(0.0765)	(0.0891)	(0.0882)

Table 8.9 (continued)
The Effects of Prior Achievement and Demographic Characteristics by Grade, Scale Score, and Model

8th Grade

	TLI		Rasch Scale				
Variable	**OLS**	**EV1**	**OLS**	**EV1**	**EV2**	**IV**	**IV - H Model**
7th Grade Achievement	0.753	0.827	0.741	0.862	0.880	0.923	0.928
	(0.0005)	(0.0006)	(0.0005)	(0.0006)	(0.0007)	(0.0009)	(0.0009)
Female	0.287	0.345	0.428	0.502	0.514	0.352	0.394
	(0.0121)	(0.0121)	(0.0159)	(0.0162)	(0.0163)	(0.0194)	(0.0189)
Disadvantaged	-0.979	-0.548	-1.505	-0.597	-0.459	-0.075	-0.061
	(0.0176)	(0.0178)	(0.0233)	(0.0238)	(0.0240)	(0.0292)	(0.0278)
Black	-1.203	-0.632	-1.878	-0.721	-0.545	0.145	0.157
	(0.0230)	(0.0233)	(0.0304)	(0.0311)	(0.0313)	(0.0378)	(0.0356)
Hispanic	-0.778	-0.436	-1.236	-0.491	-0.377	0.125	0.105
	(0.0195)	(0.0197)	(0.0257)	(0.0263)	(0.0265)	(0.0318)	(0.0304)
Other Ethnicity	0.986	0.747	2.343	1.613	1.502	1.190	1.214
	(0.0448)	(0.0451)	(0.0592)	(0.0603)	(0.0607)	(0.0726)	(0.0788)
Mixed Report	-0.646	-0.393	-1.029	-0.470	-0.385	-0.035	-0.057
	(0.0556)	(0.0560)	(0.0734)	(0.0748)	(0.0752)	(0.0853)	(0.0866)

The estimates are based on separate models at each grade level, with data pooled (at each grade level) across all years: 1995 to 2000.[24] However, separate district school effects are included for every year. (Tables 8.10 to 8.14 report micro model estimates based on separate models by grade and year.) The pooled estimates are based on samples of approximately 1.2 million observations. Since the parameter estimates are all made with great precision (small standard errors), we will not discuss whether the estimates are statistically significant. They are. Nonetheless, estimates of standard errors are reported in parentheses.

Let's begin by discussing the estimates of the Rasch model in 5th grade (second panel in Table 8.9). In particular, focus on the coefficients on prior achievement and economic disadvantage for the alternative-estimation methods, reported below for convenience.

Method of Estimation	**Pretest**	**Economic Disadvantage**
OLS	0.702	-2.225
EV1	0.847	-1.207
EV2	0.916	-0.730
IV	0.986	-0.169
IV-H	0.987	-0.150

[24] Since the value-added models used in this study all require pretest data, models cannot be estimated for the first year (1994) and for the earliest grade (3rd grade). In addition, the IV and IV-H models require data on twice-lagged achievement. These models cannot be estimated for 3rd and 4th grades and for the years 1994 and 1995.

Table 8.10

The Effects of Prior Achievement and Demographic Characteristics in 4th Grade by Year and Scale Score

EV1 Model

Texas Learning Index

Variable	Pooled	1995	1996	1997	1998	1999	2000
3rd Grade Achievement	0.70	0.74	0.70	0.67	0.69	0.65	0.73
Female	-0.09	0.05	0.44	0.13	-0.24	-0.51	-0.36
Disadvantaged	-1.40	-1.90	-1.87	-1.66	-1.33	-1.04	-0.85
Black	-1.48	-1.92	-1.87	-1.74	-1.17	-1.52	-0.43
Hispanic	0.02	-0.26	-0.22	-0.10	0.09	0.34	0.37
Other Ethnicity	1.26	1.51	1.54	1.16	1.30	1.28	0.83
Mixed Report	-0.31	-0.54	-0.44	-0.19	-0.11	-0.30	-0.28

Rasch Scale

Variable	Pooled	1995	1996	1997	1998	1999	2000
3rd Grade Achievement	0.80	0.85	0.82	0.74	0.77	0.78	0.87
Female	-0.10	0.15	0.56	0.34	-0.25	-0.82	-0.64
Disadvantaged	-1.84	-2.06	-2.22	-2.07	-1.85	-1.53	-1.44
Black	-2.01	-2.27	-2.18	-2.27	-1.83	-2.13	-0.96
Hispanic	-0.10	-0.50	-0.42	-0.27	-0.12	0.41	0.45
Other Ethnicity	2.58	2.13	2.96	2.48	2.65	3.11	2.16
Mixed Report	-0.57	-0.72	-0.77	-0.55	-0.34	-0.30	-0.69
Sample	1191841	195851	195602	196328	202298	200186	201576

OLS Model

Texas Learning Index

Variable	Pooled	1995	1996	1997	1998	1999	2000
3rd Grade Achievement	0.60	0.63	0.61	0.60	0.61	0.54	0.60
Female	-0.08	0.04	0.40	0.11	-0.25	-0.46	-0.23
Disadvantaged	-1.93	-2.61	-2.48	-2.08	-1.76	-1.52	-1.39
Black	-2.17	-2.90	-2.59	-2.20	-1.67	-2.18	-1.42
Hispanic	-0.31	-0.77	-0.59	-0.35	-0.15	0.03	0.01
Other Ethnicity	1.50	1.78	1.77	1.34	1.50	1.51	1.11
Mixed Report	-0.59	-0.99	-0.71	-0.34	-0.32	-0.59	-0.68

Rasch Scale

Variable	Pooled	1995	1996	1997	1998	1999	2000
3rd Grade Achievement	0.63	0.68	0.66	0.60	0.61	0.59	0.64
Female	-0.05	0.15	0.52	0.33	-0.26	-0.66	-0.30
Disadvantaged	-2.99	-3.19	-3.43	-3.11	-2.99	-2.68	-2.68
Black	-3.45	-3.75	-3.54	-3.38	-3.07	-3.64	-3.08
Hispanic	-0.88	-1.35	-1.21	-0.91	-0.80	-0.40	-0.51
Other Ethnicity	3.23	2.70	3.59	3.03	3.33	3.76	2.98
Mixed Report	-1.19	-1.44	-1.36	-0.95	-0.90	-1.01	-1.56
Sample	1191841	195851	195602	196328	202298	200186	201576

Table 8.11
The Effects of Prior Achievement and Demographic Characteristics in 5th Grade by Year and Scale Score

EV1 Model

Texas Learning Index

Variable	Pooled	1995	1996	1997	1998	1999	2000
4th Grade Achievement	0.76	0.79	0.79	0.73	0.74	0.72	0.72
Female	-0.32	-0.14	-0.08	-0.53	-0.85	0.16	-0.42
Disadvantaged	-0.84	-1.43	-0.84	-0.95	-0.73	-0.78	-0.52
Black	-1.29	-1.96	-1.81	-1.20	-0.65	-1.62	-0.57
Hispanic	-0.16	-0.68	-0.33	-0.08	0.18	-0.04	0.05
Other Ethnicity	0.80	0.83	0.95	0.88	1.25	0.49	0.58
Mixed Report	-0.39	-0.82	-0.54	-0.30	-0.05	-0.30	-0.36

Rasch Scale

Variable	Pooled	1995	1996	1997	1998	1999	2000
4th Grade Achievement	0.85	0.87	0.85	0.81	0.85	0.86	0.86
Female	-0.34	0.00	0.02	-0.56	-1.27	0.49	-0.70
Disadvantaged	-1.21	-1.65	-1.06	-1.29	-1.00	-1.35	-0.94
Black	-1.59	-2.04	-1.76	-1.59	-0.89	-2.35	-0.87
Hispanic	-0.33	-0.93	-0.44	-0.39	0.17	-0.46	0.05
Other Ethnicity	1.85	1.59	1.95	1.60	2.57	1.70	1.62
Mixed Report	-0.60	-0.97	-0.75	-0.59	-0.04	-0.55	-0.86
Sample	1235502	196677	203434	204943	211026	209899	209523

OLS Model

Texas Learning Index

Variable	Pooled	1995	1996	1997	1998	1999	2000
4th Grade Achievement	0.68	0.71	0.72	0.66	0.67	0.63	0.63
Female	-0.32	-0.19	-0.08	-0.52	-0.85	0.15	-0.42
Disadvantaged	-1.26	-1.96	-1.35	-1.37	-1.10	-1.16	-0.84
Black	-1.82	-2.68	-2.47	-1.66	-1.05	-2.10	-1.04
Hispanic	-0.38	-1.02	-0.62	-0.29	0.01	-0.23	-0.10
Other Ethnicity	1.02	1.16	1.18	1.07	1.41	0.71	0.78
Mixed Report	-0.59	-1.11	-0.82	-0.47	-0.16	-0.50	-0.54

Rasch Scale

Variable	Pooled	1995	1996	1997	1998	1999	2000
4th Grade Achievement	0.70	0.74	0.72	0.68	0.70	0.70	0.68
Female	-0.33	-0.08	0.05	-0.52	-1.22	0.49	-0.69
Disadvantaged	-2.22	-2.52	-2.02	-2.29	-2.11	-2.44	-2.00
Black	-2.84	-3.20	-2.99	-2.67	-2.09	-3.66	-2.35
Hispanic	-0.91	-1.52	-1.05	-0.92	-0.38	-1.04	-0.52
Other Ethnicity	2.57	2.30	2.52	2.26	3.25	2.50	2.56
Mixed Report	-1.11	-1.45	-1.31	-1.04	-0.44	-1.11	-1.43
Sample	1235502	196677	203434	204943	211026	209899	209523

Table 8.12
The Effects of Prior Achievement and Demographic Characteristics in 8th Grade by Year and Scale Score

EV1 Model

Texas Learning Index

Variable	Pooled	1995	1996	1997	1998	1999	2000
5th Grade Achievement	0.75	0.77	0.70	0.77	0.76	0.76	0.77
Female	-0.57	-0.04	-0.94	-0.76	-0.76	-0.12	-0.83
Disadvantaged	-1.03	-1.26	-1.08	-1.08	-0.82	-1.00	-0.95
Black	-0.85	-1.93	-0.84	-0.22	-0.46	-1.32	-0.34
Hispanic	-0.46	-1.43	-0.77	0.07	-0.06	-0.52	-0.13
Other Ethnicity	0.86	0.69	0.93	1.11	1.13	0.52	0.85
Mixed Report	-0.44	-1.00	-0.60	-0.10	-0.15	-0.57	-0.37

Rasch Scale

Variable	Pooled	1995	1996	1997	1998	1999	2000
5th Grade Achievement	0.84	0.88	0.80	0.85	0.82	0.87	0.81
Female	-0.60	0.13	-1.02	-0.92	-0.94	0.58	-1.35
Disadvantaged	-1.41	-1.32	-1.38	-1.55	-1.11	-1.72	-1.36
Black	-1.15	-1.87	-1.04	-0.51	-0.51	-2.30	-0.61
Hispanic	-0.75	-1.63	-1.03	-0.13	-0.01	-1.49	-0.23
Other Ethnicity	1.58	1.01	1.63	2.04	2.19	0.92	1.66
Mixed Report	-0.67	-1.10	-0.74	-0.25	-0.22	-1.12	-0.70
Sample	1271374	200428	202551	212387	217953	218170	219885

OLS Model

Texas Learning Index

Variable	Pooled	1995	1996	1997	1998	1999	2000
5th Grade Achievement	0.68	0.69	0.63	0.70	0.69	0.68	0.69
Female	-0.60	-0.07	-0.98	-0.77	-0.79	-0.16	-0.81
Disadvantaged	-1.43	-1.79	-1.54	-1.52	-1.19	-1.34	-1.26
Black	-1.38	-2.65	-1.46	-0.81	-0.87	-1.72	-0.80
Hispanic	-0.68	-1.80	-1.05	-0.17	-0.23	-0.67	-0.30
Other Ethnicity	1.08	1.01	1.19	1.34	1.32	0.73	1.01
Mixed Report	-0.65	-1.35	-0.87	-0.33	-0.29	-0.68	-0.54

Rasch Scale

Variable	Pooled	1995	1996	1997	1998	1999	2000
5th Grade Achievement	0.70	0.75	0.68	0.73	0.69	0.71	0.66
Female	-0.62	0.11	-1.07	-0.91	-0.98	0.46	-1.27
Disadvantaged	-2.40	-2.20	-2.32	-2.53	-2.13	-2.79	-2.39
Black	-2.39	-3.03	-2.25	-1.78	-1.62	-3.53	-1.99
Hispanic	-1.34	-2.28	-1.64	-0.73	-0.54	-2.01	-0.86
Other Ethnicity	2.33	1.69	2.36	2.75	2.90	1.83	2.45
Mixed Report	-1.20	-1.71	-1.28	-0.81	-0.69	-1.54	-1.25
Sample	1271374	200428	202551	212387	217953	218170	219885

Table 8.13
The Effects of Prior Achievement and Demographic Characteristics in 7th Grade by Year and Scale Score

EV1 Model

Texas Learning Index

Variable	Pooled	1995	1996	1997	1998	1999	2000
6th Grade Achievement	0.86	0.95	0.99	0.89	0.89	1.08	0.90
Female	-0.18	0.75	-0.50	0.34	0.09	-0.26	-1.72
Disadvantaged	-1.16	-1.49	-1.01	-1.08	-1.29	-1.79	-1.59
Black	-1.07	-2.32	-0.95	-0.84	-1.62	-0.81	-0.02
Hispanic	-0.41	-1.77	-0.27	-0.06	-0.95	-0.42	0.88
Other Ethnicity	0.99	1.30	1.25	2.14	1.00	2.63	4.27
Mixed Report	-0.40	-1.10	-0.59	-0.27	-0.49	-0.15	0.22

Rasch Scale

Variable	Pooled	1995	1996	1997	1998	1999	2000
6th Grade Achievement	0.95	0.85	0.88	0.86	0.84	0.92	0.80
Female	-0.24	0.61	-0.52	0.24	-0.03	-0.46	-0.84
Disadvantaged	-1.37	-1.68	-0.92	-1.00	-1.10	-1.21	-1.10
Black	-1.09	-2.57	-0.90	-0.85	-1.40	-0.46	-0.33
Hispanic	-0.43	-1.79	-0.21	-0.09	-0.68	-0.08	0.30
Other Ethnicity	2.18	0.92	0.87	1.28	0.38	0.95	1.60
Mixed Report	-0.38	-1.39	-0.53	-0.26	-0.46	-0.07	0.02
Sample	1312269	203956	207982	213514	228390	228349	230078

OLS Model

Texas Learning Index

Variable	Pooled	1995	1996	1997	1998	1999	2000
6th Grade Achievement	0.78	0.77	0.80	0.78	0.77	0.83	0.72
Female	-0.25	0.50	-0.55	0.13	-0.08	-0.53	-0.87
Disadvantaged	-1.59	-2.16	-1.43	-1.47	-1.50	-1.65	-1.44
Black	-1.60	-3.28	-1.58	-1.38	-1.84	-0.91	-0.74
Hispanic	-0.71	-2.20	-0.63	-0.40	-0.91	-0.34	0.07
Other Ethnicity	1.22	1.24	1.13	1.52	0.59	1.20	1.77
Mixed Report	-0.63	-1.73	-0.87	-0.49	-0.65	-0.26	-0.11

Rasch Scale

Variable	Pooled	1995	1996	1997	1998	1999	2000
6th Grade Achievement	0.80	0.82	0.86	0.76	0.75	0.88	0.75
Female	-0.33	0.61	-0.50	0.15	-0.01	-0.42	-1.70
Disadvantaged	-2.43	-2.32	-1.89	-2.11	-2.37	-3.15	-2.75
Black	-2.34	-3.53	-2.10	-1.98	-2.76	-2.19	-1.40
Hispanic	-1.21	-2.52	-1.05	-0.79	-1.62	-1.26	0.01
Other Ethnicity	2.96	2.01	1.86	2.87	1.78	3.72	5.01
Mixed Report	-0.97	-1.72	-1.23	-0.81	-1.04	-0.80	-0.32
Sample	1312269	203956	207982	213514	228390	228349	230078

Table 8.14

The Effects of Prior Achievement and Demographic Characteristics in 8th Grade by Year and Scale Score

EV1 Model

Texas Learning Index

Variable	Pooled	1995	1996	1997	1998	1999	2000
7th Grade Achievement	0.83	0.84	0.87	0.87	0.84	0.79	0.61
Female	0.35	1.33	0.05	0.49	-0.09	-0.37	0.64
Disadvantaged	-0.55	-1.16	-0.22	-0.39	-0.47	-0.53	-1.08
Black	-0.63	-2.34	-0.44	-0.10	0.01	0.05	-1.41
Hispanic	-0.44	-2.04	-0.02	0.00	-0.34	0.00	-0.56
Other Ethnicity	0.75	0.63	0.87	1.32	0.71	0.86	0.71
Mixed Report	-0.39	-1.40	-0.21	-0.15	-0.23	-0.13	-0.58

Rasch Scale

Variable	Pooled	1995	1996	1997	1998	1999	2000
7th Grade Achievement	0.86	0.88	0.89	0.90	0.92	0.87	0.76
Female	0.50	1.56	-0.11	0.47	0.37	-0.40	1.19
Disadvantaged	-0.60	-0.85	-0.16	-0.38	-0.42	-0.65	-0.96
Black	-0.72	-1.97	-0.45	0.04	0.06	-0.12	-1.47
Hispanic	-0.49	-1.78	0.00	0.14	-0.42	-0.09	-0.59
Other Ethnicity	1.61	0.89	1.92	2.64	1.87	1.57	1.04
Mixed Report	-0.47	-1.18	-0.27	-0.09	-0.29	-0.27	-0.61
Sample	1300942	198652	206547	213692	222799	228496	230756

OLS Model

Texas Learning Index

Variable	Pooled	1995	1996	1997	1998	1999	2000
7th Grade Achievement	0.75	0.76	0.79	0.80	0.77	0.72	0.61
Female	0.29	1.23	0.01	0.43	-0.16	-0.39	0.64
Disadvantaged	-0.98	-1.65	-0.73	-0.85	-0.90	-0.90	-1.08
Black	-1.20	-3.09	-1.23	-0.70	-0.49	-0.43	-1.41
Hispanic	-0.78	-2.52	-0.51	-0.37	-0.63	-0.27	-0.56
Other Ethnicity	0.99	0.94	1.18	1.55	0.96	1.02	0.71
Mixed Report	-0.65	-1.78	-0.58	-0.46	-0.44	-0.34	-0.58

Rasch Scale

Variable	Pooled	1995	1996	1997	1998	1999	2000
7th Grade Achievement	0.74	0.76	0.79	0.79	0.79	0.74	0.64
Female	0.43	1.44	-0.13	0.41	0.25	-0.44	1.09
Disadvantaged	-1.51	-1.59	-0.94	-1.27	-1.43	-1.64	-1.99
Black	-1.88	-3.12	-1.62	-1.09	-1.09	-1.33	-2.52
Hispanic	-1.24	-2.55	-0.79	-0.62	-1.13	-0.83	-1.24
Other Ethnicity	2.34	1.56	2.55	3.24	2.70	2.24	1.98
Mixed Report	-1.03	-1.78	-0.83	-0.73	-0.81	-0.85	-1.08
Sample	1300942	198652	206547	213692	222799	228496	230756

The pretest coefficients vary enormously across the five different estimation methods, ranging from a low estimate of 0.702 for the OLS model up to 0.987 for the IV-H model. Similarly, the estimated effect of economic disadvantage ranges from a high (negative) estimate of -2.225 in the OLS model to a low of -0.150 (essentially no effect) in the IV-H model. It is evident that the choice of estimation method makes a huge difference. Which estimates should we prefer? The choice of estimation method is complicated by two factors. First, student achievement (whether measured by the Texas Learning Index or the Rasch scale) is measured with error. Second, the Texas dataset (like most datasets based on administrative records) includes only a limited number of control variables.

The first factor, if not addressed, causes downward bias in the coefficient on prior achievement and bias in the other parameter estimates. Fortunately, the biases due to measurement error can be eliminated using the errors-in-variables method (EV), provided that (correct) information on the magnitude of the error variance is available.[25] Earlier in the paper, we demonstrated that the finite-sample approach, in contrast to the traditional asymptotic (maximum likelihood) method, yields correct estimates of the measurement error variance. As a result, the approach that corrects for measurement error using finite-sample measures of the error variance (EV1) should be preferred over the alternative approach (EV2). The latter approach over-corrects for measurement error and, as indicated above (and in Table 8.9), produces high estimates of the pretest coefficient and low estimates of the coefficient on economic disadvantage.

The second factor gives rise to two different problems. One, failure to control for persistent individual/family attributes that affect achievement growth is likely to produce upward bias in the coefficient on prior achievement (since the omitted individual effects are likely to be positively correlated with prior achievement). As a result, EV1 estimates of the coefficient on prior achievement should be viewed as upper-bound estimates of the coefficient on prior achievement. It is tempting to speculate that the biases caused by measurement error (negative bias) and (unmeasured) persistent individual effects (positive bias) cancel each other out. In this case, OLS estimates of the coefficient on prior achievement would be unbiased. Unfortunately, there is no reason to assume this. I am inclined to believe that the OLS estimates are on the low side. If so, we can conclude that the true value of the pretest coefficient in 5^{th} grade lies between 0.702 (OLS) and 0.847 (EV1).

The second problem with weak control variables is that the (unmeasured) individual effects may be correlated with the school district indicators. In other words, the groups of students in different districts may differ with respect to unmeasured student/family characteristics. If so, estimation methods that yield unbiased pretest coefficients could paradoxically yield biased estimates of school district effects (and the effects of educational inputs and policies). One possible approach to reducing (or eliminating) this source of bias is to allow the pretest variable to serve, in effect, as a proxy for omitted individual effects. This suggests (as argued earlier in the paper) that the EV1 method might produce the best estimates of school effects (even though it would produce upward-biased estimates of the effects of prior

[25] The EV method yields parameter estimates that are consistent (asymptotically unbiased). Given the large size of the Texas data set, it is reasonable to assume that the EV estimates fully correct for the biases due to measurement error.

achievement). The bottom line is that the EV1 estimates are probably most useful if the objective of the analysis is to obtain valid (consistent) estimates of school district effects (and the effects of educational inputs and policies). Nonetheless, given the uncertainty that arises from having a weak set of control variables, we report district-level (level-two) results in the next section based on level-one estimates from both the OLS and EV1 approaches.

Finally, it is interesting to note that both models based on instrumental variables (IV and IV-H) produce estimates of pretest coefficients that exceed the upper-bound estimates given by the EV1 approach. This is evidently due to the fact that twice-lagged achievement is highly correlated with the omitted individual effect (and possibly a serially correlated error component) and thus is not a valid instrumental variable. The interesting point about these results is that they indicate at face value that there is essentially no effect on student achievement growth of being economically disadvantaged. This result is a technical artifact of using a model that produces a hugely upward-biased pretest coefficient (approximately equal to 1.0). The same result would be obtained from a linear growth model (a model that imposes the assumption that the pretest coefficient is equal to 1). At all grade levels, our analysis indicates that the upper-bound estimate (from the EV1 model) of the pretest coefficient is substantially less than 1 (about 0.85). Hence, one important methodological finding of this study is that linear growth models (quite widely used) would be inappropriate for the Texas achievement data.

In light of the above discussion, we will focus below on estimates obtained using the EV1 and OLS methods. As indicated in Table 8.9, the strongest demographic predictors of achievement growth are economic disadvantage and Black status. Both are negatively related to achievement growth. For example, with respect to the EV1 estimates for the Rasch scale, the effects of economic disadvantage and Black status are equal to -1.2 and -1.6, respectively. These effects are somewhat larger based on the OLS estimates. The effects of gender and Hispanic status are quite small at all grade levels.

The coefficient estimates should be interpreted as the effect of an explanatory variable on grade-to-grade growth in mathematics achievement as measured by the TLI or Rasch scale. An example of an explanatory variable is a demographic characteristic such as economic disadvantage or the USI program. Thus, a coefficient estimate of one corresponds to one point on the TLI or Rasch scale.

Next, we examine whether the effects of economic disadvantage and race/ethnicity varied over time—see Tables 8.10 to 8.14. To facilitate comparison of these effects, coefficient estimates based on the Rasch scale and EV1 estimation method are presented in Figures 8.14 (the effect of economic disadvantage), 8.15 (the effect of being Black), and 8.16 (the effect of being Hispanic).[26] Separate estimates are presented for each grade and the average effect (across grades) over time (the latter in bold). The effects of economic disadvantage (relative to

[26] The regression model results for the Rasch and TLI scales were quite similar. Hence, we report only the results for the Rasch scale. This is perhaps a surprising result given that the univariate statistics (see the means and standard deviations reported in Tables 8.1 to 8.5) for the two scales exhibited very different patterns. This difference is due to the fact that the regression models control for prior achievement and thus implicitly adjust for differences in the shape of the achievement growth curves produced by the Rasch and TLI scales. The TLI scale exhibits greater curvature over time than the Rasch scale. As a result, the posttest-on-pretest coefficients tend to be smaller in the models based on the TLI scale than the Rasch scale.

nondisadvantaged students) were relatively constant over time in grades 5, 6, and 7. This also was true for the average across all grades. The effect for the average hovered around –1.25 from 1996 to 2000. This indicates that grade-to-grade growth in mathematics achievement was, on the average, 1.25 points lower for economically disadvantaged students as compared to economically advantaged students. What is interesting are the differences in the change between 4th grade and 8th grade over time; the 4th grade effects decreased slightly (became less negative) and the 8th grade effects decreased from 1995 to 1996 and then gradually increased (became more negative) over the next four years. In general, as the number of disadvantaged students increased (Figure 8.13), the effect on achievement remained negative and constant over time, except in the 4th grade and 8th grade.

In contrast, the effects of Black status (relative to White status) varied somewhat by grade and over time (see Figure 8.15 for clarification). The average effect of Black status (across grades) improved from about –2.0 to –1.0 from 1995 to 2000. In the intervening years, the effects differed quite a bit across grades, with a relatively large negative effect (about –2.2) in the 4th grade from 1995 to 1999 and no effect in the 8th grade from 1997 to 1999. Overall, the number of Black students tested in Texas remained constant (Figure 8.11). Black students performed lower than White students in 8th grade for three years.

Figure 8.14. The effect of being economically disadvantaged over time by grade, Rasch scale (EV1 Estimates).

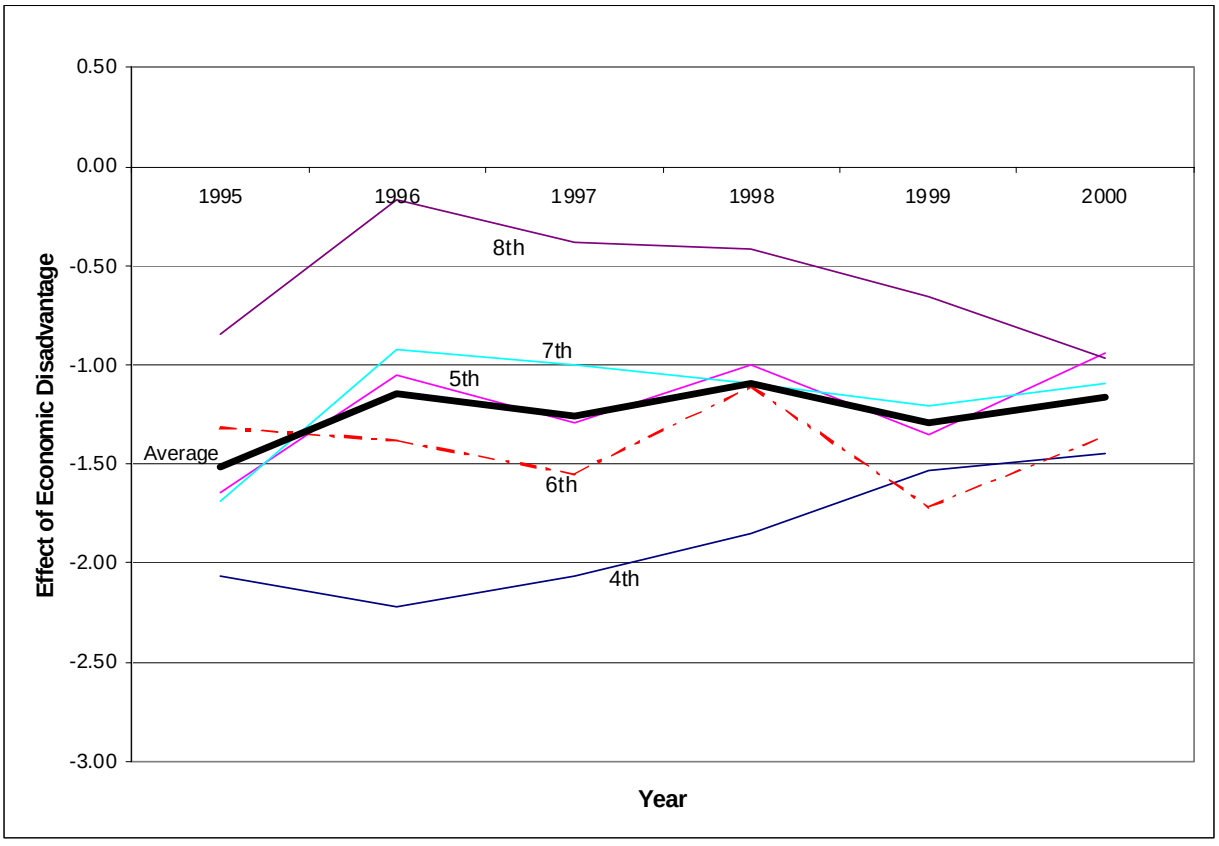


Figure 8.15. The effect of being African American over time by grade, Rasch scale (EV1 Estimates)

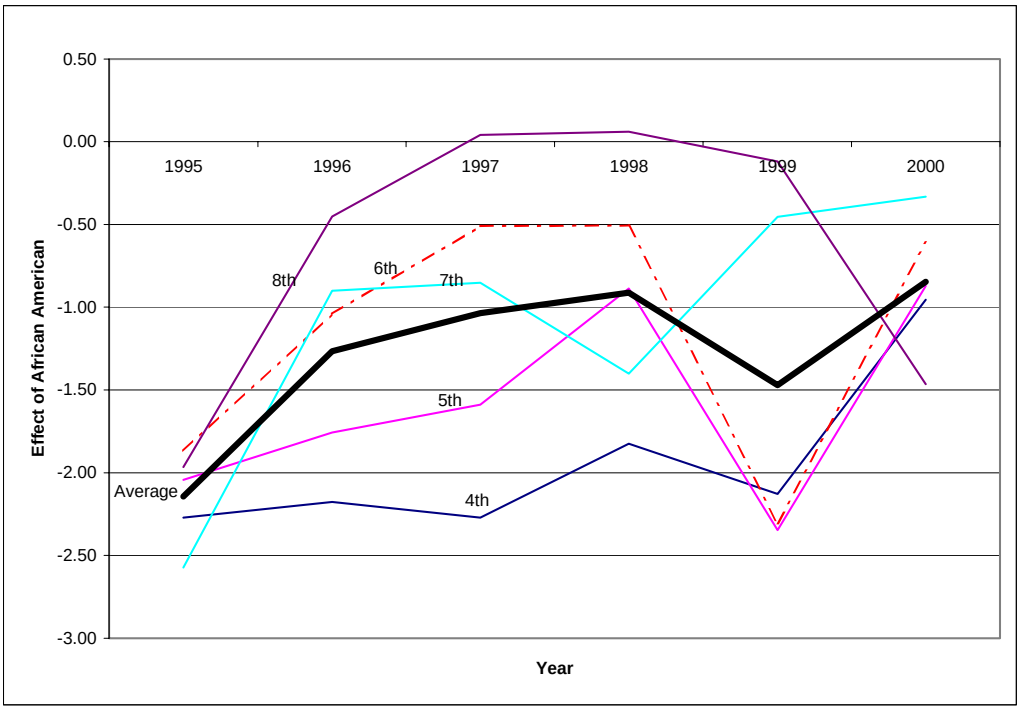
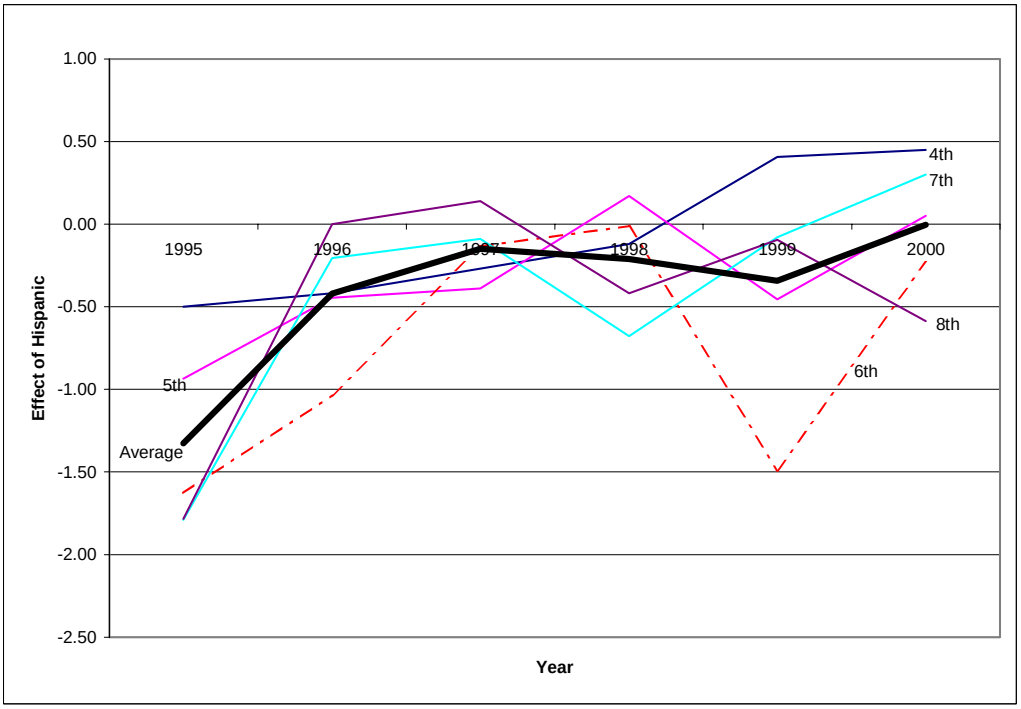


Figure 8.16. The effect of being Hispanic over time by grade, Rasch scale (EV1 Estimates).



Although the effects of Hispanic status (relative to White status) were generally quite small—typically, between 0.0 and –0.5—at all grades and in all years, the 4th grade Hispanic students had a more positive effect when compared to students in the other grades in 1999 and 2000. This is noticeably different from the disadvantaged and Black analyses. The first year (1995) was somewhat of an exception: the effects varied from –0.5 in 4th grade to –1.8 in 8th grade. Even though the number of Hispanic students tested increased significantly over time (Figure 8.10), the average performance of Hispanic students from 1997 to 2000 was nearly comparable to White students. The adjusted gain scores by 4th grade Hispanic students in 1999 and 2000 actually were slightly higher than those of 4th grade White students. Looking across the complete set of coefficient estimates by year, it appears that the coefficients are reasonably stable over time. Nonetheless, there was a general rise in performance by Black and Hispanic students after 1995. The (negative) Black effect declined by about 50% from 1995 to 2000, and the small, but negative Hispanic effect (averaged across all grades) essentially vanished by 2000. However, there was little improvement by economically disadvantaged students over the six years analyzed. Because there was some stability in results over time, it is reasonable to focus on the pooled (across years) estimates. In the next section, we draw on estimates of school district productivity derived from the pooled data to assess the effectiveness of the Urban Systemic Initiative (USI).

Estimates of the Effectiveness of the Urban Systemic Initiative

Tables 8.15 to 8.18 present estimates of the effectiveness of the Urban Systemic Initiative (USI) based on the level-two model discussed earlier in the paper. Separate USI effect estimates are presented based on level-one estimates from the OLS and EV1 approaches and the Rasch and TLI scales. The USI effect estimates are reported by grade and year. As indicated in the top panel in each table, the standard errors associated with each effect estimate (by grade and year) are quite large. This is due to the fact that the number of districts that participated in the USI was relatively small. As a result, it is best to interpret the USI estimates as a set rather than focus on estimates in any particular grade and year. The separate estimates by grade and years are in general not statistically significant. Nonetheless, it is apparent from inspecting the four tables that the USI appears to have contributed modestly to student achievement growth. Moreover, the USI effect appears to be have increased over time, from essentially no effect during the initial years of the program to modest, positive effects in 1999 and 2000. One interesting result is that the USI effect estimates appear to be generally larger in the early rather than later grades.

The above results—small but positive effects at the district level—indicate that it would be interesting to investigate the effectiveness of the USI at the school (and possibly teacher) level. It is possible that the program may have been implemented with varying degrees of success across schools. In order to conduct such an analysis, it would be necessary to draw on school (and possibly teacher) level data to implement a three- (or four-) level model of achievement, where, as above, level one is the student level, level two is the school (or teacher level), and level three is the district level. It would be straightforward to implement such a model using the methods developed in this study.

Table 8.15
Effectiveness of the Urban Systemic Initiative by Grade and Year for the TLI Scale and OLS Model

TLI Scale, OLS Model

Grade	1995	1996	1997	1998	1999	2000
4	0.565	0.728	0.765	0.868	1.533	1.228
	(1.327)	(1.328)	(0.799)	(0.799)	(0.799)	(0.800)
5	-0.323	-0.239	0.702	0.669	0.977	0.355
	(1.212)	(1.213)	(0.730)	(0.729)	(0.730)	(0.730)
6	-0.874	0.373	0.505	0.518	0.810	0.720
	(1.031)	(1.032)	(0.621)	(0.621)	(0.621)	(0.621)
7	-1.871	0.734	0.541	0.251	-0.034	0.512
	(0.945)	(0.947)	(0.570)	(0.570)	(0.570)	(0.571)
8	-1.942	1.730	0.037	0.482	0.759	1.211
	(0.954)	(0.955)	(0.575)	(0.575)	(0.575)	(0.575)

Note: Standard Errors in Parentheses

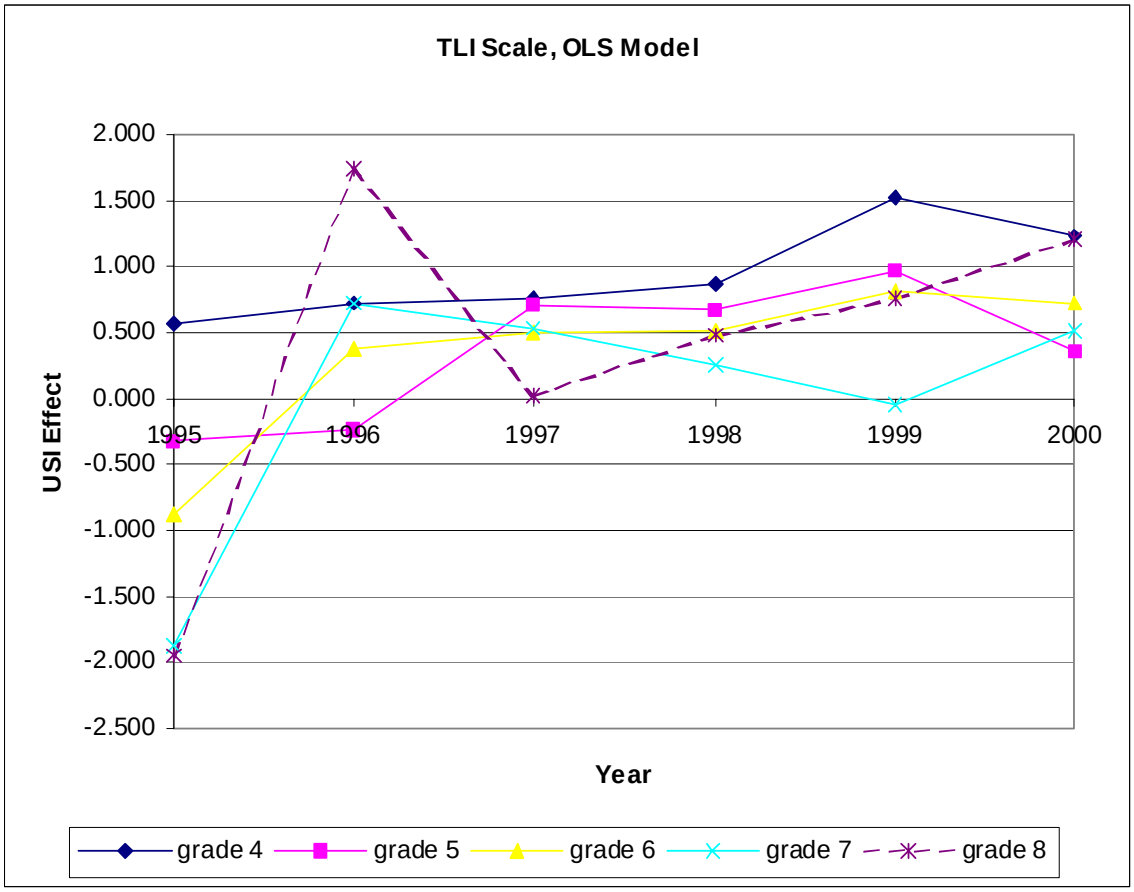


Table 8.16

Effectiveness of the Urban Systemic Initiative by Grade and Year for the Rasch Scale and OLS Model

Rasch Scale OLS Model

Grade	1995	1996	1997	1998	1999	2000
4	0.465	0.593	0.982	1.020	1.995	1.554
	(1.679)	(1.681)	(1.012)	(1.012)	(1.012)	(1.013)
5	-0.231	-0.164	0.810	0.773	1.387	0.389
	(1.611)	(1.612)	(0.971)	(0.970)	(0.971)	(0.972)
6	-1.431	-0.512	-0.018	0.126	0.445	0.191
	(1.478)	(1.480)	(0.890)	(0.891)	(0.890)	(0.891)
7	-1.601	0.715	0.808	0.311	-0.747	0.017
	(1.331)	(1.334)	(0.803)	(0.803)	(0.803)	(0.804)
8	-1.617	1.803	-0.252	0.051	0.280	1.004
	(1.143)	(1.145)	(0.690)	(0.690)	(0.690)	(0.691)

Note: Standard Errors in Parentheses

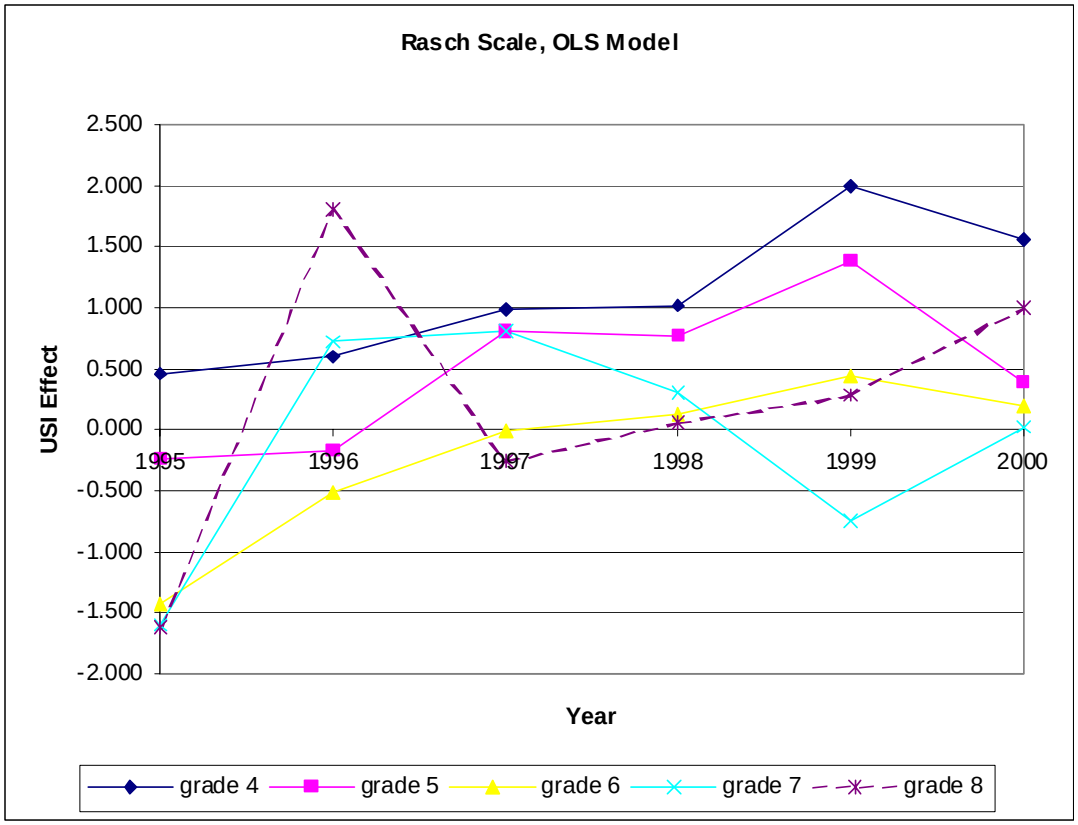


Table 8.17

Effectiveness of the Urban Systemic Initiative by Grade and Year fpr the TLI Scale and EV1 Model

TLI Scale, EV1 Model

Grade	1995	1996	1997	1998	1999	2000
4	0.318	0.599	0.646	0.740	1.389	0.977
	(1.379)	(1.380)	(0.830)	(0.830)	(0.830)	(0.831)
5	-0.439	-0.372	0.662	0.585	0.862	0.177
	(1.256)	(1.256)	(0.756)	(0.755)	(0.756)	(0.756)
6	-0.960	0.305	0.433	0.432	0.678	0.542
	(1.067)	(1.067)	(0.642)	(0.642)	(0.642)	(0.643)
7	-2.018	0.722	0.457	0.153	-0.146	0.341
	(0.978)	(0.980)	(0.590)	(0.590)	(0.590)	(0.590)
8	-2.053	1.767	-0.046	0.378	0.663	1.117
	(0.979)	(0.981)	(0.590)	(0.590)	(0.590)	(0.591)

Note: Standard Errors in Parentheses

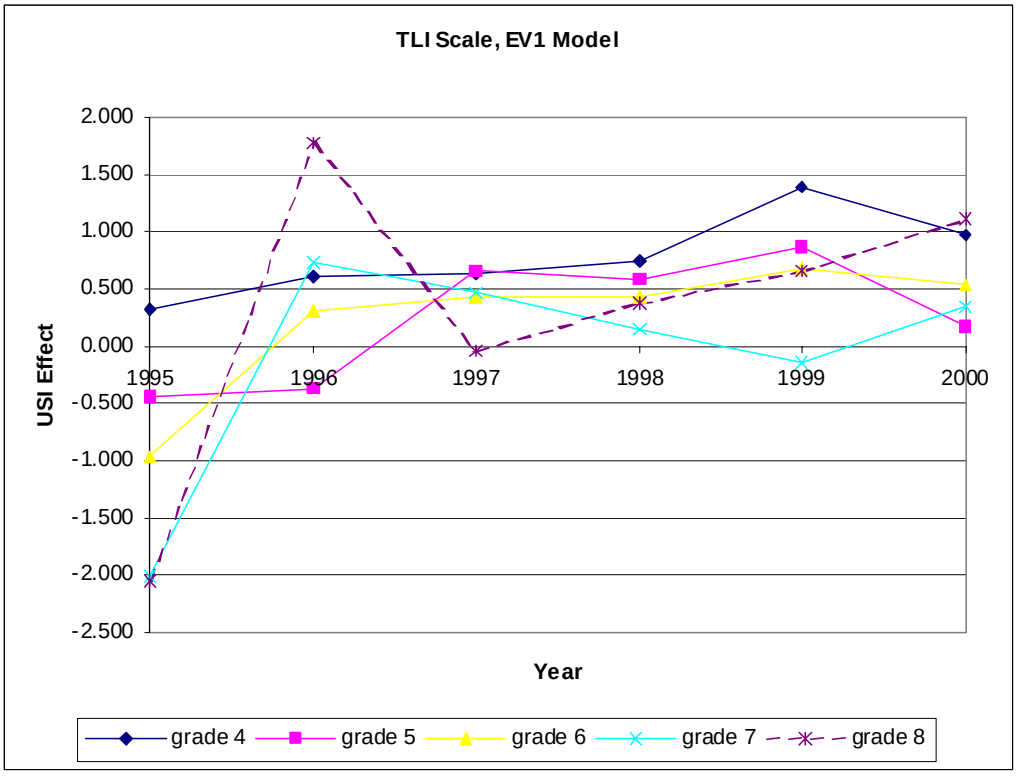
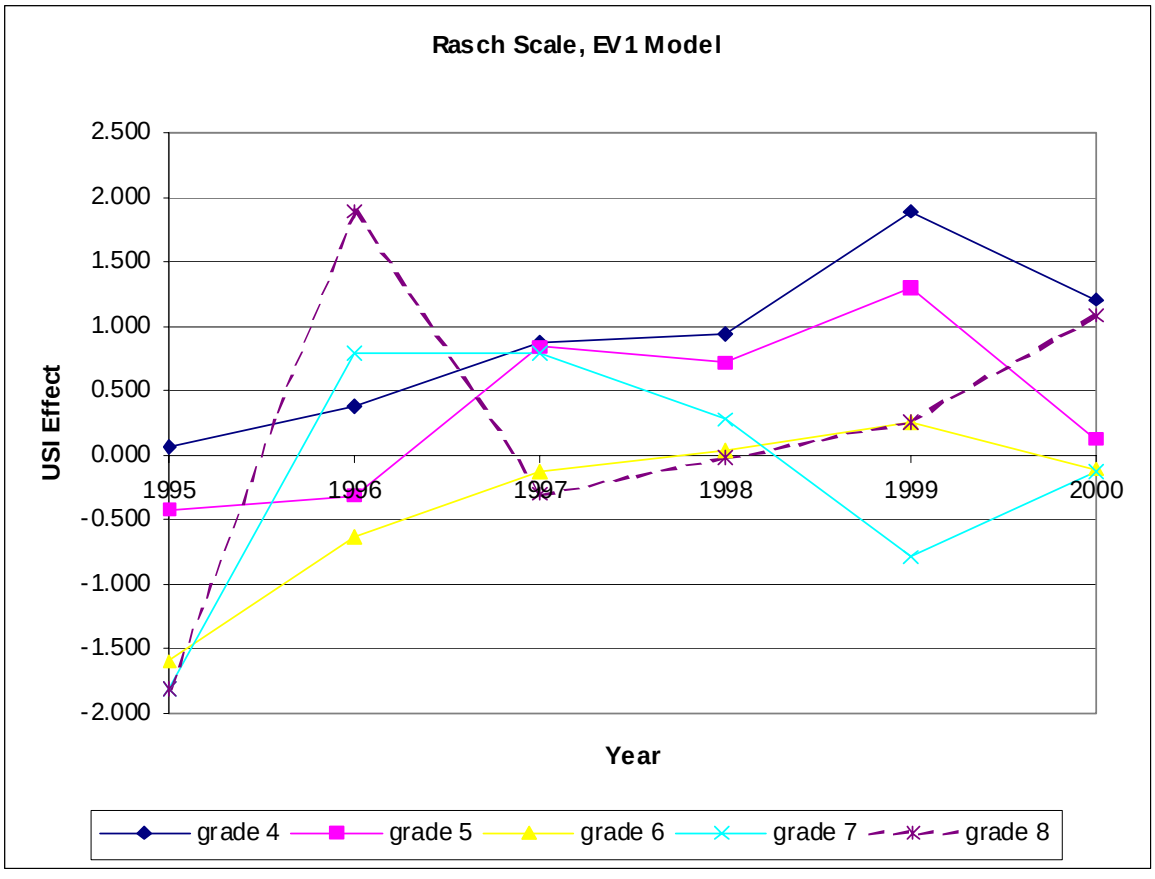


Table 8.18
Effectiveness of the Urban Systemic Initiative by Grade and Year for the Rasch Scale and EV1 Model

Rasch Scale, EV1 Model

Grade	1995	1996	1997	1998	1999	2000
4	0.059	0.375	0.880	0.934	1.881	1.212
	(1.793)	(1.795)	(1.080)	(1.080)	(1.080)	(1.081)
5	-0.422	-0.315	0.841	0.713	1.293	0.121
	(1.708)	(1.709)	(1.029)	(1.029)	(1.029)	(1.030)
6	-1.593	-0.640	-0.122	0.026	0.257	-0.114
	(1.557)	(1.558)	(0.937)	(0.938)	(0.937)	(0.938)
7	-1.803	0.789	0.791	0.285	-0.792	-0.124
	(1.418)	(1.422)	(0.855)	(0.855)	(0.855)	(0.856)
8	-1.810	1.896	-0.305	-0.027	0.266	1.079
	(1.202)	(1.204)	(0.725)	(0.725)	(0.725)	(0.726)

Note: Standard Errors in Parentheses



Conclusions

This study has developed a powerful tool set for analyzing the determinants of student achievement, particularly with respect to large data sets. These methods could be practically applied to analyze the state data sets that will hopefully be assembled as part of the new federal testing requirements.

This study directs attention to the fact that policy conclusions may be sensitive to the way in which test scores are scaled. We found that the TLI and Rasch scales frequently yielded quite different results. In subsequent research, we hope to develop statistical tests that will allow researchers to choose between alternative scales. A preferred scale is one that (in conjunction with a properly specified model) yields effect estimates that can be interpreted causally.

We demonstrated that it is important to employ statistical methods that correctly adjust for measurement error in achievement. Interestingly, we showed that the traditional method for estimating measurement error variances yields (based on asymptotic maximum likelihood formulas) quite biased estimates of measurement error at the extremes of the achievement distribution. Fortunately, it is feasible to apply finite sample methods to obtain unbiased estimates of measurement error variances with respect to Rasch scale estimates and test score scales based on this scale (for example, the TLI). In subsequent research, we plan to generalize the finite sample approach to a wider class of test theory models—in particular, the three-parameter logistic and two-parameter partial credit models. These latter models are frequently used to scale tests based on multiple choice and constructed-response items.

We also demonstrated that it is important to be careful when analyzing data with weak control variables. Based on our analysis, we concluded that it was best to present estimates based on two estimating methods (OLS and EV1) rather than on a single method. We argued that the true model estimates (given possible biases due to weak controls) were likely to be in the range of estimates given by the OLS and EV1 estimates.

References

Allen, M.J., & Yen, W.M. (1979). *Introduction to Measurement Theory*, Monterey, CA: Brooks/Cole Publishing Co.

Anderson, T. W., & Hsiao, C. (1981). Estimation of dynamic models with error components. *Journal of the American Statistical Association*, *76*, 598-606.

Anderson, T.W., & Hsiao, C. (1982). Formulation and estimation of dynamic models using panel data. *Journal of Econometrics*, *18*, 47-82.

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

Fuller, W. A. (1987). *Measurement error models.* New York: John Wiley & Sons.

Greene, W. H. (2000). *Econometric analysis* (4th ed.). Upper Saddle River, NJ: Prentice Hall.

Hanushek, E. A. (1974). Efficient estimators for regressing regression coefficients. *The American Statistician, 28*, 66-67.

Hsiao, C. (1986). *Analysis of panel data.* Cambridge: Cambridge University Press.

Kolen, M. J., Zeng, L., & Hanson, B. A. (1996). Conditional standard errors of measurement for scale scores using IRT. *Journal of Educational Measurement, 33*, 129-140.

Lee, W.-C., Brennan, R. L., & Kolen, M. J. (2000). Estimators of conditional scale-score standard errors of measurement: A simulation study. *Journal of Educational Measurement*, *37*, 1-20.

Lord, F. M. (1980). *Applications of item response theory.* Hillsdale, NJ: Erlbaum.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings." *Applied Psychological Measurement*, *31*, 453-461.

Meyer, R. H. (1992). *Applied versus traditional mathematics: New econometric models of the contribution of high school courses to mathematics proficiency* (Discussion Paper 966-92). Madison, WI: Institute for Research on Poverty.

Meyer, R. H. (1999). The production of mathematics skills in high school: What works? In S. E. Mayer, & P. E. Peterson (Eds.), *Earning and learning: How schools matter* (pp. 169-204). Washington, DC: Brookings Institution.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14*, 97-116.