
A Review and Analysis of the Tennessee Value-Added Assessment System

R. Darrell Bock, University of Chicago

Richard Wolfe, Ontario Institute for Studies in Education

Thomas H. Fisher, Florida Department of Education



STATE OF TENNESSEE
COMPTROLLER OF THE TREASURY

William R. Snodgrass
Comptroller

STATE CAPITOL
NASHVILLE, TENNESSEE 37219-5043
PHONE (615) 741-2501

April 3, 1996

The Honorable John S. Wilder, Lieutenant Governor
Speaker of the Senate
The Honorable Jimmy Naifeh
Speaker of the House of Representatives
The Honorable Andy Womack
Chairman of the Senate Education Committee
and
The Honorable Gene Davidson
Chairman of the House Education Committee

Gentlemen:

Transmitted herewith are the results of the evaluation of the Tennessee Value-Added Assessment System (TVAAS), contracted last year by the Office of Education Accountability. The evaluation is presented in two parts. The first was written by Dr. Darrell Bock of the University of Chicago and Richard Wolfe of the Ontario Institute for Studies in Education. The second part was written by Dr. Thomas Fisher of the Florida Department of Education. Brief biographies of the authors follow this letter.

A summary of the reports, prepared by the staff of the Office of Education Accountability, is available under separate cover. To obtain copies of the summary, please write or call the Office of Education Accountability, 500 Deaderick St., Suite 1360, Nashville, Tennessee 37243-0268, phone 615/532-1111.

Very truly yours,

W. R. Snodgrass
Comptroller of the Treasury



Comptroller of the Treasury, Office of Education Accountability. Authorization Number 307243, reprinted August 1996, 200 copies. This public document was promulgated at a cost of \$8.02 per copy.

The Consultants

R. Darrell Bock

R. Darrell Bock was a Professor of Behavioral Sciences and of Education at the University of Chicago from 1974-1993. He is presently a faculty fellow for the Division of Social Science at the University of Chicago and serves as the Senior Study Director of the Methodology Research Center at the National Opinion Research Center. He was employed as an assistant and associate professor at the University of North Carolina before going to the University of Chicago. He is a Visiting Fellow at the Institute of Education, University of London. His honors include the 1991 Educational Testing Service award for "Distinguished Contributions to Educational Measurement" and a 1990 award from the National Council on Measurement in Education for "Contributions to the Design and Analysis of Educational Assessment." He is a Fellow of the American Statistical Association and the Royal Statistical Society.

Richard G. Wolfe

Richard Wolfe is the Head of Computing for the Ontario Institute for Studies in Education and is cross-appointed with the Department of Curriculum. He holds a bachelor's degree and a master's degree in Mathematics from the University of Wisconsin and did doctoral work at the University of Chicago. He has served as a member of the Ontario Minister of Education's Council on Education and Technology and as an associate editor of *Evaluation Review*. Wolfe has served as the chairman of the sampling and methodology committee for the International Association for the Evaluation of Educational Assessment. He has worked for a decade on improvement of the teaching and learning of mathematics in the Dominican Republic.

Thomas H. Fisher

Thomas Fisher is the Program Director for the Student Assessment Services Section of the Florida Department of Education. He has served as the report reactor for the U.S. General Accounting Office on issues related to the establishment of performance standards for the National Assessment of Educational Progress. Dr. Fisher has served as a member of the Advisory Committee on Minimum Competency Testing, National Institute of Education; an advisor to the Southern Regional Education Board on the implementation of interstate achievement testing; an Advisor to the Council of Chief State School Officers on the design of a state-by-state achievement testing program; and served as a member of the Technical Advisory Committee for the Texas Academic Skills Project. He holds an Ed.D. degree from Wayne State University in Detroit, Michigan, and has done postdoctoral work in statistics and psychometrics at Florida State University.

Part 1

Audit and review of the Tennessee Value-Added Assessment System (TVAAS): Final Report

R. Darrell Bock, University of Chicago
Richard Wolfe, Ontario Institute for Studies in Education

March 15, 1996

Contents

1	Background	1
1.1	Comparison with results from the National Assessment Program (NAEP)	10
1.2	The TVAAS concept	10
2	Measurement issues	13
2.1	Are successive forms of the TCAP tests equally difficult?	15
2.2	Do standard deviations of students' IRT scale-scores vary systematically with age during the school years? Are achievement levels and achievement gains correlated in these years?	24
3	Data Quality Issues	29
3.1	Introduction	29
3.2	Comparison of the School Enrollment and Tested Student Populations (All Tennessee Data)	29
3.3	Summary of the Completeness of Student Testing Records (11-County Data)	31
3.4	Completeness of the Subtest Records (11-County Data) .	33
3.5	Percentage of Student Test Records with Teacher Assignment Information (11-County Data)	33
3.6	Discussion	36
4	Data Analysis Issues	39
4.1	The TVAAS Models	39
4.2	The Analysis	42

5 Reliability Issues	45
5.1 Teacher gain scores	48
5.2 School gain scores	50
5.3 Regressed estimates of teacher effects	53
5.4 Empirical evaluation of the stability of school and teacher gains estimated by TVAAS	56
6 Standards and Reporting Issues	61
6.1 Teachers and schools	62
6.2 School systems	63
7 Conclusions and Recommendations	69
7.1 Conclusions	69
7.2 Recommendations	72
Appendix A. CTB/McGraw-Hill construction of successive forms of the TCAP tests.	78
Appendix B. State Mean Scores Smoothed Over Cohorts	82

Preface

This report was prepared pursuant to a contract between the Tennessee Office of Education Accountability, Comptroller of the Treasury, and the Ontario Institute for Studies in Education for a technical review of the TVAAS data analysis procedures. The review is based on the descriptions of the procedures supplied to us in written documents and an oral presentation by Professor W. L. Sanders, University of Tennessee, who created and implemented the value-added assessment system. In addition, Dr. Fretta Bunch, Director of the State Testing Center, supplied information about the Tennessee Comprehensive Assessment Program (TCAP). Her agency conducts the state achievement testing and delivers the test scores to the University of Tennessee Value-Added Research and Assessment Center for data analysis. Accompanying the technical report is a further review, independently prepared by Dr. Thomas H. Fisher of the Florida State Bureau of School Improvement and Instruction, concerning contractual, legislative, and policy issues relating to TVAAS. Neither of these reviews deals with questions bearing on the educational content of the tests or their alignment with the State of Tennessee curricular guidelines: these are the responsibility of Tennessee Department of Education and are outside the scope of the contracted reviews.

Chapter 1

Background

To view the value-added assessment system in proper perspective, it is useful to examine the average performance of Tennessee school students on the TCAP tests since they were introduced in 1990. These tests, which are supplied in annually updated forms by the California Test Bureau/McGraw-Hill (CTB), are described in more detail in section 2. For present purposes, it is sufficient to point out that separate grade-level test booklets are supplied by CTB and administered by TCAP to all students in grades 2–8 of Tennessee public schools during the last weeks of the school year. The test booklets contain sections devoted to tests in five subject-matters: reading, language, math, science, and social studies. The tests in the first three subject areas contain two types of items—so-called “norm-referenced” items and “criterion-referenced” items. The science and social studies tests contain only norm-referenced items. TVAAS makes use of test scores on the norm-referenced items only. These scores are expressed on a special scale, constructed by CTB, that ranges from 0 to 999 and applies across all grade levels.

Because the successive grade-level tests are reported on a common scale, the scores can be used to measure a student's growth in achievement from one school grade to another. The availability of this type of scale for reporting test performance is essential to TVAAS, which is based on the measurement of annual gains (“value-added”) rather than on the test scores themselves. To distinguish between these two types of measures, we refer to the difference between a student's test score in a given grade and that in the previous grade as the *gain* for the given grade; the test score as such we call the *score-level*.

Another distinction we make is between the year when the testing takes

place and the groups of students who take particular grade-level tests; these groups are called *cohorts*. We will label the cohorts by the year in which the students enter second grade, corresponding to the years of testing 1990 through 1995. The test forms of the set of seven grade-level tests administered in these years will be labeled by their CTB designations, A through F.

The progress of learning achievement of Tennessee public-school students in grades 2–8 from 1990–1995, as reflected in the arithmetic average (mean) of their score-levels on the CTB norm-referenced tests, is shown in Tables 1.1 through 1.5 of this section. The rows in the upper section of the tables correspond to successive annual cohorts of students moving through the school system during the six years of testing. The columns correspond to the grade-level tests. The diagonals of the upper section correspond to the years of testing and also to the TCAP test forms supplied by CTB.

In addition, the bottom section of the table shows selected percentiles of the student score distribution for the U.S. based on a 1988 representative sample collected by CTB. This sample provided data for the item pools from which the TCAP norm-referenced tests are constructed.

To facilitate interpretation of Tables 1.1–1.5, we plot the statewide average score levels for tests in Figure 1. Score levels are plotted for each grade and test for the six successive cohorts of students. Inspecting Figure 1, we see a substantial amount of irregular variation in score levels from one cohort of students to another. The irregularity is most pronounced in the social studies test and least in mathematics and reading. The presence of this haphazard variation has important implications for value-added assessment; we discuss it further in section 2.1.

Despite the variability, there is with one exception an overall trend for score levels of the tests to increase since the introduction of the assessment in 1990. The exception is reading, which shows since 1990 almost flat levels in all grades. Compared to the 1988 National Median (50-th percentile), however, the state average levels in reading are not unfavorable. In only a few instances are they below the national median; elsewhere they are a few points above it.

In the other four subject-matter areas the generally upward trend in scores is more pronounced in grades 4 through 8, and is especially consistent in language and mathematics. These increasing score levels over the six-year period can be interpreted as an outcome of teachers improving the alignment of their instruction with the objectives measured by the tests, and of more effective focusing of student learning on the objectives. Teachers' efforts in this direction

Table 1.1
 TCAP Norm-referenced Means by Cohort and Grade-level Tests (with
 National Percentiles of the Student Score Distribution)

READING								
Cohort	Grade-level Tests							Form
	2	3	4	5	6	7	8	
84							761.69	A
85						750.98	762.78	B
86					740.70	746.96	764.97	C
87				723.52	740.50	750.97	762.21	D
88			703.62	720.34	738.78	753.64	760.42	E
89		683.40	703.98	728.44	740.12	753.58	760.13	F
90	652.38	676.87	701.09	726.95	741.77	749.00		
91	655.04	685.58	707.91	725.23	741.06			
92	651.42	681.49	705.19	720.92				
93	655.19	683.39	706.00					
94	649.77	684.86						
95	651.53							
Percentiles								
77	686	713	731	751	775	777	787	
50	650	680	701	722	740	749	759	
23	605	640	663	688	709	717	728	

Table 1.2
 TCAP Norm-referenced Means by Cohort and Grade-level Tests (with
 National Percentiles of the Student Score Distribution)

LANGUAGE								
Cohort	Grade-level Tests							Form
	2	3	4	5	6	7	8	
84							759.52	A
85						748.35	759.77	B
86					735.86	744.96	766.40	C
87				726.15	741.46	751.98	770.74	D
88			709.50	724.26	743.89	754.43	764.00	E
89		698.02	711.79	738.04	745.75	756.28	768.53	F
90	668.80	698.31	713.03	737.33	741.57	756.69		
91	671.40	697.22	715.60	733.33	745.47			
92	680.97	698.85	715.30	732.48				
93	678.79	697.23	718.86					
94	672.41	698.78						
95	676.03							
Percentiles								
77	698	725	737	754	769	779	787	
50	667	696	707	724	739	749	757	
23	630	661	672	688	705	714	723	

are more successful in those subject-matters that involve knowledge of content as well as skills. Reading is not such a subject: it is a pure skill applicable to any content. Without increasing the amount of student time devoted to practicing the skill, it is difficult to improve average reading performance in a given school grade over a period of years. In addition, reading is not taught as a specific skill in grades 5 through 8. A combination of these effects is plausible explanation for the absence of upward trend in reading scores.

A notable feature of Figure 1 is that language scores show a strong upward trend in all grades except 2 and 3. These tests heavily involve the mechanics of language—correct spelling, punctuation, grammar, usage, etc. Learning objectives are clear and easy to focus on in instruction. As a result, especially in the upper grades, students are averaging well above the national median.

The mathematic averages in Figure 1 are notable for the wide range they exhibit between grades 2 and 8. Although grade 2 scores seem low relative to the other tests, comparison with the national median in Table 1.3 shows they are well above the national median. At third grade, however, they drop back closer to the median and remain there through grade 8.

In science, the upward trend during the six years is most apparent in grades 6 through 8. It probably represents improved organization of the science curriculum and better teaching methods and materials at these grade levels. Interestingly, scores in science are generally above the national median and somewhat more so in lower grades than higher grades. The same is true of social studies, even though the great amount of variability makes some of the trends difficult to discern. Despite the variability, average scores tend to be five to eight points above the national median in most grades.

Finally, a clear trend revealed by Figure 1 is the tendency on the CTB scale for gains between grades to be greater at lower grade levels than at higher grades. This is an interesting parallel with the general trend in childrens physical growth from birth to twelve years of age: annual gains in height and weight decrease over these years prior to the adolescent growth spurt. Language and mathematics differ somewhat from the overall trend in that there is a generally larger increase between grades 7 and 8 than between grades 6 and 7. These trends in rate of growth and achievement have important implications, which we pursue in section 2.3, for the conduct of value-added assessment. A point that is immediately apparent, however, is that comparisons between schools or teachers in terms of the average gains of their students, must be made grade-for-grade, and subject-for-subject. This is already the practice in

TVAAS reporting.

Table 1.3
TCAP Norm-referenced Means by Cohort and Grade-level Tests (with
National Percentiles of the Student Score Distribution)

MATH								
Cohort	Grade-level Tests							Form
	2	3	4	5	6	7	8	
84							780.93	A
85						761.21	779.11	B
86					749.99	759.13	782.98	C
87				729.57	747.53	766.71	782.49	D
88			706.33	731.85	753.12	767.28	784.16	E
89		680.66	708.50	736.35	755.13	767.14	783.28	F
90	631.45	676.54	712.88	735.80	753.96	768.48		
91	635.06	684.52	713.72	734.74	757.79			
92	636.52	687.63	711.39	736.27				
93	632.66	687.64	711.83					
94	629.02	684.09						
95	633.75							
Percentiles								
77	655	707	728	754	773	791	809	
50	615	675	701	726	745	760	778	
23	571	640	669	695	712	727	743	

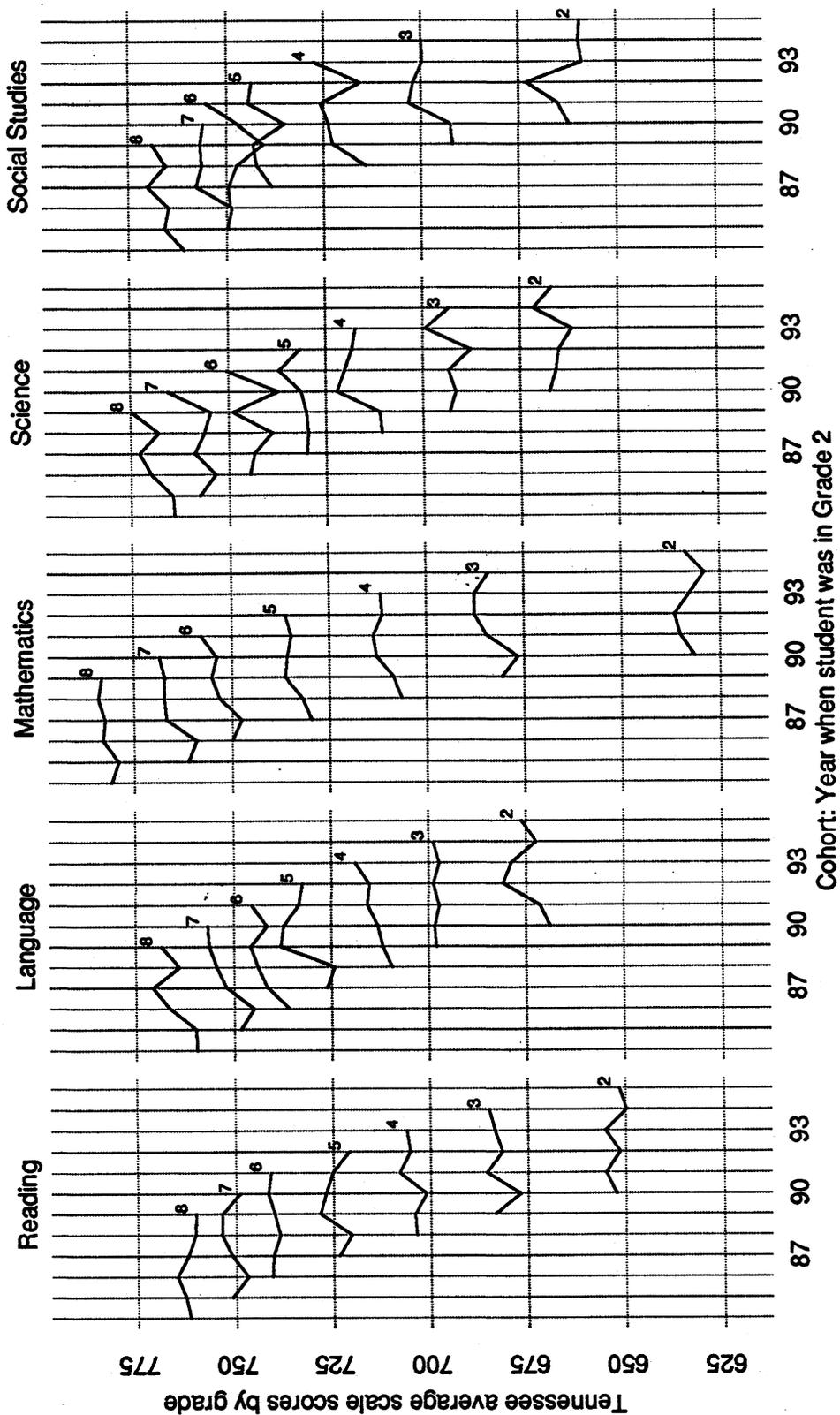
Table 1.4
 TCAP Norm-referenced Means by Cohort and Grade-level Tests (with
 National Percentiles of the Student Score Distribution)

SCIENCE								
Cohort	Grade-level Tests							Form
	2	3	4	5	6	7	8	
84							764.01	A
85						757.46	764.37	B
86					744.62	753.51	769.87	C
87				729.91	743.44	758.68	772.89	D
88			710.61	729.55	738.94	756.21	767.91	E
89		693.12	711.14	729.98	748.90	754.55	774.63	F
90	667.50	691.51	722.16	731.29	737.34	765.67		
91	665.87	693.32	720.11	737.06	750.49			
92	665.16	687.60	718.26	731.65				
93	661.72	699.42	717.35					
94	671.60	693.30						
95	667.06							
Percentiles								
77	690	724	740	762	776	787	795	
50	655	690	709	732	745	756	765	
23	621	654	674	694	709	731	731	

Table 1.5
TCAP Norm-referenced Means by Cohort and Grade-level Tests (with
National Percentiles of the Student Score Distribution)

SOCIAL STUDIES								
Cohort	Grade-level Tests							Form
	2	3	4	5	6	7	8	
84							760.95	A
85						749.73	765.80	B
86					749.06	748.65	764.77	C
87				738.44	749.50	757.77	770.12	D
88			714.17	742.20	747.33	756.30	765.53	E
89		691.73	722.67	743.12	740.56	756.66	768.93	F
90	662.18	692.45	723.78	735.27	747.11	755.98		
91	664.95	703.16	725.91	744.46	755.25			
92	673.03	702.03	715.63	743.66				
93	658.83	699.65	727.53					
94	659.63	699.76						
95	659.38							
Percentiles								
77	693	725	747	765	774	779	790	
50	652	691	713	735	745	749	761	
23	611	652	671	700	710	718	728	

Figure 1. TCAP Averages by Cohort, Grade, and Subject Matter



1.1 Comparison with results from the National Assessment Program (NAEP)

It would be helpful to have some confirmation of the above trends with tests based on other item content than that of CTB. To some extent this information exists in the Tennessee results from the NAEP state assessment program. Thus far, however, state NAEP results are limited to fourth grade reading in 1992 and 1994, fourth-grade math in 1992, and eighth-grade math in 1990 and 1994; however, Tennessee did not participate, in the 1990 state NAEP.

For what they are worth, the average proficiency scores in reading for Tennessee students on the NAEP scale, which ranges from 0 to 500, were 213 and 214 in 1992 and 1994, respectively. (A footnote of the NAEP report indicates that Tennessee did not satisfy one of the specified guidelines for school sample participation rates.) Since the reading mean for the nation was 216 in 1992 and 213 in 1994, Tennessee was at about the national level in reading in these years, a result roughly consistent with the comparison of TCAP reading scores with the CTB 1988 norms.

The NAEP math results are less in agreement, however: the Tennessee mean for 1992 fourth-grade math was 209, compared to 217 for the nation, and the eighth-grade mean in that year was 258, compared to 266 for the nation. Perhaps the most plausible explanation for the discrepancy between reading and math is limited alignment of the NAEP 1992 item content in mathematics with that of the TCAP tests or with the state curricular guidelines. If so, Tennessee students would have had insufficient opportunities to learn these areas of mathematics. The same would not be true of reading, a subject that is highly general in nature, and readily learned in many situations not closely tied to test content specifications or curricular guidelines.

1.2 The TVAAS concept

TVAAS is an effort under the leadership of Professor William L. Sanders, University of Tennessee, to apply contemporary data-analysis methodology and computer technology to the problem of evaluating the performance of school systems, schools, and individual teachers. It has the potential to serve the purposes of accountability of school officials for learning outcomes in their

schools and evaluation of the pedagogical effectiveness of teachers nurturing those outcomes.

The central concept of Professor Sander's system is that, in respect to learning outcomes, a teacher, school or school system should be held accountable only for the amount of gain in achievement their students accomplish as a result of each year of their schooling. This means that, if students enter a given school or classroom at any level of achievement, high or low, the school or teacher is responsible only for raising their level, on average, by an amount set in some standard by the State Department of Education. This principle is advanced as the only fair and objective way that teaching effectiveness can be assessed in the presence of the wide variation in average levels of students' educational attainment typical of communities and school catchment areas in Tennessee, or indeed in any state.

Not having the data management systems that TVAAS has implemented, which is capable of following students through their school years, or from school to school within systems, other state accountability assessments have relied on statistical predictions of achievement levels expected in schools drawing students from differing home and community backgrounds. But reliable background information on students is difficult to obtain, and the relationships among the variables are never strong enough to predict accurately the achievement levels of individual students prior to instruction. These indirect procedures cannot provide a secure basis for measurement of gain or evaluation of teacher and school performance. Professor Sander's approach overcomes these difficulties by directly measuring gains in individual student's test performance from year-to-year.

Analysis of students' gains at the school and teacher level by TVAAS methods is based on a number of major assumptions, however, each of which needs verification. It also requires that certain quality standards be met by the tests employed, by the conduct of the testing, and by the collection and analysis of the data. The objectives of this audit and review are to examine these assumptions and to make necessary checks on data quality. We have had the complete cooperation of the University of Tennessee Value-Added Research and Assessment Center in making data available for audit and review.

Chapter 2

Measurement issues

Traditionally, the score on a multiple item test is the number or percent of items answered correctly. This type of score is quite satisfactory for placing students in order of merit for advancement or certification, but it is not suitable for measuring gains in achievement. To measure gain as a difference in test score before and after a course of instruction, we must have scores defined on a scale with uniform units throughout the range. Such scores are said to be on a *metric* scale. A further desirable property for purposes of statistical analysis and interpretation is that the distribution of scores on the metric scale among examinees should have the same shape and spread whatever the location of its average value on the scale. Number-correct or percent-correct scores do not have this property: their standard deviations, measuring spread, are greatest towards the center of the distribution and decrease markedly toward the extremes of none or all items correct. A still more stringent requirement of the value-added assessment system is that the scores of the full series of on-grade tests are represented on a metric scale encompassing achievement levels from grade 2 to grade 8. Ideally, the standard deviations of the score distributions on this scale should be constant for all grades or at least related to grade level in a simple way.

In the educational measurement field, there are two widely used methods for representing achievement test results on a metric scale over a number of school grades. The first transforms the test number-correct scores into so-called “grade equivalents”; the second, based on concepts from item response theory (IRT) estimates so-called “IRT scale scores” directly from the pattern of correct or incorrect responses to the test items. Grade equivalents transform

the test scores so that the mean grade-equivalent at each grade is equal to the nominal value of the school grade, e.g., 1, 2, 3, . . . , 8. The most general IRT method produces a scale on which interactions between items and the responses of examinees at the same level of achievement are, insofar as possible, distributed normally, simultaneously for all items (after allowing for chance successes if the items are multiple-choice). For a discussion of grade-equivalent scaling, see Peterson, Kolen, & Hoover, (1989); for IRT scaling, see Lord (1980) and Lord & Wingersky, (1984).

Both types of scales suppress the curvilinear relationship between mean scores and standard deviations that occurs with number-correct scores, but the grade equivalent scale introduces its own pattern of relationship—namely, increasing variation with increasing grade, a phenomenon sometimes referred to as the “fan-spread” effect. Taken at face value, this effect implies that individual differences in achievement increase throughout schooling. It also implies a positive correlation between achievement level and gain: students who are initially at a higher level of achievement gain more rapidly than those who are initially at low levels. Schultz and Nicewander (1995) have recently demonstrated, however, that the fan-spread effect is an artifact of grade-equivalent scaling. By simulating test scores from ability distributions that increase in mean score level with grade but have *constant* standard deviation, they found that the transformation of the number-right scores to grade equivalents produces the familiar pattern of *increasing* standard deviation. At the same time, a positive correlation between rate of gain and score level is introduced that was not present in the untransformed scores. Since there is no a priori reason that cognitive growth should have a straight line relationship with grade level, or even with age during childhood (physical growth does not), no psychological, educational, or other substantive interpretation should be placed on the increasing standard deviation or positive correlation between level and gain typically seen in grade-equivalent scores.

IRT scales for tests of achievement over the school years are better behaved in this respect. Although they can show increasing, decreasing or constant standard deviation with grade level, the extent of correlation of scale-score level with gain is not as great as seen in grade equivalents. Thus, the TCAP norm referenced tests, which have been IRT scaled by CTB from their 1988 national sample, are well-suited to an assessment system based on gains. Indeed, the TVAAS data base is the best resource presently available for investigating the properties of IRT metric scales for well constructed achievement tests

administered annually in a testing program that follows individual students through their school years. We have examined the scaling properties of the TCAP test in these data, and present the results below after discussing a prior question.

2.1 Are successive forms of the TCAP tests equally difficult?

The achievement gains for TVAAS are differences between scores on a form of the test administered in the current year minus the score on a different form of the corresponding test administered in the previous year; the overall difficulty of corresponding tests in the two forms must therefore be equal for the results to be meaningful. If a test form is more difficult than those that precede and follow, it will systematically make the first measure of gain too low and the second too high. Assuming that any such variation in test difficulty will be random among the grade-level tests from year to year, we would therefore expect the average gains for successive cohorts of students to vary in excess of variability arising from the sampling of students or teachers. We saw in Figure 1 that this was the case in the TCAP data.

CTB constructs successive forms of the TCAP test by procedures that attempt to minimize the differences in form difficulty. As described in Appendix 1, items for each form are selected by stratified random sampling from large item pools, and they are positioned in each form so that items of similar content and difficulty appear always at the same location; the latter is a safeguard against so-called *context* effects on item difficulty. If the composition of the item pools and the procedures for sampling items are held constant and the number of items on each test is large, the average difficulty of the forms is then expected to be consistent. Thus, in the TVAAS measured based on norm-referenced items in the TCAP forms, any excessive random variation of form difficulty would be more likely in the shorter tests—namely, Science and Social Studies, which contain only twenty items—than in Reading, Language, and Math, which each contain forty items. This effect is also apparent in Figure 1.

Of the seven TCAP forms thus far constructed, items for the first four were drawn from two randomly parallel item pools arising from the 1988 national

survey. These forms contain no duplicate items. The three subsequent forms (E, F, G) contain not more than twenty-five percent of items that had previously appeared in a TCAP form, and none that had appeared in the form just previous to the current form (see Appendix 1). Given the large number of items appearing on the tests, and the very low exposure of any particular item, there is virtually no possibility of students encountering items that had specifically been taught or discussed in class or elsewhere; thus, no variation in effective test difficulty could be expected from this source.

Nevertheless, the possibility of random variation due to item selection remains, especially in the twenty-item tests. A standard procedure for verifying the equivalence of test difficulty is the so-called method of *equivalent-groups equating*. In this procedure, the alternative test forms are assigned randomly to examinees drawn from the same population. Typically this is done by simply packaging the two forms together in rotation and distributing them to students in the classroom in any order. The test booklets must of course have the same external appearance and the same arrangement of content within forms, so that one set of test taking instructions applies to both. This arrangement of test administration guarantees that the scores from the two forms can be considered to arise from the same population of examinees. Thus, if the location and shape of the score distributions for each form are identical within sampling error, the forms can be considered equivalent and interchangeable for testing purposes. If they are not, conversion tables can be constructed to adjust the mean and standard deviation of the distribution of the new form with that of the old, or, if the sample size is large enough, to equate the percentiles of the new forms to the old. The former is referred to as *linear* equating and the latter as *equipercentile* equating. In large scale assessment studies, random equivalent groups equating can be carried routinely by randomly distributing a small percentage of the previous years test forms among those of the current year.

For the TCAP test, no such equating studies have been conducted, but we can still seek evidence of possible variation in forms difficulty by examining statewide average gains across the grade-level test results for each cohort represented through the 1990 through 1995 assessments.

For the science and math tests, which are, respectively more and less at risk of discrepancies and difficulty between test forms, we show the corresponding cohort by grade level gains in Tables 2.1 and 2.2. The enormous size of the statewide sample represented in the data justifies tabulating the gains to two

decimal places on the CTB scale. In addition, we plot the gains for all tests in Figure 2.

Considering that large changes in the composition of the statewide population of teachers and students, or even of teaching practices, are unlikely in a single grade for a single year in a particular subject-matter, one would expect fairly smooth trends in gain across cohorts. For the forty-item math test this appears to be the case, the main exception being the grade 6 to 7 gain on the 1986 cohort, which appears to be too low. Looking back at Figure 1 (see also Figure 1 of Appendix B), this appears to be the result of Test Form B for grade 7 being too difficult and Test Form A for grade 6 being somewhat too easy; that accounts for the low gain of 1986 cohort followed by an extra high gain by the 1987 cohort, which had the benefit of test form C being more in line with the difficulty level of the later forms for that grade.

Table 2.1
State-wide Gains by Cohort

SCIENCE						
Cohort	Grade					
	3	4	5	6	7	8
85						6.91
86					8.89	16.36
87				13.53	15.24	14.21
88			18.94	9.39	17.27	11.70
89		18.02	18.84	18.92	5.65	20.08
90	24.01	30.65	9.13	6.05	28.33	
91	27.45	26.79	16.95	13.43		
92	22.44	30.66	13.39			
93	37.70	17.93				
94	27.70					

Figure 2. TCAP Gains by Cohort, Grade, and Subject Matter

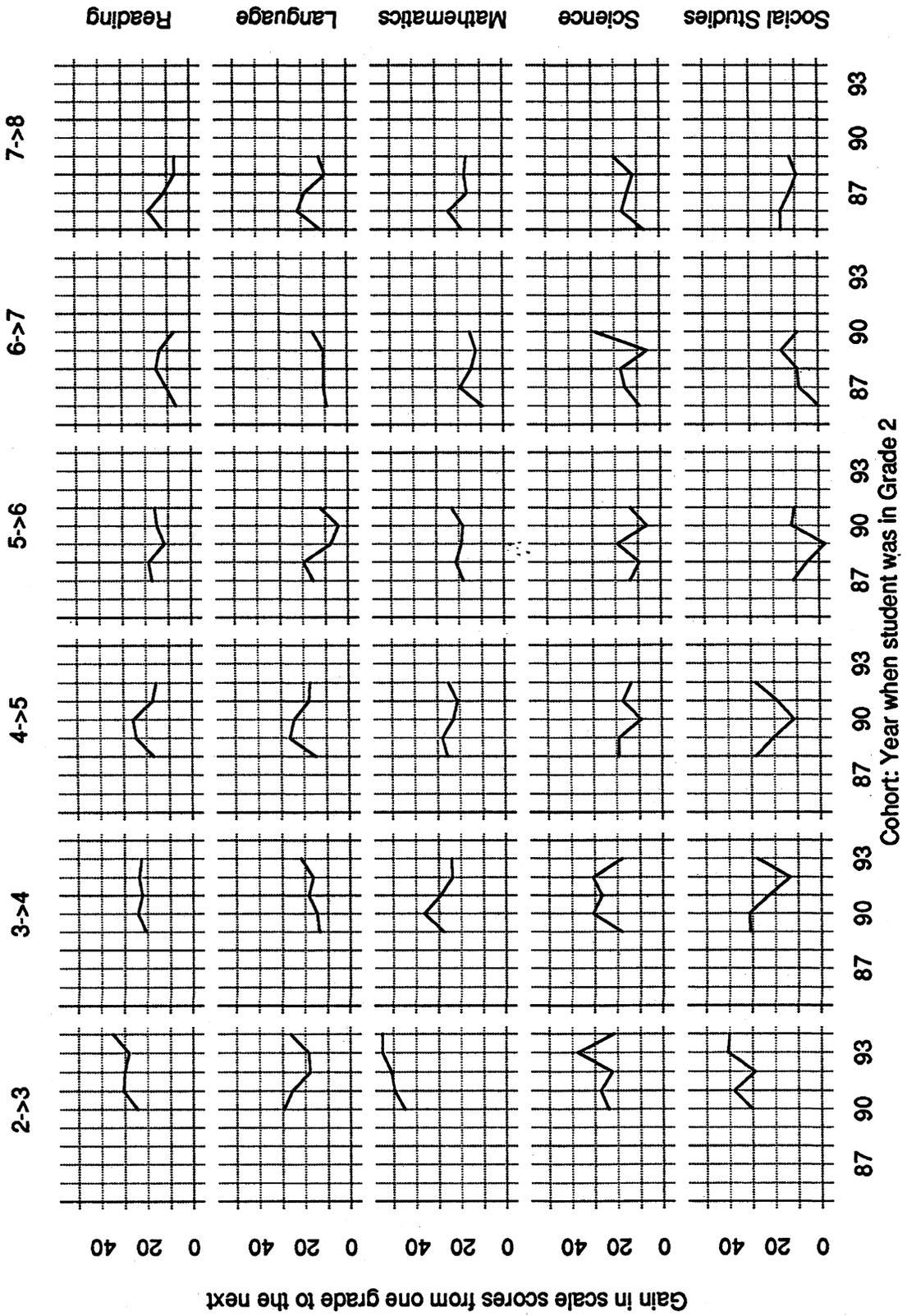


Table 2.2
State-wide Gains by Cohort

Cohort	MATH					
	Grade					
	3	4	5	6	7	8
85						17.90
86					9.14	23.85
87				20.96	19.18	16.27
88			25.52	21.27	14.16	16.88
89		27.84	27.85	18.78	12.01	16.14
90	45.09	36.34	22.92	18.16	14.52	
91	49.46	29.20	21.02	23.05		
92	51.11	23.76	24.88			
93	54.98	24.19				
94	55.07					

In fact there is a tendency for as similar effects in the grade 8, 6, and 3 gains indicating form B tests that are somewhat too difficult. This could indicate either some more general problem with Test Form B or some effect of conditions of test administration in these grades during the second year of the testing program. The latter explanation gains some plausibility, however, when one notices in Figure 1 that the grade 7, 1986 cohort scored below all other cohorts on all five tests in the year that Test Form B was administered (1991). Since all five tests for a given grade are contained in the same test booklet, there could be a problem with that specific booklet, or perhaps in the instruction given to the grade 7 teachers who administered the test that year. In any event, whatever produced this effect must be a condition existing in some degree in all schools, or perhaps in greater degree in a few large school systems, that is specific to that grade and year. An effect of the test form is perhaps most likely, but other explanations cannot be ruled out.

Another exceptionally large gain among the 40-item tests is that for language in the 1989 cohort between grades 4 and 5. Looking again at Figure 1, we see that this gain may be the effect of the grade 5 test in form C being too easy, thus producing the large gain during the third year of testing.

Table 2.3
 TVAAS 11 county longitudinal data
 Student S.D.'s and correlation of level and gain

READING								
Cohort	Grade							Correlation: Level vs. Gain
	2	3	4	5	6	7	8	
86					41.9	41.3	41.9	0.000
87				43.8	41.9	40.9	41.5	-0.081
88			40.2	42.3	36.8	39.9	39.2	-0.036
89		46.7	45.9	40.2	39.7	38.8		-0.232
90	50.2	48.9	42.9	43.0	41.7			-0.223
91	60.4	48.3	46.9	46.3				-0.293
92	57.7	43.6	48.0					-0.241

Table 2.4
 TVAAS 11 county longitudinal data
 Student S.D.'s and correlation of level and gain

LANGUAGE								
Cohort	Grade							Correlation: Level vs. Gain
	2	3	4	5	6	7	8	
86					38.9	46.8	46.3	0.259
87				42.2	41.1	50.6	43.6	0.138
88			39.9	42.6	41.5	48.2	45.1	0.218
89		40.3	41.8	39.8	40.5	99.1		0.222
90	41.8	46.8	44.5	40.2	41.5			-0.045
91	42.3	45.5	43.4	45.3				0.088
92	44.2	42.8	46.8					0.085

Table 2.5
 TVAAS 11 county longitudinal data
 Student S.D.'s and correlation of level and gain

MATHEMATICS								
Cohort	Grade							Correlation: Level vs. Gain
	2	3	4	5	6	7	8	
86					43.3	43.8	45.8	0.098
87				41.4	42.7	44.7	47.1	0.200
88			39.4	41.7	40.6	46.2	45.3	0.234
89		48.4	43.2	38.9	43.3	49.3		0.086
90	47.8	41.1	42.0	41.4	42.7			-0.016
91	50.7	44.2	42.6	40.9				-0.176
92	51.7	44.7	40.9					-0.253

Table 2.6
TVAAS 11 county longitudinal data
Student S.D.'s and correlation of level and gain

SCIENCE								
Cohort	Grade							Correlation: Level vs. Gain
	2	3	4	5	6	7	8	
86					46.0	39.8	36.0	-0.271
87				45.8	52.4	42.4	40.7	-0.185
88			45.9	41.8	47.9	41.7	38.6	-0.129
89		48.8	42.7	41.9	43.4	45.8		-0.013
90	53.9	45.3	48.4	43.5	47.6			-0.056
91	51.7	43.5	44.1	42.1				-0.121
92	52.3	50.3	45.7					-0.090

Table 2.7
TVAAS 11 county longitudinal data
Student S.D.'s and correlation of level and gain

SOCIAL STUDIES								
Cohort	Grade							Correlation: Level vs. Gain
	2	3	4	5	6	7	8	
86					41.2	40.8	39.6	-0.043
87				44.9	44.2	43.0	41.6	-0.074
88			45.4	39.4	47.7	43.2	39.2	-0.089
89		50.9	53.5	45.9	39.1	41.4		-0.286
90	60.8	55.5	49.4	47.2	46.7			-0.209
91	64.9	53.6	46.2	47.0				-0.256
92	54.8	54.8	52.9					-0.008

not for mathematics. The reading tests show relatively few problems: other than the too difficult grade 7, form B, test mentioned above, and a possibly too easy grade 5, form C, test, little else seems notable.

The observation that most implicates problems in forms equating with year-to-year variability is the many more problems seen in the twenty-item science and social studies tests. The smaller number of items in these tests does not allow random variation in the difficulties of the replaced items to average out in the test score. Although the quality of teaching of science and social studies may vary more than in other subjects, it is difficult to explain how the wide variations in score levels and gains seen in Figures 1 and 2 could occur in different cohorts and different grades systematically throughout the state without there being some artifact in the test forms themselves. It is similarly difficult to imagine that changes in curricular goals or instructional practices could have statewide effects for single years and single subjects in this way. Forms equating problems are the more plausible conclusion.

There are three ways in which this inference could be tested. If the original data files containing item response records from all tests have been retained, the scaling procedure could be re-calibrated by so-called "nonequivalent-groups equating" making use of the twenty-five percent of common items between successive forms (see Bock and Zimowski, 1996). If the test results based on this equating showed relatively smooth trends in score level and gain from cohort to cohort, the existence of equating problems in the present scores would be strongly confirmed. Alternatively, a similar test could be carried out by random equivalent-groups equating described above. This could be accomplished by inserting a certain proportion of test booklets from previous forms randomly in with those of the 1996 assessment. Because the current and previous forms would have then been administered to the same population, one would expect them to show the same state averages within the limits of sampling error. Differences in average score levels in excess of sampling error would indicate a forms equating problem, and the differences could be used to correct the state means from previous years. This should produce more regular trends in state average score levels and gains.

Finally, a weak test for the presence of form equating effects would be to smooth the trends in the state mean scores over cohorts and estimate form effects from the residuals. We have shown such calculations in Appendix B. If the application of corrections to the school gain estimates made the results more consistent and interpretable, this would be further evidence that form

equating discrepancies have occurred.

2.2 Do standard deviations of students' IRT scale-scores vary systematically with age during the school years? Are achievement levels and achievement gains correlated in these years?

To answer these questions, we needed longitudinal data for individual students. To obtain such data, we selected from an eleven-county sample supplied by Professor Sanders all students who participated in the 1990-1994 testing and had uninterrupted sequences of test scores in grades 2 through 8. Represented in this group are students from the 1986 through 1992 cohorts.

Tables 2.3–2.7 show the scale-score standard deviations for each cohort and grade level in the data. Because these figures measure deviation about the mean for each cohort and grade-level, they are unaffected by any variation in the difficulty of the tests. Across the rows of the tables, they are also unaffected by year-to-year changes in the composition of the student population: each row of the table contains results from precisely the same students measured at one year intervals. The sample sizes are all in excess of 3,600—large enough to justify reporting the standard deviations to one decimal place.

The right-hand margin of these tables shows the correlation between students average score levels and their average gain during the years they are in the sample. (Average gain is measured by the slope of the best-fitting straight line relationship between grade level and corresponding achievement score.) Notice that these correlations represent relationships at older ages in the early cohorts and younger ages in the later cohorts.

Inspecting the standard deviations within each cohort of the tables, we see systematic changes in age as represented by the grade levels. The pattern of variation tends to differ at earlier and later ages, and also differs among the subject matters evaluated by the tests, as follows:

Reading There is a clear trend for generally *decreasing* standard deviation with age; the decreases are greater at younger ages and all but disappear when the students reach grades seven and eight.

Language Trend in the standard deviations is more variable than in reading, but is generally *increasing*, especially in the earlier cohorts in grades 4–7; however, the 1990 cohort shows no consistent trend.

Math The standard deviations show a generally curvilinear relationship with age; they are consistently *decreasing* from grades 2–4 and *increasing* from grades 5–8.

Science Standard deviations are consistently *decreasing* in the 1986–88 cohorts, but are variable elsewhere except for a consistent decrease between grades 2 and 3.

Social Studies Standard deviations, consistently *decreasing*, are although less marked in the 1986–88 cohorts.

The correlations between achievement level and achievement gain also show consistent trends over the cohorts as follows:

Reading Correlation *zero* or *negative*, with the largest negative values among the younger students in cohorts 1989–92.

Language *Positive* among the older students in cohorts 1986–89; essentially *zero* among younger students.

Math *Positive* among the older students, except in the 1986 cohort. *Negative* for younger students in the 1991–92 cohorts; essentially *zero* in the middle cohorts, 1989–90.

Science *Negative* in all instances, but near *zero* in the middle cohorts.

Social Studies All correlations *negative*, but essentially *zero* in the older cohorts and in the 1992 cohort.

As expected, the correlations tend to be negative when standard deviations are decreasing with age and positive when they are increasing. It is impossible, however, to identify the source of these effects in the present data. They could arise either from artifacts of the IRT scaling procedure, or they could represent real differences in achievement levels reached by students at given ages, or both. Decreasing standard deviation and negative correlation would certainly result

if the IRT transformation stretched the scale too greatly toward the low end. Conversely, if some students were placed in more or less demanding programs of instruction based on their prior achievement levels, increasing variation might be expected during the middle-school years with resulting positive correlations. A combination of these effects could explain the observed trends if middle school tracking of students occurred in mathematics and English composition courses only. The increased individual differences in achievement in math and language that would result would then differentiate them from the reading, science, and social studies tests that show declining standard deviations and negative correlations.

The latter effects could also be real. Perhaps by the time children enter first grade, they already differ (according to their family, preschool, and kindergarten backgrounds) in their nascent reading skills and real-world knowledge represented in science and social studies. As they move through the school grades, they may become more similar in these skills and knowledge with the of equal opportunities to learn provided by the schools; the effect would be magnified if there is tracking of these subjects in lower grades but not in upper grades. The plausibility of these conjectures could be investigated by a technical examination of item characteristics of the tests and studies of student transcripts, but this activity is beyond the scope of the present review.

Although the magnitude of all of the correlations is less than .3, a good number of them are large enough to have implications for the comparison of gains between teachers whose students differ in average achievement level. To judge the effect of such correlation, we need the actual regression equation predicting average gain per year from average achievement level. For example, the correlation for reading in grades 3, 4, and 5 in the 1992 cohort was -0.24138 ; the corresponding mean achievement level over those years was 696.77—close to the third grade mean; the average rate of gain was 24.54 score points per year; the corresponding standard deviations were 45.342 and 19.95. Thus, the regression equation predicting gain from level is

$$\text{Gain/yr} = \frac{-0.2414 \times 19.950}{45.343}(\text{Level} - 696.77) + 24.54 .$$

Substituting in this equation we find that a third grade teacher whose students are 25 points above the mean (approximately one-half standard deviation) can expect, other things being equal, a gain of 21.89 points. A teacher

whose students achievement level is 25 points below the mean, could expect a gain of 27.20 points. The difference, 5.3 points, is an appreciable fraction of the average gain at this level. It is large enough to suggest that adjustments for expected gain as a function for student score level should be included when teachers are compared for subject matters and grade levels where the magnitude of the correlation exceeds, say 0.15.

With such adjustments, student gains would be essentially independent of their average test score levels. This means that demographic groups, systems, schools, and classrooms would have the same potential for gain regardless of their average levels of test performance.

Chapter 3

Data Quality Issues

3.1 Introduction

The quality of the TVAAS reports cannot exceed the quality of the data that were used in producing them. We need to be assured of the completeness and accuracy of the database. Did all schools participate? Did all students take the tests? Were their tests completed and scored? Was information about students properly collated across the 5-year testing period? Were the student records linked with the appropriate teachers and schools? These questions are especially important in the TVAAS, because the longitudinal series and the student-teacher-school linkages are essential to the analysis. Although the analysis model and calculations are capable of interpolating the results over “missing” data, systematic gaps are likely to cause systematic biases.

We have examined the quality of the TVAAS database from several perspectives, using both the all-Tennessee report and statistical files and the detailed 11-county database that Professor Sanders prepared for us.

3.2 Comparison of the School Enrollment and Tested Student Populations (All Tennessee Data)

There appears to be only minimal linkage between the education statistics database of the Department of Education and the testing databases collected

by the State Testing and Evaluation Center and processed and maintained and at the TVAAS center. In particular, there is no external, automated confirmation that the numbers of schools, teachers, and students in the TVAAS system correspond at any given stage of the testing, analysis, or reporting to the number officially in the educational system.

To investigate this, we obtained from the Department of Education a set of computer tapes with basic school-by-school education statistics for the years 1991–1994. The data from the 1994 tape was selected, extracted, and aligned with the TVAAS school-level data for the 1994 testing. The alignment was complicated because of many small differences in coding and identification. The TVAAS database included some special schools (e.g., alternative schools) that were not present on the education statistics data file. There were also several schools that evidently merged and others that were created between 1994 and 1995.

Eventually, we were able to make an almost complete match. Two schools in the education statistics database did not appear in the TVAAS file (of the total of 1551). Six schools appeared in the TVAAS file but not the education statistics files. Some further checking revealed that two additional schools appeared in the TVAAS testing records but not in the school reporting files. While these seem like major discrepancies, they probably correspond to known and special circumstances.

A more general question is the overall numerical match between the education statistics data and the testing data. Some results are presented in Table 3.1. There is approximately a 3 percent “shortage” in the testing data in each of the grades 3–8. This is not unexpected given the usual difficulties of student illnesses, absences, and transfers.

A further examination shows that in some locations or school types the completion of the testing program may not be as thorough. Ten percent of the counties had shortages of more than 5 percent. Ten percent of the schools had shortages of more than 10 percent. The schools with all-white enrollment had about 3 percent shortage, while those with more than 50 percent non-white students had shortages of about 5 percent through grade 6 and about 10 percent in grade 8.

These high figures may be due to epidemics, high net student mobility, or some kind of systematic exclusion from testing. Perhaps some schools or some kinds of schools experience greater amounts of student mobility and consequent within-year changes in enrollment. But without better linkage of

the education statistics and the testing program data, these questions are hard to monitor.

3.3 Summary of the Completeness of Student Testing Records (11-County Data)

The TVAAS has created and maintains what may be one of the largest databases in the world of linked, longitudinal school test-score information. It now comprises 1) a six-year time series of Grade 2–8 student scores on a battery of scaled, comparable tests and 2) linkage of the students' test information with identification of the teachers, schools, and school systems responsible each year for their education. Each year, the TVAAS obtains new test scores for all current students and must identify the students already in the database, update their records, and create new records for new students. This is accomplished through matching of name, birth date, and sex and without benefit of a constant identification code. (Codes are assigned internally in the TVAAS system, not in the schools or the data collection.)

TVAAS reports that more than 90 percent of the incoming records in a year are matched. While this is a worthy accomplishment, our concern is for the net accuracy of the database. If, for example, students were unmatched or mismatched 10 percent of the time, then over the five-year span that students are kept in the sample, we would expect about 40 percent of the students to have some matching error.

For our audit of the accuracy and completeness of the student testing records, we examined the testing records in the 11-county database. Some results based on all students who were tested in 1995 are given in Table 3.2. The "regular" pattern for a student tested in grade 5, for example, would be testing in grade 4 in 1994, grade 3 in 1993, and grade 2 in 1992. We find that 67.5 percent of the students have records with that pattern. Other patterns would include (a) no earlier testing, (b) testing in grade 3 but not grade 2; (c) testing in grade 2 but grade 3 missing, (d) repeated testing in grade 4 or 3 or 2, etc. Summary statistics are given for those kinds of patterns: we found for these students tested in grade 5 in 1995, 7 percent had gaps, e.g., missed the grade 3 testing, 3.4 percent showed repeating grades, and 38.6 percent were somehow incomplete. These figures overlap somewhat.

Table 3.1
Comparison of the School Enrollment and Tested Student Populations

Grade	2	3	4	5	6	7	8
Enrollment from Statistical Records	67260	66953	65614	65656	66715	67659	64282
Number of Students in the TVAAS file	65448	65061	63674	63823	64838	65004	62548
Tested Percentage, i.e., this percent of enrolled students were tested	97.3	97.2	97.0	97.2	97.2	96.1	97.3
10% of the counties had testing percentages less than this number	94.2	94.0	92.9	93.4	93.0	94.1	93.9
The lowest county testing percentage was this number	87.1	80.8	76.0	86.8	85.2	89.5	79.4
10% of the schools had testing percentages less than this number	88.9	87.5	88.2	88.0	88.7	87.1	88.4
The 276 schools with all-white enrollment had this percentage of testing	97.8	97.1	97.5	98.0	97.6	98.1	98.2
The 217 schools with less than 50% white enrollment had this percentage of testing	94.4	94.1	94.4	94.5	94.9	89.2	91.3

The finding is that from grade 4 on, there are at least 25 percent and up to 40 percent of the student records that fall outside of the regular pattern.

There are two conclusions from these results. First, any analysis procedure that needs to deal comprehensively with the student database will require good mechanisms for partially handling matched data and irregular longitudinal patterns. Second, because there is no linkage between the TVAAS database and data processing with a real student statistics and tracking system, one cannot tell if the gaps and inconsistencies correspond to the reality of student mobility, to failures in the matching system, or to systematic exclusions. Any of these conditions can potentially introduce biases into the reports.

3.4 Completeness of the Subtest Records (11-County Data)

A large part of the statistical and computational difficulty in the TVAAS statistical processing involves keeping track of and compensating for potentially incomplete testing records. Did many students not have scores on all five subtests in a given testing period? We see in Table 3.3 that the percentage is very small in all grades and in most grades has gotten over the years. Our conclusion is that missing data within student and year is of little importance and could perhaps allow some simplification in the analysis procedures.

3.5 Percentage of Student Test Records with Teacher Assignment Information (11-County Data)

The teacher-student assignment records have been collected for three years, 1993–1995. The data collection requires much detailed, repetitious work by the teachers. For example, they need to re-bubble their social security number for each batch of students and each subject area. We understand that the teachers do not like the task, and with no results released yet that use the information, there may also be some resentment. The data collection process could be considerably streamlined by re-designing the form and by developing good precoding systems.

Table 3.2
Summary of Completeness of Student Testing Records (11-County Data)

Grade of student when tested in 1995	Number of students	Percent of students with exactly that pattern	Percent with one or more gaps in their testing records	Percent with repeated grades	Percent with incomplete record
2	10701	98.3	0.00	1.7	0.00
3	10687	83.6	0.10	2.6	14.0
4	10793	75.1	3.20	2.9	22.2
5	10512	67.5	7.00	3.4	29.8
6	10643	61.4	11.0	2.7	38.6
7	10728	60.6	10.8	3.1	39.4
8	10385	61.5	10.8	2.4	38.5

Grades in which student should have been tested in previous four years.

	1994	1993	1992	1991
	-	-	-	-
2		-	-	-
3		2	-	-
4		3	2	-
5		4	3	2
6		5	4	3
7		6	5	4

Table 3.3
Completeness of the Subtest Records (11-County Data)

Grade	Year of testing				
2	1.99	1.96	1.74	2.42	1.22
3	2.58	2.39	2.43	2.11	1.48
4	3.00	2.57	2.25	1.97	1.69
5	2.40	1.85	1.95	1.99	1.82
6	4.34	3.44	2.61	2.54	2.49
7	4.15	3.20	2.64	2.64	2.95
8	3.52	2.74	2.49	2.51	3.23

Note: These are the proportions of students for whom some but not all of the five subtests are recorded.

Our concern for the analysis of the teacher data was how completely and accurately the assignment records have been collected and entered into the database. The question asked is what percentage of the testing records have a connection to teachers. We see in Table 3.4 that the answer is 70–86 percent, depending on the grade and subject area. The lower percentages occur in grades 7 and 8, especially in reading. This is understandable since there is no explicit teaching of reading at that level.

The general rate of about 85 percent completion may be due, especially at the level of grades 2–6, to student mobility, because teachers are to report assignment only when their attention to a student exceeds a 75 percent of days in the school year. However, when we made a detailed, school by school examination of the data, assignment data was sometimes missing for whole classes rather than for some students from each class. That this occurs is not unexpected because the data monitoring process assures processing of the forms as received, but no systematic verification of what forms should have been received.

Teacher assignments are, of course, quite different in the circumstances of an elementary school with whole classes taught by one teacher, in elementary schools with shared teacher responsibilities, or in intermediate schools with

subject-matter teachers. These difference should be perhaps mirrored in the data collection procedures, monitoring, and processing.

3.6 Discussion

Assuring accurate and complete data is an important part of gaining public confidence in the testing and reporting program. If there are missing schools, missing students, or missing tests, there may be the appearance or reality of manipulation and unfairness in the data and reports. The ability of the analytic model to “handle” missing data is not a great comfort. If information is missing or misallocated systematically, there will likely be distortion in the results.

In the tables discussed above, some problems in data quality are evident. Our overall impression is that, while the database is well constructed, it lacks independent and thorough checks against other statistical and school records. For example, there is no crosscheck between testing records and enrollment records, and no audit of teacher assignments against teacher rosters. The teacher assignment information should be considered to be of unknown quality, partly because a relatively high proportion of students are unassigned, especially in grades 7 and 8, and partly just because this is a novel and difficult data collection procedure that may take several applications and reporting cycles before it is stabilized.

In summary, we conclude that:

1. The match with educational statistics is good, although there is some suggestion of lower match in some areas, and lack of data makes matching rates difficult to monitor;
2. The student-record linkage seems to be satisfactory, although the proportion of broken records suggests some problems. It is clear that there will not be clean longitudinal records for many students and that the degree of irregularity varies with school and school type. Some biases may result from this.
3. The subtest records are nearly always complete, so that some simplifications could be made in the analysis.

4. The linkage from students to teachers is never higher than about 85 percent, and worse in grades 7-8, especially in reading. This is again concentrated in schools, school types, and teacher types, creating the potential for biases. The forms for collecting this information must be made much easier for the teachers to use. The data processing of the linkage data could be improved. Also, there needs to be some thought about what assignments of teachers to student instruction are meaningful in the upper grades.

Table 3.4
 Percentage of Student Test Records with Teacher Assignment Information
 (11-County Data)

	Grade						
	2	3	4	5	6	7	8
Reading							
1993	84	83	83	85	81	74	70
1994	85	85	84	86	81	74	69
1995	85	84	85	84	79	74	70
Language							
1993	84	83	83	85	81	81	79
1994	84	85	83	85	83	85	79
1995	85	84	85	85	82	80	80
Mathematics							
1993	84	83	83	85	82	80	81
1994	85	85	84	85	83	83	84
1995	85	84	85	85	81	80	80
Science							
1993	83	83	83	84	82	81	82
1994	84	85	84	85	82	84	81
1995	85	84	85	85	81	79	76
Social Studies							
1993	83	83	83	85	82	77	80
1994	84	85	84	86	84	82	82
1995	85	84	85	85	82	79	77

Chapter 4

Data Analysis Issues

The TVAAS method of estimating annual achievement gains is based on three separate quantitative models for system, school, and teacher effects, respectively. “System” refers here only to counties; the analysis estimates each county’s annual achievement gain in each of the five subject-matter areas. The analysis under the school model does the same for each school within each county; special and municipal districts do not figure in the analysis. In the system model, each individual student’s progress is followed even when he or she moves between schools within the county. The analysis under the school model estimates similar gains for each school within the county and follows each student who is in the school for more than a specified percentage of the school year. Finally, the analysis under the teacher model estimates achievement gains of students some percentage of whose instruction in each school subject is the responsibility of a given teacher. The average of these gains is referred to as the “teacher gain”. Initially, teacher gains are estimated as deviations about the system mean gain, but the system mean is added in to each score so that teacher gains can be compared between systems and over the state generally.

4.1 The TVAAS Models

The system and school models are typical two-stage hierarchical models for educational survey data. The annual achievement measurements within students comprise the first stage, and the students comprise the second stage. In

addition, the model contains the overall system mean or school mean, as the case may be. The teacher model is three-stage hierarchical with teachers as the third stage. Multilevel models in education are now well known from the work of Aitken & Longford (1986), Bryk & Raudenbush (1992), Goldstein, (1987), and others; see Bock (1989). They are now the accepted standard procedures for analyzing growth and other effects in social and educational survey data, including large-scale assessment. The development of the underlying statistical theory, however, has been in progress for more than twenty-five years. Prominent contributors are Henderson, (1984), Harville, (1977), McLean & Sanders, (1988), Dempster, Rubin, & Tsutakawa, (1981), Laird and Ware, (1982), and Longford (1986, 1994, 1996), among others. Professor Sanders' development of TVAAS is based largely on Henderson and Harville's work on the so-called "mixed-model analysis" for biometric data.

All of these formulations of multilevel analysis distinguish between *fixed* effects, such as a system mean, and *random effects* such as those associated here with teachers or students. Random effects are assumed to be representative of a population distribution of values or scores. The procedure for estimating the realized values of random effects for individual teachers within counties makes use of the information supplied both by the student's test scores and the knowledge that the teacher's score is drawn from the distribution of scores within the county. These types of estimates are referred to variously as "shrunk" in the biometrics literature, as "regressed" in the education literature, and as "empirical Bayes estimates" in the statistics literature. For the most part, the analysis of hierarchical models is based on the assumption that the random effects are normally distributed. The analytical procedure simultaneously estimates the mean and standard deviation of this distribution along with the estimation of realized values of the random effects. Standard errors for the estimated parameters and effects are also made available by most of these procedures.

In addition to these general attributes, the TVAAS models, especially the teacher model, have some additional features not commonly seen in hierarchical models:

1. These models may be described as multivariate repeated measures. They handle observations consisting of the student's response to the five tests in up to five consecutive annual testings; in other words, each test may be represented by as many as five repeated measurements from each student.

In the analysis, these potential twenty-five measurements are treated as a single multivariate observation without distinction between the tests and the repeated measures. This is different than treatment of multivariate data, (see Anderson, 1958; Roy, 1957; Pothoff and Roy, 1964), where the qualitatively distinct measures, such as the tests, are treated differently than the repeated measures in order to allow for the possibly different scales and units with which they are measured. Ignoring the distinction is allowable in the TVAAS application, however, because the TCAP norm referenced tests are all standardized in the same way. The TVAAS treatment has the advantage of handling instances when a student does not give usable answers to all five tests in each administration (although this rarely happens; see section 3), whereas the classical treatment requires complete data in this case. In addition, handling the two types of measures symmetrically facilitates the possible combination of estimated effects for the separate tests into a single measure for teachers in the lower grades who instruct in all subjects (see section 5).

2. Even though the potential for combining teacher effects across tests and providing a standard error for result is inherent in the teacher model, it is not followed through in estimating the assumed multivariate normal distribution of teacher gains for the five tests. Even at the lower grades where this is possible, the effects for different tests are assumed uncorrelated and only their standard deviations (actually variances) are estimated; the correlations (actually covariances) are arbitrarily assumed to be zero. The reason for this is that in the upper grades, where teachers typically instruct in only one subject, there is no basis for estimating any such correlations or covariances. However, there is no reason why grades 2 through 4 and grades 5 through 8 should not be treated differently in this respect.
3. The most unusual aspect of the TVAAS formulation is in the definition of the teacher gains: they do not represent just students' average gain during the year of the teacher's instruction, but extend beyond to following years when the students are taught by other teachers. They are coded in the model in a form described as "layered". In effect, the gain attributed to any given teacher can represent gain from the previous year to the average of the current year and up to three subsequent years. No clear

rationale for this convention is given in the description of the methodology. While it is true that the teachers effect on students, favorable or unfavorable, might influence their achievement gains in subsequent years and that the averaging effect would make the estimated teacher gain more stable from year-to-year, the sensitivity of the estimate as an indicator of a specific teacher's performance would be blunted.

The layering approach also seems inconsistent with the basic concept in TVAAS that students' achievement gains are not strongly affected by their level of achievement; thus, a teacher is not disadvantaged by possible poor performance of the preceding teacher, and therefore should not be advantaged or disadvantaged by the performance of a subsequent teacher or teachers. This aspect of the TVAAS model needs better justification if it is to be a permanent feature of teacher evaluation.

4.2 The Analysis

The data analysis procedures described by Dr. Sanders and his associates, including the numerical methods for fitting the several TVAAS models, are extremely general and capable of handling virtually any irregularity that might occur in the student's records or the assignment of students time to teachers. This has the advantage of minimizing any prescreening or editing of the data, but it also makes the computational algorithms extremely complex and places heavy demands on computer capacity and computation time. Not all of this generality is made use of in computing the gains and standard errors required in the annual reports. For example, the system and school models compute all correlations among the five tests and as many as five repeated measures for each test, even though the final estimates of annual system and school gains and their standard errors require only the correlations between adjacent years. Similarly, the computations include students who are in the system only for one year and contribute only one set of test scores, even though these scores should not contribute to the estimation of gains.

Another aspect of the estimation of system and school gains that is questionable is the ignoring of the clustering of students within classrooms when computing standard errors of the estimated gains. The models as described assume that all students are responding independently and are not subject to correlation within classrooms due to sharing teacher and classroom activities

in common. There is a considerable literature concerning education evaluation (see, for example, Aitken and Longford, 1986) that emphasizes the importance of accounting for clustering effects in judging the significance and errors of estimation. Both students and teachers should be considered units of sampling in the analysis of such data, and the error estimates should include components of variance from both sources. Because the clustering effect almost always introduces a correlation in the data, standard errors for systems and school will be larger when teacher and classroom effects are accounted for.

In the analysis of the teacher model, it is primarily the "layered", multi-year concept of teacher effect on gain that adds complexity to the data analysis. Absent the layering, a much simpler analysis of teacher gains could be carried out on a grade-by-grade basis. The model for the analysis would be much more compact, so much so that system, school, and teacher gains could be estimated by conventional procedures for hierarchical data. A review of the cost benefits of the present TVAAS estimation procedures versus smaller scale methods focused on the specific quantities reported in the assessment would be desirable.

A standard practice in the development of statistical estimation algorithms is the comparison of its results with an independent procedure using the same data. A test of this kind carried out by Professor Walter W. Stroup, University of Nebraska, using a SAS matrix implementation of the Henderson algorithm, agreed fully with the corresponding TVAAS analysis. We carried out similar tests by classical analysis of variance methods in the samples of TVAAS data from eleven Tennessee counties. In these large samples, the two methods of analysis should give approximately the same results, and in fact they did. We also compare results of the TVAAS analysis of a sample of data with SAS procedures and the multilevel computer programs of Bryk and Raudenbush and of Goldstein and found good agreement. We are satisfied that the TVAAS computer procedures are performing correctly.

Chapter 5

Reliability Issues

The Tennessee Value-added Assessment System has not previously had the benefit of a systematic analysis of the reliability and stability of teacher, school, and system measurements of gain. In this section, we report results of such analysis based on the data on the 11-county sample (see section 2.2). The analysis describes variation in gain expressed as deviations from the sample mean gain; it is therefore unaffected by any systematic year-to-year variation due to faulty test-forms equating (see section 2.1).

There is inevitably considerable uncontrollable variation in students' test performance and in the instructional process within classrooms and schools. The estimation of achievement gains, either at the teacher-classroom or school level, must therefore be robust enough to separate the effects of interest from the background of random variation. In the language of communication theory, the estimation procedure must be able to extract the signal from the noise.

There are four main sources of "noise" in the TVAAS data:

1. The sampling of items for the TCAP norm-referenced test forms from year to year.
2. Effects of different groups of students in teacher-classrooms from one cohort to another.
3. Effects of different teachers, within schools, teaching the same subject matter.
4. Changes in school administration, instructional programs, teaching staff, resources, and other school-level influences from year to year.

For an overview of the magnitude of these components and their effect on the accuracy of estimating gains, we made use of the classical analysis of variance method of estimating components of variance in hierarchical (nested) data structures (see Graybill, 1961; Bock, 1985 reprint; see also Wiley and Bock, 1967). For this purpose, we selected a special set of schools and teachers from the 11-county data: we chose schools that had at least two fourth-grade teachers who had taught two consecutive cohorts of students. We then performed the analysis on gains (fourth-grade test score minus third-grade score) of 2,645 students in the classrooms of 81 teachers in 29 schools. The results of the components of variance analysis for the resulting unequal- N , nested analysis are shown in Table 5.1 in the section labeled ANOVA. In this method of analysis the estimates are not constrained to be positive, and near-zero expected values can appear as inadmissible negative values. As this happened in a number of instances, we also performed the estimation by the "maximum-likelihood" method that constrains the estimates to be zero or positive and properly estimates all values. The results are shown in the section labeled ML.

Notable are the zero or very small components of *between-school* variation for the reading, language, and math tests. This does not mean that the average gains in the test scores do not vary between school, but only that variation can be attributed to differences among teacher-classrooms, students within classrooms, and variation from one cohort of students to another. This is not as true of the science and social studies tests, suggesting that there is more variation in the emphasis that various schools give these topics in their instructional programs. It is also notable that *between-student* variation is larger for science and social studies, but that can be explained, as in section 2, by the reduced reliabilities of these tests, which contain 20 rather than the 40 items of the other tests.

Having only eleven counties in our sample of data, we could not reliably estimate a between-system variance component. Given the very small between-school components, however, we would expect the between-system components also to be small. This should be checked in the statewide data and examined in relation to the system by cohort variation. We comment on the stability of gains estimates at the system level in section 6.

Table 5.1
 Components of Variance for Gains from Grade 3 to Grade 4: 2 cohorts, 29
 schools, 81 teachers, and 2645 students

	Reading	Language	Math	Science	Social Studies
ANOVA					
Between Schools	-0.78	4.73	-10.30	24.10	37.76
Between Teachers within schools	24.77	37.68	86.59	32.74	0.04
School by Cohort	20.53	13.84	55.42	36.94	-9.07
Teacher by Cohort	28.67	11.57	11.90	-17.2	74.07
Between Students within classrooms	976.54	852.10	882.29	1434.72	1414.00
ML					
Between Schools	0	2.90	0	26.66	31.83
Between Teachers	27.62	40.09	90.54	29.51	8.63
School by Cohort	13.33	12.37	51.76	26.54	0
Teacher by Cohort	28.99	8.36	7.68	0	58.99
Between Students	979.81	854.33	884.91	1429.00	1916.19

5.1 Teacher gain scores

On the assumption that the components of between school variation for gains in science and social studies will decline as teaching of these topics becomes more uniform among schools, we show in Table 5.2 a similar component of variance analysis, again by both the ANOVA and maximum-likelihood methods, for teachers ignoring their school and system identifications. In addition, we supplemented the analysis of fourth-grade gains, which includes only teachers who taught all five subject-matter areas, with the results shown in Table 5.3 for eighth-grade gains in mathematics, where in many schools the teacher instructs only in that subject. Apart from school effects, all the variance components are positive by both methods of estimation and are very similar; however, for purposes of interpretation we prefer those of the more widely used maximum-likelihood method. Because the eighty-one teachers drawn from the 11-counties are undoubtedly somewhat more heterogeneous than those in an actual school system, the size of the teacher components may be slightly larger than in the TVAAS results based on single counties.

Because TVAAS evaluates gains for individual teachers in terms of average gain over the three most recent years, both student variation and teacher by cohort variation contribute to the instability of the average. The components of variance estimates enable us to compute a reliability-like index, which we will refer to as a “stability coefficient” (SC), of the teacher gain measurement. Since the size of this estimate depends upon the number of students contributing to the data for each teacher, we will assume a class size of twenty-five. On that assumption, the expected variance of teacher gains is the sum of the teacher component plus the teacher by cohort component divided by 3, plus the student component divided by 3 times 25. Thus, from the ML values in Table 5.2, the calculation for Reading is

$$\begin{aligned} SC_R &= \frac{28.04}{28.04 + 41.31/3 + 479.82/3 \times 25} \\ &= 0.51 \end{aligned}$$

Of the terms in the denominator, the first is the reliable variation and the second and third are error variation; thus, the stability coefficient is the ratio of the teacher component to the total. Similarly, the corresponding expected standard error is the square root of the sum of the two error terms. Finally,

Table 5.2
Components of variance for gains from Grade 3 to Grade 4: Between
Teachers, Teacher by Cohorts, (81 teachers; 2645 students)

	Reading	Language	Math	Science	Social Studies
ANOVA					
Teachers	24.01	42.28	76.58	56.17	36.73
Teacher by Cohort	48.62	25.01	65.75	18.63	65.26
Students	976.54	852.10	882.29	1434.72	1914.07
ML					
Teachers	28.04	43.58	85.10	56.12	40.03
Teacher by Cohort	41.31	20.08	60.53	15.58	59.30
Students	979.82	854.29	884.61	1435.00	1916.15

and the expected standard deviation of teacher scores is the square root of the sum of the three terms. These quantities for the five tests are shown in the upper section of Table 5.4.

It could be argued, however, that teachers who instruct in all five subject-matters should be evaluated in relation to their average gains over five tests. If we estimate the components of variation for that measure, we obtain the results shown on the right in Table 5.4. The stability for the combined measure of teacher gain (0.79) is a reasonably high value, comparable to the reliability of student achievement tests with 30 to 40 items. As we discuss in section 6 in connection with reporting, this figure is high enough to permit the identification of teachers who are sufficiently high or low in their 3-year average gain scores to merit special attention.

Although teachers in grades 5 through 8 who instruct in only one subject-matter cannot benefit from averaging gains over the five tests, they have the advantage of measurement based on a larger number of students. Suppose a math teacher teaches four different classes of 25 students each year. Then the 3-year average gain score is based on 300 students, and the student variance component in the stability formula is divided by that number. The result, 0.81, calculated from the ML values in Table 5.3, is shown in the bottom section of

Table 5.3

Components of variance for **Math** gains from Grade 7 to Grade 8: Between Teachers, Teacher by Cohorts, Students within Cohorts and Teachers (43 Teachers, 5135 students)

	Teachers	Teacher by Cohort	Students
ANOVA	30.18	14.38	744.25
ML	28.22	13.30	744.99

Table 5.4. This value is also high enough to insure accurate identification of teachers with the more extreme 3-year average gains.

To illustrate how the estimated teacher gains might be used, suppose one wished to identify with 95 percent confidence those teachers that are below the 20-th percentile or above the 80-th percentile of the teacher gains distribution (see Figure 3). If the gains measure has stability coefficient 0.80, the lower and upper cut points would have to be set at the 8-th and 92-nd percentiles, respectively, in order to insure 90 percent confidence that a teacher was correctly classified in the lower and upper 20 percent of the teacher population. In other words, the system would single out eight percent of teachers as meritorious and eight percent of teachers as problematic with respect to their 3-year average gains.

This result, of course, describes only the average situation. The teacher-gains standard error actually depends on the number of students the teacher has taught. Teachers with larger numbers of students would be classified with greater confidence; those with very few students would rarely be confidently classified as extreme.

5.2 School gain scores

The foregoing stability analysis can also be applied to average gain for schools, in particular grades and subject-matter test scores. In this case, we assume the teachers are not identified and the variation attributable to them is amalgamated with that attributable to schools. Since the school-mean gain score is an

Table 5.4
 TEACHER 3-year Average Gains: Stability Coefficients (SC), Average Standard Errors (SE), and Teacher-Effect Standard Deviation (SD) (Based on ML estimated teacher and teacher by cohort variance components)

	Reading	Language	Math	Science	Social Studies	All Tests
Grade 4						
SC	.51	.70	.73	.70	.47	.79
SE	5.18	4.25	5.65	4.93	6.73	3.17
SD	9.24	7.41	10.82	8.97	9.24	6.93
Grade 8						
SC			.81			
SE			2.63			
SD			5.98			

average of the gains of teachers teaching the same subject-matter at the same grade-level, the teacher variance component in the stability formula for schools is divided by the number of such teachers. With teachers unidentified, the total variation in school-mean gains is then the sum of the between-school component, plus the between-teacher component divided by the number of teachers, plus the school-by-cohort component, plus the teacher-by-cohort component divided by the number of teachers, plus the student component divided by total number of students in the classrooms in all the teachers.

If the 3-year average school-mean gain is considered, the school-by-cohort, teacher-by-teacher cohort, and student component must be divided by 3, and the teacher and teacher-by-cohort components must be divided by the number of different teachers during that period. To illustrate the calculation, let us assume that there are two teachers per school, each with 25 students per class per year, and that the same teachers teach for all three years. Then for fourth grade language, for example, the stability coefficient for schools, as computed

from the component values in Table 5.1, is

$$SC_L = \frac{2.90 + 40.09/2}{2.90 + 40.09/2 + 12.37/3 + 8.36/2 \times 3 + 854.33/2 \times 3 \times 225}$$

$$= 0.67$$

Because the between-school component is small and there is some school-by-cohort variation, this value is less than the corresponding stability coefficient for teacher gains (see Table 5.4). If there were three teachers per school, it would be even smaller. Absent appreciable between-school variation not attributable to teachers, the effect of averaging over teachers in the school mean score reduces the proportion of reliable variation when school-by-cohort variation is present: for this reason, it can happen that teacher-gains can be measured with acceptable reliability while school gains cannot.

As is apparent in Table 5.5, the measurement of school-gains is appreciably less stable than teacher-gains in reading, language, and math. Science and social studies are exceptions, however, because their between-school variation is greater. We have suggested above, however, that this variation may be due to differing school programs in science and social studies in fourth grade. It may disappear as consensus grows in the state on how these subjects should be taught in primary school.

Table 5.5
SCHOOL Grade 4, 3-year Average gains: Stability Coefficients (SC),
Standard Errors (SE), and Standard Deviations (SD)

	Reading	Language	Math	Science	Social Studies
SC	.47	.67	.65	.69	.62
SE	3.97	3.55	4.94	4.29	4.75
SD	5.44	5.84	8.35	7.72	7.66

The foregoing results raise a number of questions about the public reporting of school-level gains:

First, for those schools in which only one teacher is teaching the same subject-matter in the same grade for three consecutive years, the school- and

teacher-level results are identical; moreover, they can easily be identified with the teacher from the school report. This would violate the confidentiality of teacher evaluation.

Second, for larger schools with more than one teacher per subject per grade, the lower stability of these school-gain scores will require somewhat more extreme percentile cut points to identify high and low schools than those required to identify high and low teachers. Unless the system has many schools, the number of them so identified may be very small, or even zero.

Third, because the range of school-gain scores is small relative to their measurement error, they are likely to be overinterrupted, or misinterpreted by the public. The average test score levels for school that are more commonly reported for schools by school systems are much more stable relative to their measurement error and typically place the schools in approximately the same rank order from one year to another. The school-level gain scores will not enjoy that consistency (see section 6). Although our data sample was not large enough to carryout a similar analysis at the school system level, we believe that the same considerations would apply there (see section 6 for a discussion of problems associated with reporting average gains for systems).

5.3 Regressed estimates of teacher effects

The preceding analysis, based as it is on classical methods, makes use of the average of teacher-classroom mean gains as the measure of teacher performance. In practice, however, these unconditioned differences of teacher-classroom mean scores encounter numerical instabilities when class sizes are small, which is not uncommon in the Tennessee data. For this reason, TVAAS uses a so-called “shrunk” mean that takes into account the extent of variation that is likely, given the empirically observed standard deviation of the teacher-gains distribution. This type of estimate, which was first introduced by Truman Kelly (1947), and is called the “regressed” estimate in the educational measurement literature; in the recent statistical literature, it is referred to as the “empirical Bayes” estimate or, sometimes, the “mean of the predictive distribution”. We will refer to it here as the regressed estimate. Kelly expressed its formula in terms of the reliability coefficient and standardized variables, but the more general form in the present application is as follows,

$$d_i^* = \frac{1}{\left(\frac{N_i}{\sigma_e^2} + \frac{1}{\sigma_t^2}\right)} \cdot \frac{N_i}{\sigma_e^2} (\bar{d}_i - \mu),$$

where \bar{d}_i is the observed mean gain score for the teacher, μ is the mean gain in the population of teachers, N_i is the number of students for teacher i , σ_e^2 is the error variance component, σ_t^2 is the teacher variance component, and d_i^* is the regressed estimated gain. d_i^* is expressed here as a deviation from the population mean. If the school system average gain is substituted for that mean, the regressed estimates will be expressed in the form of deviations about that mean, and will tend to average to zero over all teachers in the system.

To give some impression of the difference between the observed class mean and the corresponding shrunken mean, we have shown these values for ten teachers randomly drawn from our fourth grade sample (Table 5.6). In each case the class means for the two cohorts for each teacher are shown in their raw form, and the mean for the first cohort is shown as deviations (differences) from the grand mean. The regressed mean for the first cohort is also shown deviated about the mean. These means are for the math test only.

The deviated regressed means have been supplied by Professor Sanders who applied the TVAAS procedure to our sample. It is based on his estimates of the variance components, which are obtained by a slightly different estimator than the ANOVA and ML estimators above, but are consistent with our results. The standard errors of the regressed estimates are also those of Professor Sanders.

The manner in which the regressed estimates shrink the observed deviated means back to their population mean when the number of observations is small is clearly seen in Table 5.6. This protects the estimation procedure from the wide variations one may encounter in the raw means, (for example, for Teacher 22 and Teacher 78), when the class size is small. This protection is essential to any type of teacher-performance measure based on gains in students' test scores.

The TVAAS model for schools (see section 4) does not include regressed estimation of school-level gains on grounds that the number of students involved makes their use unnecessary; this is certainly true for large schools. However, in small schools, where small numbers of students per class make regressed estimates of teacher effects advisable, there will likely be only one teacher per subject-matter per grade, and so the same reasoning should apply

Table 5.6
 Observed class means for **Math** gains compared to TVAAS shrunken estimates of teacher effects: Ten randomly selected **4-th Grade** Teachers from the set of 81 teachers in the 11-county sample

Teacher Number	N	Class Mean	Deviated Mean ¹	Deviated Regressed Mean	Model Standard Error
12	16	29.63	3.65	1.39	3.74
	21	35.90			
49	9	16.44	-9.54	-2.70	4.08
	20	16.85			
22	5	1.40	-24.58	-3.13	5.43
	7	49.43			
28	19	12.42	-13.56	-3.31	3.86
	15	27.40			
69	21	20.52	-5.46	-4.22	3.70
	17	14.65			
40	21	37.67	11.69	6.70	3.57
	21	50.43			
56	15	36.60	10.62	3.16	3.90
	18	31.00			
5	20	48.80	22.82	21.66	3.74
	17	57.41			
31	17	40.47	14.49	6.25	3.63
	23	18.61			
78	8	-16.25	-42.33	-10.92	4.72
	11	9.09			
¹ Grand Mean of 11-county sample		25.98			

to the estimation of school-level gains. This approach would also have the advantage of expressing the school gains as deviations about the system gain, which would then be unaffected by any problems in the equating of test forms. In addition, an analysis by a three-stage model that assumes random effects for students, teachers, and schools would have a more accurate standard error for the school-level effects than an analysis considers the sampling of students the only source of error variation.

5.4 Empirical evaluation of the stability of school and teacher gains estimated by TVAAS

To evaluate empirically the performance of the TVAAS model as implemented and applied to the actual data of the assessment, we have analyzed year-to-year variability of the reported school gain scores for 1993, 1994, and 1995, and the corresponding teacher regressed scores, supplied by Professor Sanders for the complete 11-county sample. The results are shown in Table 5.7 and 5.8, respectively, for grades 4, 6, and 8. Table 5.7 contains the average reported standard errors of *school* mean gain scores for three consecutive years at the three grade levels, along with empirical standard errors computed by subtracting the mean for the reported scores from each of the three years, summing the squares over all years and schools for each variable and each grade, dividing by the degrees of freedom (two times the number of schools), and taking the square root. In order to eliminate possible effects on form equating discrepancies, we expressed the school gain scores supplied by TVAAS as deviations from the state means for the respective years. The standard errors are therefore those that would be expected if the forms were perfectly equated. In addition, we computed the standard deviation of the reported school means as shown.

As we explained in section 3, the standard errors computed under the school model can be expected to be underestimates because they neglect the intraclass correlation in scores of students within the same teacher-classrooms; they also excluded larger-scale variation between successive annual cohorts of students. The empirical standard errors reflect both of these sources and are consistently larger than the model values. Indeed, they are an appreciable

Table 5.7

Statewide Mean Standard Errors for **School Gains** based on the TVAAS model compared with Empirical Standard Errors computed from Estimates for Three consecutive years per school (with standard deviations of 3-year school averages)

	Reading	Language	Math	Science	Social Studies
Model					
Grade 4	2.98	2.75	3.04	3.58	3.99
Grade 6	2.45	2.34	2.40	3.08	3.13
Grade 8	2.01	2.30	2.30	2.59	2.57
Empirical					
Grade 4	5.20	5.16	6.05	6.60	6.54
Grade 6	3.82	3.89	5.00	5.88	5.58
Grade 8	3.07	3.72	4.47	4.10	4.08
Standard Deviations					
Grade 4	7.08	8.37	9.87	9.06	9.17
Grade 6	5.30	5.98	8.48	7.23	6.96
Grade 8	3.36	4.83	6.48	5.00	4.27

fraction of the standard deviations between schools in Tennessee. The results raise the question whether it is advisable to publicly report these scores at their present levels of accuracy. A more prudent course would be to examine the distribution of school gains for the state as a whole and look for additional evidence that the schools with extremely high or low gains are memorable in other ways that would explain their positions in the distribution. If so, these schools could be identified publicly in terms of their percentiles without publicizing gain scores for all schools.

Table 5.8 contains a similar analysis of standard errors of regressed estimates of *teacher* gain scores for 1993, 1994, and 1995, along with the standard deviations for the 11-county data. Because the standard errors depend strongly on class size, which may not be normally distributed, we report here medians rather than means of the standard errors. The regressed scores are already protected from outliers, however, and their standard deviations as estimated in the usual way. Here, the model standard errors are much closer to their empirical values, for they correctly reflect the sampling of students and exclude only the between-cohort variation, which is suppressed somewhat by the averaging over the three years. The empirical standard errors are generally larger, but not greatly so. Although in Table 5.8 we show the standard deviation of the teachers regressed scores, it is important to understand that, when class sizes are small, the regression effect decreases considerably the population standard deviation. It is also the case, that the “layered” model by which the regressed estimates were calculated (see section 4) reduces variation between teachers. For these reasons, the standard deviations in this table are smaller than those for the observed gains, shown in Table 5.4, which are estimated as the sum of variance components in students’ gain scores.

Table 5.8
 Comparison of Median Standard Errors for **Teacher** effects based on the TVAAS model with empirical standards computed from three consecutive years of gains estimates per teacher (with standard deviations of 3-year effect averages)

	Reading	Language	Math	Science	Social Studies
Model					
Grade 4	2.32	2.29	2.48	2.22	2.31
Grade 6	1.64	1.60	2.15	1.53	1.43
Grade 8	1.57	1.75	1.94	1.47	1.37
Empirical					
Grade 4	2.02	2.57	3.79	2.37	2.02
Grade 6	1.30	2.56	3.59	2.14	1.54
Grade 8	1.18	2.80	4.31	1.69	1.71
Standard Deviations					
Grade 4	2.27	3.15	4.17	2.18	2.04
Grade 6	1.54	3.26	5.13	2.11	2.02
Grade 8	1.73	4.00	6.26	2.27	1.95

Chapter 6

Standards and Reporting Issues

An essential part of accountability assessment, whether at the system, school, or teacher level, is an objective and fair criterion for TVAAS identifying instances of superior, acceptable, or unsatisfactory performance. As the system is presently operating, mean gain scores for grades 3 through 8 are assigned to one of the following four categories: 1) equal to above the national norm gain, 2) below the national norm by one standard error or less, 3) below the norm by more than one but no more than two standard errors, and 4) below the norm by more than two standard errors.

There are a number of problems with this approach. The first arises from the national norm gain obtained by subtracting corresponding pairs of years in the 50-th percentile level for grades 2 through 8 from 1988 National Survey conducted by CTB-McGraw-Hill (see Tables 1.1–1.5). Problems are evident when we examine the national 50-th percentile in relation to the state mean scores for Tennessee. For reasons that can only be conjectured, students in Tennessee average substantially above the national median in language, math, and science in grade 2, but much less so in grades 3 and later. As a result, there is a persistent tendency for third grade gains in Tennessee systems and schools to be classified undeservedly in category 4. Similarly, the national reading gains between grades 6 and 7 and grades 7 and 8 are 9 and 10 points, but the corresponding figures for the 1988 and 1989 cohorts in Tennessee are 15 and 7 and 14 and 6. These discrepancies between the Tennessee average gains and the national gains, which may simply reflect different programs in reading instructions in these years, will be reflected in persistent category 1 classifications for grade 7 and category 4 classifications in grade 8, with possibly no

real justification in terms of school performance. Added to these difficulties is the fact that the national norms are now eight years old and perhaps losing their relevance as standards of comparison.

6.1 Teachers and schools

A further problem with the CTB 1988 national percentiles is that they describe student-level variation. This variation is much wider than at higher levels of aggregation such as classroom and school. Only the 50-th percentile of these different levels agree. This is not adequate for use with teacher and school gains, where knowledge of more extreme percentiles is necessary (see section 5). While the 1988 national 50-th percentiles are a useful benchmark for descriptive comparisons with score-levels and average gains in Tennessee, they are not suitable as criteria for decisions that might trigger actions, favorable or unfavorable to teachers, schools or systems.

The only workable solution to the problem of standards for gain is to create state norms for teachers and schools. This approach has the advantage that, in addition to fiftieth percentile gains, one can also compute percentiles at higher and lower levels so that teacher and school results can be reported more informatively than at present. Standards set in terms of these percentiles would not then show the above anomalies at the state level and would be easier to interpret. They could also be updated more frequently than is possible for national norms. The CTB national average gains would still be available as a point of comparison, but they would not be used directly to classify schools or systems.

A further problem with the present method of setting standards is the use of gain zones defined in terms of the standard errors of reported scores. This practice has the very undesirable effect of having systems or schools with very large numbers of students, and thus, very small standard errors, being organized to category 4 when their deviation from the norm is too small to have any practical importance. The solution to this problem is well-known and widely applied in the testing field. (It is used, for example, in the TCAP's student reports for both the norm-referenced and criterion-referenced sections of the test.) Test results are shown graphically in the form of scores, and confidence intervals around the scores, plotted with respect to the corresponding percentiles. Standards are represented by high, middle, and low percentile

intervals defined by state educational authorities. A result is then judged by the interval in which the score falls, and whether the confidence interval about the score does or does not cross a boundary between the intervals. We illustrate a hypothetical teacher report in Figure 3. The teacher is in the high gain interval for Language with better than 90 percent confidence.

6.2 School systems

We would raise the question of whether it is wise to emphasize gains in school system reports to the public. Probably it is safe to say that all other state assessment and achievement testing programs that report results for districts and schools do so in terms of *average score levels*. Although it is certainly true that this type of reporting encourages the media to rank schools and invite invidious comparisons, it is information that community groups and parents want to know. If gains only were reported, the media probably would rank them also, but they would bear little if any relationship to levels of performance indicated by average test scores. From an egalitarian point of view, this might be a desirable attribute of the reporting system, but it may not be sufficient justification for withholding the score-level information.

TVASS reports for school systems presently include score levels as well as gain scores. To illustrate the different information provided by these two types of we show, in Figures 4 and 5, results for two contrasting systems selected from the TVAAS data . System A is large and mostly urban and System B is smaller and mostly rural. Results are presented for Grades 2, 4, 6 and 8 and for Reading, Mathematics and Science. The results in these figures are simple test score averages and average differences; they are essentially the same as the system score-levels and gain scores that would be estimated by TVAAS.

The comparison of average test results in Figure 4 reveals large differences between System A and System B in both test score level and in trend. At the lower grades, the scores in System A are substantially higher in all three subject areas. That gap closes in the higher grades. In most cases, the score levels for a grade in System A increase across the years (cohorts). Except for Grade 8 Science, the trend in System B is that of constant or decreasing level over the years. The net result is that the gaps between System A and System B are getting larger.

In this three-faceted set of data, we need to look not only at differences

across subject areas and across grades, but also across years, which corresponds to cohorts of students moving through the school systems. For example, our examination of trends in the test score levels of the Grade 8 students is based on the same actual years of testing (1990-1995) as our examination of the trends for Grade 2 students, but we need to keep in mind that these cohorts are six years apart, as indicated by the staggering of the lines in the figures. We cannot easily infer that when the current crop of Grade 2 students finish Grade 8, in the year 2001, we will see the same system differences as we see with the Grade 8 students tested in 1995.

Figure 5 uses the same data as Figure 4, but compares System A and System B in terms of year-to-year gains in test scores. The information is separated according to Reading, Mathematics, and Science, and the gains are calculated for Grade 3 to 4, Grade 5 to 6, and Grade 7 to 8. No gain can be calculated for Grade 2 since the testing begins at that level.

The differences between grade and subject area and the trends across years (cohorts) are smaller for gains than for average score levels. The distinction between System A and System B is not very noticeable in the gains. In particular, the variability from year to year within a grade and subject seems to be about as large as the differences between systems and the trends. The gain comparisons, are not readily interpretable and usable. That is, the 1) year-cohort variations are about as large as any 2) substantive effects of education. We believe that the variations are technical and errorful; they have to do with test calibration and error. (This interpretation is supported by the evidence of correlation between the systems in the variation across cohorts within a given grade and subject area: in each year the same new, somewhat miscalibrated test form is used in both systems.) Looking again at the mean differences in Figure 4, a similar amount of year-to-year variability is apparent, but the substantive trends and differences are relatively larger.

Exactly these year-to-year gains are highlighted in the current TVAAS reports for schools and systems. The instability of the gains relative to their differences and trends doubtless is the cause of much of the controversy about them: To the extent that the variation is random, it bound to be uninterpretable. The differences from year to year in gain are large but apparently not very meaningful. Our preference would be to regard this year to year fluctuation as part of the error in gain measurement (the empirical standard errors). If this were done, the relatively small differences and trends in gain would be recognized as insignificant. The differences and trends in average

score levels would still provide important interpretations.

If gains are to be reported to the public, we suggest that they should be reported in a way that makes their variability apparent. One method is to depict them in a form suggested in section 6 for teacher reports—that is, plotted as empirical s.e. confidence bars on a horizontal scale with all other schools in the district. Alternatively, yearly gains of all system schools could be presented on a control chart that included the three-year moving average gain for each school. The extent of temporal instability of the estimates would then be directly apparent and consistently high or low schools, if any, could be identified visually.

Perhaps the best course to take would be to report both levels and gains, along with a clear explanation of what they each mean and imply. It would be necessary to explain that, even at the end of second grade, districts and schools will vary in their mean levels of test performance; some will be higher ranking and some lower ranking. Although a low ranking school may show gains at the state or national average at every grade, its relative position will show little or no change and it will remain below average in terms of average test scores. To improve its position, it would have to show well above average gains. The focus on gains is valuable, however, because it makes clear on a grade-to-grade basis the task facing a low-ranking school and its teachers.

Similar considerations have a bearing on the question of how standards are set in terms of gains. If the objective of state educational policy is to move the distribution of school average test scores upward while maintaining the existing separation between schools, then an overall standard for gain above the current state mean should be set. If the objective is to achieve greater equality of school test performance, then higher goals for gain must be set for lower ranking schools, and resources may have to be increased or reallocated to reach the objective.

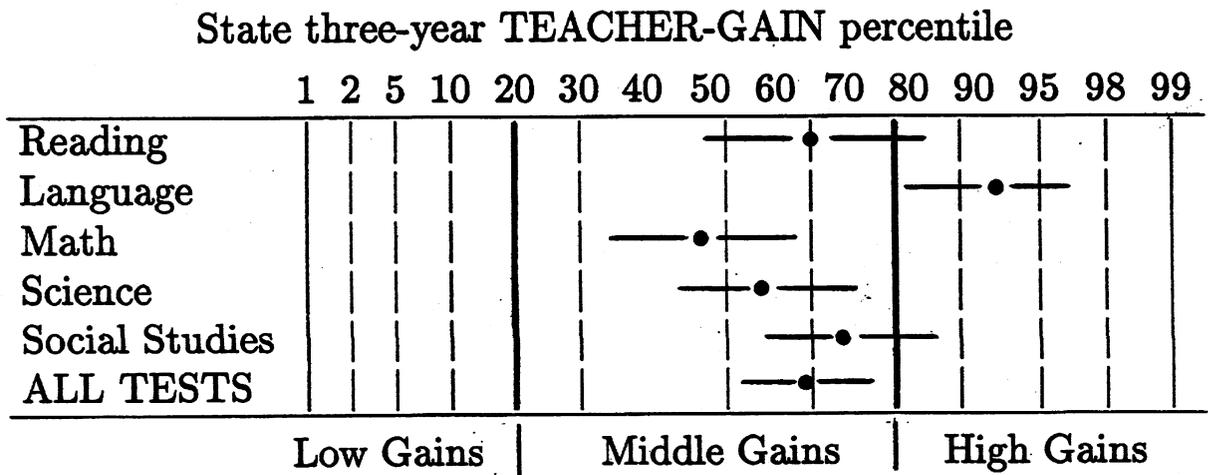
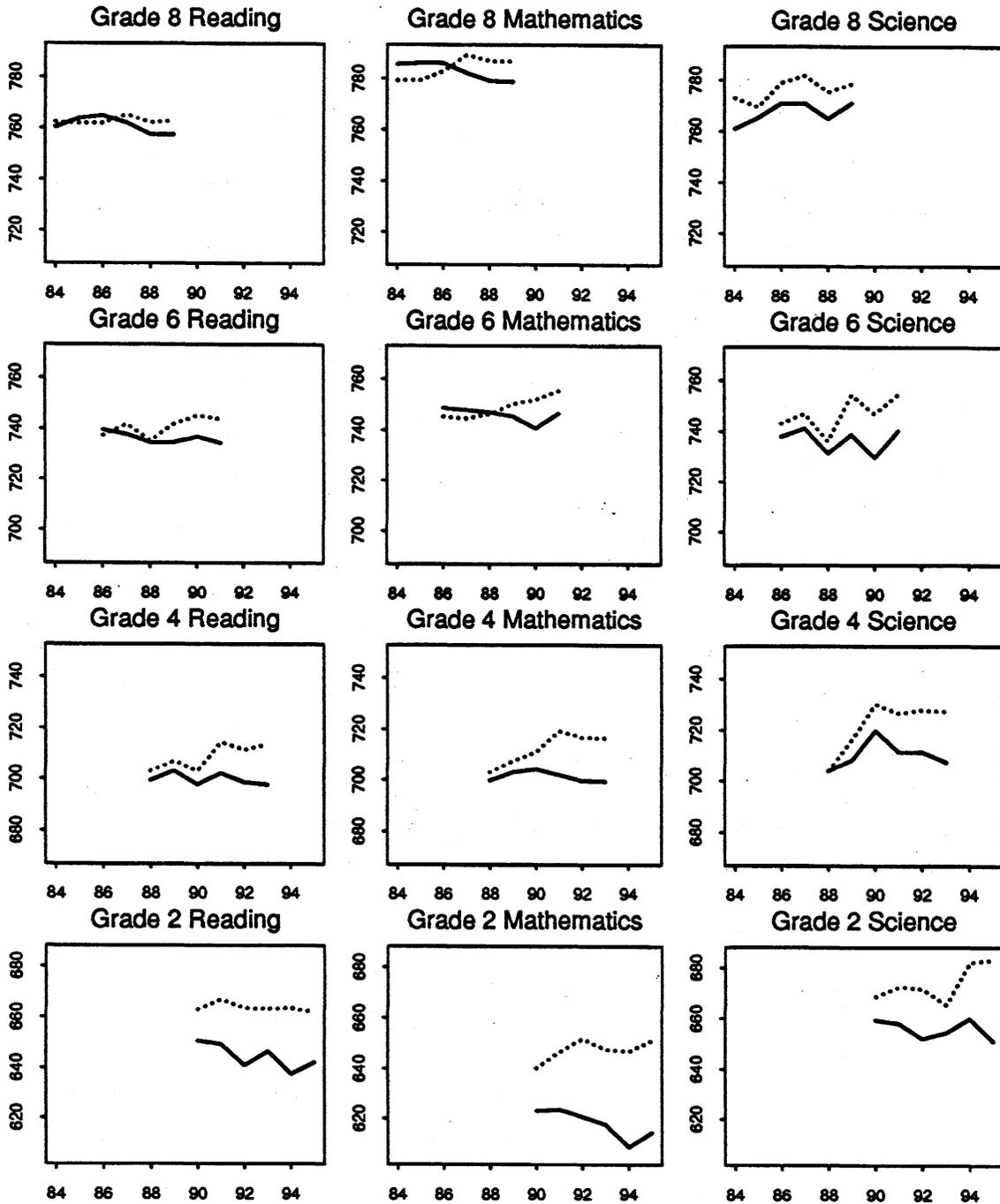


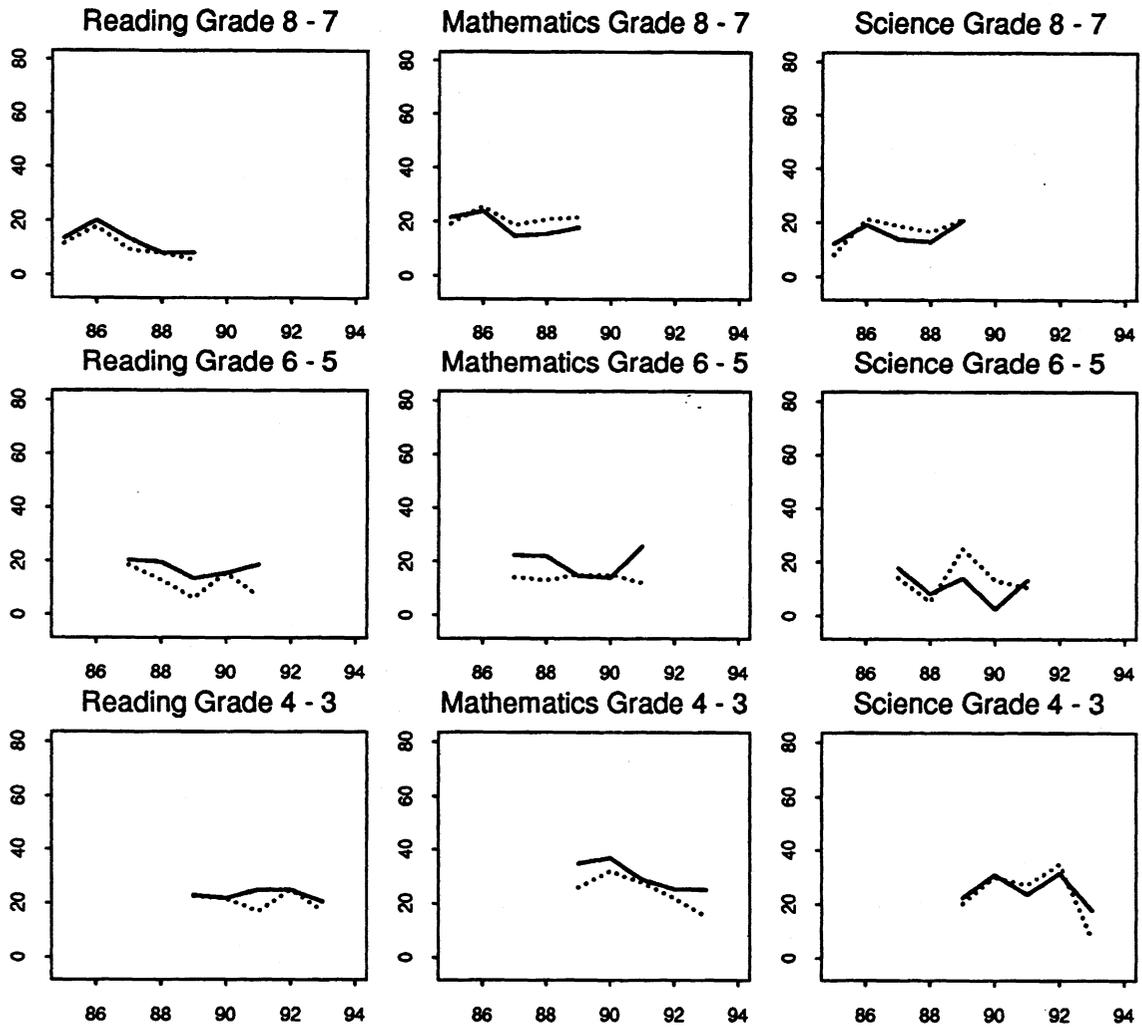
Figure 3. An example of a gains report for a fourth-grade teacher

Figure 4. Comparison of Average TCAP for Two Systems



System A (solid) and System B (dotted)

Figure 5. Comparison of Gain in TCAP for Two Systems



System A (solid) and System B (dotted)

Chapter 7

Conclusions and Recommendations

On the basis of a review of the assumptions and data analysis procedures of the Tennessee Value-added Assessment System, analysis of its operating characteristics, and the quality of the source data, we have the following conclusions and recommendations.

7.1 Conclusions

1. We agree with the central concept of the assessment system that the only present fair, objective, and dependable method of evaluating teacher effectiveness based on test scores is the measurement of achievement gains shown by students during the period a teacher is responsible for their instruction in the subject-matter measured by the test.
2. The educational data collection and management system implemented for TVAAS, in combination with the Tennessee Comprehensive Assessment Program annual achievement testing in grades 2–8, is virtually unique among the states in its ability to keep a continuing record of students' achievement test scores as they move from grade to grade or school to school in each county of the state. According to our audit of these data presented in section 4, the completeness and timeliness of the data acquisition is adequate for the purposes of the assessment system, but completeness could be improved along the lines suggested in

recommendation number 1.

3. The TCAP achievement tests supplied by CTB/McGraw-Hill in annually updated forms for the reading, language, and math subject areas have measurement and scaling properties suitable for measurement of annual student gains; however, the science and social studies tests have too few items for dependable use in this application. The successive annual forms of these tests, and probably the other three tests, show some evidence of imperfect equating of difficulty between years in certain subjects. Steps suggested in recommendation number 2 need to be taken to assure a high level of accuracy in the equating in order to reduce instabilities from year to year in estimating system, school, and teacher gains. These include 1) using item response data from the previous TCAP testings to rescale the test forms by the nonequivalent-groups equating base on common items in tests given in different years, and 2) equating forms by random equivalent-groups equating by inserting a certain number of test forms from previous years in the current year's testing. For low-stakes reporting of student score-level, and school mean score levels, however, forms are well enough equated.
4. We find the TVAAS statistical model for teachers effects on student achievement gain plus students individual differences in gains to be reasonable and entirely consistent with similar hierarchical models widely used in educational studies to represent outcomes of instruction. However, the TVAAS model represents teachers' contributions to gains, not in terms of difference between students' achievement scores the previous year and the teacher's current year, but as differences between the previous year and the average of the teacher's current year and two following years. Inasmuch as the teacher is not directly responsible for student gains in those following two years, we believe this feature is inconsistent with the basic principle of the value-added assessment system. It is included in the model for purposes of increasing the stability of the teacher evaluation, but we suggest below that other steps should be taken for this purpose that do not reduce the sensitivity of the results to the gains for which the teacher is responsible.
5. The TVAAS computerized procedures are in good agreement with other estimation procedures applied to the same data. The TVAAS estimation

procedures are highly general and could be specialized and simplified without impairing their accuracy.

6. Our analysis of the TVAAS reports of gain estimates for *schools* in the last three years indicates that the model-based standard errors reported for three-year average gains are underestimates of empirical year-to-year variability. Our analysis of variation in an 11-county sample of schools led to the same conclusion. The most likely reason for the discrepancy is that variation due to teacher effects is not included in the school model; however, the possibility that gain-score outliers among small schools may be inflating the empirical estimates of variability should be investigated.

The presence of substantial year-to-year instabilities in the annual school gain estimates and their three-year averages makes clear interpretation of the school gain reports difficult. Any public reporting of school gains should therefore be accompanied by graphical displays, such as described in section 5.2, that convey their year-to-year variability. In contrast, reports of average score levels would be much more stable between schools and better understood by the public; but their graphical display would also be helpful.

7. Both the analysis of TVAAS estimates of *teacher* gain effects and our own analysis of a large sample of gain scores in teacher classrooms showed that, although these estimates were also variable from year to year, the results were stable enough to permit identification of teachers with notably meritorious or problematic instructional effectiveness, as measured by test-score gain.
8. We question the wisdom of the TVAAS use of standards for gains based on mean gains from the 1988 National Normative Study conducted by CTB/McGraw-Hill. Statewide means for achievement in the various subject-matters at certain grade levels differ from those norms in ways that will introduce systematic discrepancies between estimated teacher and school gains and the standard. In addition, the TVAAS use of standard errors of estimated gains to signal departure from the standard does not distinguish between the practical size of the estimated gain and the accuracy of its estimations, the latter of which is affected by classroom size and other factors irrelevant to teacher performance. We suggest below an alternative method of reporting teacher gains estimates.

7.2 Recommendations

1. The quality of the data on which the value-added assessment is based should be improved by the following steps: 1) the number of items in the science and social studies tests should be increased from twenty to forty to match the test length of the reading, language, and math norm-referenced tests; 2) students should be assigned a uniform identification code upon entry into the school system—possibly the student's social security number or pseudo-number, where necessary, or an up to twelve-alphabetic character unique name or designation of the student's own choosing; 3) teacher reporting of students' time in their charge should be simplified by generated student rosters for each teacher containing student and teacher identification and requiring only a single mark per student to indicate the percent of time spent with teacher.
2. To improve the quality of the test scores, equating of test difficulty of successive forms of the TCAP tests should be improved to meet the stringent demands of gains estimation. This can be accomplished by:
 - 1) CTB test forms D through F can be rescaled to a form C bench mark by performing nonequivalent-groups equating based on the 25 percent of common items in successive forms. The original item response files from the TCAP assessment are required for this purpose. The results from the rescaling will provide corrections for forms D, E, and F. A similar procedure should be carried out on form G before the 1996 assessment results are reported. (See section 2.1)
 - 2) equivalent-groups equating can be carried out by randomly inserting, statewide, 1,500 to 2,000 copies of the forms from previous years' testings among the copies of current year's tests. Scores on the current year's tests may then be adjusted so that the mean and standard deviations of scores from the inserted forms equal those of the previous year's testing (see section 2.1).

Either of these equating methods can be used to adjust the summary results for the 1993 through 1996 assessments.

3. The statistical model for teacher effects should be modified to reflect students' achievement gains only during the time the teacher is responsible for their instruction.
4. Estimates of gain attributable to teachers should be computed and reported to responsible authorities only, and steps should be taken to increase the accuracy of the estimates by improving the equating of test forms, insuring adequate local control of test administration, and improving the completeness of data acquisition.
5. Teacher gains should be reported and judged from graphical displays, or their numerical equivalent, similar to those used in reporting scores to students. The high and low intervals representing notable or problematic levels of gain should be set by the state Department of Education as described below. The accuracy of the teacher's gains estimate should be represented by confidence intervals about the respective three-year average gain score estimates (see Figure 3).
6. Reports of assessment results for school or systems should emphasize score levels rather than gains. Any school or system average gains reported to the public should be accompanied by graphical displays that show, for each grade and subject matter, the extent of variation in gain between schools or systems and the instability of the gain score estimates from year-to-year. Either the confidence bar plots on the control charts described in section 6.2 can be used for this purpose.
7. Standards for teacher gains should be set by the state Department of Education based on consideration of the distribution of teacher gains for each test at each grade level in the state as a whole and in the light of information on national gains and score levels for the test. Percentile points for classifying high and low ranking teachers must be set at realistic levels in practical terms. It should be confirmed that they are well above and below the national median average gain of the CTB norms for students. These classification rules should be reviewed annually in relation to the validity information described in the following recommendation.
8. The use of results for individual teacher gains by the state Department of Education should be limited to an advisory role until the 1998 assess-

ment. Until that time, the operating characteristics of the system and the standards for classifying teacher performance should be evaluated annually. To obtain information for this purpose, the state department of education should request school systems to have school principals identify teachers who other evidence indicates are notably effective or ineffective in instruction. These nominations should be made before the annual assessments are reported. Concordance between these judgements and teacher gain estimates will be evidence that the system is performing satisfactorily. Demonstration of satisfactory operation of the system should be a prerequisite to its mandated administrative use.

References

- Aitkin, M., & Longford, N. (1986). Statistical modeling issues in school effectiveness studies (with Discussion). *Journal of the Royal Statistical Society, Series A*, **149**, 1–43.
- Anderson, T. W. (1958). *An introduction to multivariate statistical analysis*. New York: Wiley.
- Bock, R. D. (1985 reprint) *Multivariate Statistical Methods in Behavioral Research*. Chicago: International Educational Services.
- Bock, R. D. (Ed.) (1989). *Multilevel Analysis of Educational Data*, San Diego, CA: Academic Press.
- Bock, R. D. & Zimowski, M. F. (1966). Multiple group IRT. In Hambleton, R. & van der Linden, W. (Eds.) *Handbook of Item Response Theory*. New York: Springer-Verlag, p.p. 425–439.
- Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Newbury Park, CA: Sage Publications.
- Dempster, A. P., Rubin, D. B., & Tsutakawa, R. K. (1981). Estimation in covariance component models. *Journal of American Statistical Association*, **76**, 341–353.
- Goldstein, H. I. (1987). *Multilevel Models in Educational and Social Research*. London: Oxford University Press.
- Graybill, F. A. (1961). *An Introduction to Linear Models*. New York: McGraw-Hill.
- Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of American Statistical Association*, **72**, 320–338.
- Henderson, C. R. (1984). ANOVA, MIVQUE, REML, and ML Algorithms for Estimation of Variances and Covariances, in *Statistics: An appraisal*.

- Proceedings 50th Anniversary Conference, Iowa State University Statistical Laboratory*, H. A. David and H. T. David. (Eds.) Ames, IA: Iowa State University Press.
- Kelly, T. L. (1947). *Fundamentals of statistics*. Cambridge, MA: Harvard University Press.
- Laird, N. M., & Ware, J. H. (1982). Random effects models for longitudinal data. *Biometrics*, **38**, 963–974.
- Longford, N. T. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested effects. *Biometrics*, **74**, 817–827.
- Longford, N. T. (1994). Random coefficient models. Oxford Statistical Series No. 11. London: Clarendon Press.
- Longford, N. T. (1996). Models for uncertainty in educational testing. New York: Springer-Verlag.
- Lord, F. M. (1980). Applications of item response theory to practical testing programs. Hillsdale, NJ: Erlbaum.
- Lord, F. M. & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed score “equating”. *Applied Psychological Measurement*, **8**, 453–461.
- Peterson, N. S., Kolen, M. J. & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational Measurement*, 3rd ed., pp. 221–262. Washington, D.C.: American Council on Education.
- Pothoff, R. F., & Roy, S. N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, **51**, 313–326.
- Roy, S. N. (1957). *Some aspects of multivariate analysis*. New York: Wiley.
- Schulz, M. E. & Nicewander, W. A. (1995). *Grade Equivalent, Thurstonian, and IRT representations of growth*. Iowa City: American College Testing Program.

- Wiley, D. E. & Bock, R. D. (1967). Quasi-experimentation in educational settings: comment. *School Review*, **75**, 353–366.

Appendix A.
CTB/McGraw-Hill
construction of successive
forms of the TCAP tests.

INTEROFFICE MEMORANDUM

DATE: October 26, 1995

TO: Wendy Yen

FROM: Gary Dudney & Richard Schwarz

SUBJECT: Item Selection Procedures for the Tennessee Comprehensive Assessment Program (TCAP)

Item Selection Pool

The *Comprehensive Tests of Basic Skills, 4th Edition (CTBS/4)* and the *California Achievement Tests, 5th Edition (CAT/5)* are the two test series that were created to be strict parallel forms of each other. They are parallel in terms of item writing specifications, objective coverage, and all important psychometric measures as described below. CTBS/4 and CAT/5 each have two "shelf" editions of the Survey, and two "shelf" editions of the Battery, which is a longer test. In addition, CTBS/4 and CAT/5 have in common one particular test edition called the "Benchmark". The Benchmark was used to place the CTBS/4 and CAT/5 items on the same scale. Thus, despite the use of the different names for the CTBS/4 and CAT/5 test series, their items taken collectively comprise the CTB item pool. This pool is referenced to the CTBS/4 scale to provide the continuity required for the TCAP program.

TCAP Forms

Existing short forms of CTBS/4 and CAT/5 were used to supply the norm-referenced portions of the Tennessee Comprehensive Assessment Program (TCAP) test battery for the first four years of the program. The following table shows which short forms of CTBS/4 and CAT/5 were used for TCAP:

<u>TCAP Form</u>	<u>CTB Short Form</u>
A	CTBS/4 Survey, Form A
B	CTBS/4 Survey, Form B
C	CAT/5 Survey, Form A (on the CTBS/4 scale)
D	CAT/5 Survey, Form B (on the CTBS/4 scale)

Beginning with TCAP, Form E, fresh short forms of twenty norm-referenced items per subtest per level were selected from the CTB pool for each TCAP subtest. These forms were all constructed to be strictly parallel following the procedures described below.

Appendix A
The Item Selection Procedure

Several rules governed the selection process for Forms E through G concerning which items were to be given preference. One rule was to give preference to all items that had not just appeared in the previous TCAP Form. Another rule was to limit to 25% the total number of items that had previously appeared in any TCAP form. And another rule was to spread the items selected evenly across all four editions of the complete battery. These rules assured that relatively few items would reappear in subsequent forms so that students and teachers could not become familiar with them. This also insured that no one form of CTBS/4 or CAT/5 would be overly represented. The pools were sufficiently robust in almost all cases to avoid violating these rules during selection. The Tennessee Department of Education and the University of Tennessee were consulted and informed on all the rules to be followed and conformity to these rules.

The item selection process was accomplished through the use of CTB's item selection software called ITEMSYS. This program is a tool for constructing tests with pre-determined statistical characteristics from a pool of calibrated items, taking into account various constraints, including the item parameters, content representation, number of passages or stimuli, item fit, item bias, and previous usage. ITEMSYS makes an interactive connection possible between content experts and the item database. This allows the editor to adjust content and receive immediate feedback on the statistical consequences of those item selections. ITEMSYS is particularly good at helping a content expert construct comparable test forms, which was the basic task in selecting the test components for successive forms of TCAP.

The same criteria that is applied to the selection of CTB's norm-referenced achievement test series were applied to the TCAP selections. Floor and ceiling targets were established by CTBS/4. Nominal ranges for each level of the test were based on the CTBS/4 nominal ranges. The nominal range went from the scale score at the 5th percentile in the Fall for the earliest target grade for that level to the 95th percentile in the Spring. Expected number correct scores for a given scale score were compared across the nominal range for a previously used target form and the current TCAP selection. The two expected number correct scores could not depart from one another more than 0.5 for any scale score within the nominal range with very few exceptions. Selections were required to maintain ordinality across the appropriate range of scores for each test level. Standard error curves (i.e., reciprocal of the square root of the test information function) for the current selection and the target form were also compared in order to obtain a close match. This also ensured that the SEM curve was minimized throughout the score range embraced by each test level.

During the development of the items comprising the CTB item pool careful attention was given to questions of ethnic, racial, and gender bias. Reviewers representing various ethnic groups identified items that they considered to reflect possible bias in language, subject matter, or representation of people. Such material was eliminated or was modified. The Linn and Harnisch (1981) procedure was used to detect item DIF.

Editors were instructed to minimize DIF by avoiding the selection of items exhibiting DIF. Item selections for a given grade and content level had to show less DIF than the item pool from which they were selected.

Objective coverage within each subtest was monitored so that it matched the objective coverage in CTBS/4 or CAT/5 proportionally within a few percentage points. Any other content nuances in CTBS/4 or CAT/5 were matched in the TCAP selection. (For the norm-referenced selection, there was no attempt to match Tennessee curriculum. Editors were instructed to pay no attention to Tennessee curriculum as items were selected for these tests.) After selection was complete, copies of the selected test items were sent to Tennessee for confirmation. Occasionally, the review committee in Tennessee objected to items for content reasons and these items were either defended or replaced. Replacing items entailed going back into ITEMSYS and being held to the original criteria once again for the new selection.

Finally, testbooks were produced according to our usual criteria and reflected formats common to the CAT/5 and CTBS/4 series. In an effort to minimize context effects, items were assigned a position in their respective subtests based on their position in the source textbook, that is, items appearing near the beginning of the source test were placed near the beginning of the TCAP test. For this purpose, the test was divided into thirds.

Appendix B. State Mean Scores Smoothed Over Cohorts

On the assumption that statewide averages of test scores should change smoothly with respect to successive cohorts of students, we have fitted quadratic curves to the data in Tables 1 through 5. Adding further refinements to the model, such as a term for cubic trend, did not produce a statistically significant improvement in fit.

Smooth values for the state averages are shown in numerical form in Table B1 and in graphical form in Figure B1. Differences between the observed and fitted values (residuals) indicate possible problems in forms equating. The residuals are shown in Table B2, where values greater than plus or minus four are highlighted as large enough to merit attention. All residuals are represented in Figure B1 as vertical projections from the smoothed curves. The possible equating problems discussed in section 2 are apparent in the residuals, including the greater number of large values in the twenty-item science and social studies tests.

Table B1
Smoothed TCAP Scores by Subject, Form, and Grade

Subject	Grade	Fitted Values					
		1990 A	1991 B	1992 C	1993 D	1994 E	1995 F
Reading	2	652.4	652.7	652.9	652.9	652.6	651.9
	3	682.1	682.4	682.7	682.9	682.9	682.6
	4	704.0	704.3	704.6	704.9	705.1	705.0
	5	723.6	723.8	724.1	724.4	724.7	724.9
	6	740.0	740.0	740.2	740.6	740.9	741.2
	7	750.8	750.5	750.6	750.8	751.1	751.4
	8	762.5	762.0	761.7	761.8	762.0	762.3
	Language	2	673.1	674.3	675.1	675.5	675.5
3		695.4	697.0	698.2	699.0	699.4	699.4
4		710.5	712.3	713.9	715.1	715.9	716.3
5		727.7	729.8	731.6	733.1	734.3	735.1
6		737.4	739.7	741.8	743.6	745.1	746.3
7		746.7	749.1	751.4	753.5	755.3	756.8
8		758.9	761.5	763.9	766.2	768.3	770.1
Mathematics		2	631.9	633.3	634.1	634.2	633.4
	3	680.9	682.5	683.9	684.8	684.9	684.1
	4	707.2	709.0	710.6	712.0	712.9	713.0
	5	730.1	731.7	733.5	735.2	736.6	737.4
	6	749.0	750.3	752.0	753.8	755.5	756.9
	7	761.7	762.5	763.9	765.6	767.3	769.0
	8	780.0	780.1	780.9	782.3	784.0	785.7
	Science	2	664.7	665.6	666.4	667.0	667.4
3		690.9	691.9	692.9	693.6	694.2	694.7
4		714.1	715.3	716.3	717.2	718.0	718.6
5		728.8	730.1	731.2	732.3	733.2	733.9
6		740.9	742.3	743.5	744.7	745.7	746.6
7		754.4	755.8	757.2	758.4	759.6	760.6
8		765.5	767.0	768.4	769.7	771.0	772.1
Social Studies		2	662.7	663.7	664.2	663.9	662.8
	3	696.4	697.7	698.7	699.2	698.9	697.8
	4	718.9	720.4	721.7	722.7	723.2	722.9
	5	737.9	739.4	740.9	742.2	743.2	743.7
	6	744.7	746.0	747.5	749.0	750.3	751.3
	7	751.1	752.0	753.3	754.8	756.3	757.6
	8	763.7	764.1	765.0	766.3	767.8	769.3

Table B2
Residual TCAP Scores by Subject, Form, and Grade
(Residuals greater than ± 4.0 are in bold type)

Subject	Grade	Residuals					
		1990 A	1991 B	1992 C	1993 D	1994 E	1995 F
Reading	2	.0	2.3	-1.5	2.3	-2.8	-.4
	3	1.3	-5.6	2.9	-1.4	.5	2.3
	4	-3	-3	-3.5	3.0	.1	1.0
	5	.0	-3.4	4.4	2.5	.5	-4.0
	6	.7	.4	-1.5	-4	.9	-.1
	7	.2	-3.6	.4	2.9	2.5	-2.4
	8	-8	.8	3.2	.5	-1.5	-2.1
	Language	2	-4.3	-2.9	5.9	3.3	-3.1
3		2.6	1.3	-1.0	-2	-2.2	-6
4		-1.0	-6	-8	.5	-6	2.5
5		-1.5	-5.5	6.5	4.2	-1.0	-2.7
6		-1.6	1.8	2.1	2.1	-3.6	-9
7		1.7	-4.1	.6	1.0	1.0	-.1
8		.6	-1.7	2.5	4.5	-4.3	-1.6
Mathematics		2	-4	1.8	2.4	-1.6	-4.4
	3	-2	-6.0	.6	2.9	2.7	.0
	4	-9	-5	2.2	1.7	-1.5	-1.2
	5	-5	.1	2.8	.6	-1.8	-1.2
	6	1.0	-2.8	1.1	1.3	-1.5	.9
	7	-5	-3.4	2.8	1.7	-2	-5
	8	1.0	-1.0	2.1	.2	.2	-2.5
	Science	2	2.8	.2	-1.2	-5.3	4.2
3		2.2	-4	.5	-6.0	5.2	-1.4
4		-3.5	-4.2	5.8	2.9	.2	-1.3
5		1.1	-5	-1.2	-1.0	3.9	-2.3
6		3.7	1.2	-4.6	4.2	-8.4	3.9
7		3.1	-2.3	1.5	-2.2	-5.0	5.0
8		-1.5	-2.6	1.5	3.2	-3.1	2.5
Social Studies		2	-6	1.2	8.8	-5.1	-3.2
	3	-4.7	-5.3	4.4	2.8	.7	2.0
	4	-4.7	2.3	2.1	3.2	-7.5	4.6
	5	.6	2.8	2.3	-6.9	1.3	.0
	6	4.4	3.5	-2	-8.4	-3.2	3.9
	7	-1.3	-3.3	4.5	1.5	.4	-1.7
	8	-2.7	1.7	-2	3.8	-2.3	-4

Figure B1. Fitted and Residual Scale Scores

