# Using a Longitudinal Student Tracking System to Improve the Design for Public School Accountability in California

Edward H. Haertel[1]
Stanford University
August 2005

Abstract

This discussion paper considers possible state accountability designs, especially "Value-Added Models" relying on the California Longitudinal Pupil Achievement Data System (CALPADS). It is intended for a general audience of educational professionals and policy makers, and to inform the deliberations of the PSAA Advisory Committee. Technical challenges are described but are not discussed in depth. Three basic state accountability designs are distinguished and their strengths and weaknesses are briefly described. The paper concludes, as have other studies of the issue, that accountability designs based on tracking individual students' gains have sufficient promise and potential to warrant continued investigation. However, substantial technical hurdles stand in the way of statewide implementation within the next few years, especially for high-stakes accountability purposes in a state as big and complex as California. As with many technical innovations, while potential benefits are real, they may fall well short of some popular claims and expectations.

[1] Edward H. Haertel is a Professor in the School of Education, Stanford University. He chairs the Technical Design Group (TDG), which has advised the State of California since 1999 on technical issues in the design and implementation of the API. Dr. Haertel also chairs the API Subcommittee of the Public Schools Accountability Act (PSAA) Advisory Committee. Comments on an earlier draft by the staff of the Policy and Evaluation Division, California Department of Education, are gratefully acknowledged. Any remaining errors or omissions are the responsibility of the author. The opinions expressed are solely those of the author, and no endorsement by the Technical Design Group, the PSAA Advisory Committee, or any other individuals or institutions should be inferred.

# Using a Longitudinal Student Tracking System to Improve the Design for Public School Accountability in California

Edward H. Haertel
Stanford University
August 2005

Over the years, various accountability designs have been implemented statewide for California's public schools. Details have varied, but all have had one thing in common: None has ever relied on linking individual students' test scores from one year to the next. Thus, while some designs have employed year-to-year gains in test scores at the school level, California has never used a design that required information about the annual test score gains of individual students. An obvious, fundamental requirement for calculating individual students' year-to-year gains is some way to match up each student's current-year scores with his or her own test scores from one or more previous years--a "longitudinal student tracking system." Senate Bill 257 mandates the implementation of such a statewide student tracking system, known as the California Longitudinal Pupil Achievement Data System (CALPADS). Under CALPADS, Statewide Student Identifiers (SSIDs) will be required on all state assessments beginning in the 2005-06 school year. Thus, it is important to consider how these identifiers might be used to improve the design of California's statewide school accountability system. In particular, considerable interest has been expressed in "value-added" models, incorporating measurements of individual students' growth over time (e.g., Doran & Izumi, 2004).

This discussion paper provides an overview of design options for California's statewide school accountability system, with particular attention to designs taking advantage of CALPADS.[2] The alternatives are described in general terms. Many details would have to be specified, and significant technical challenges would arise, in the actual implementation of any such system. Indeed, significant challenges have arisen, and have been met, in the process of implementing and maintaining California's *current* school accountability system. One message of this paper, however, is that designs that rely on linking individual student records over two or more years would bring much greater challenges. The potential benefits of such models certainly justify further theoretical exploration and perhaps district-level pilot studies, but great caution is urged in moving toward statewide implementation of a high-stakes accountability system using such a model.

## Three Basic Designs for Accountability Systems

Three basic designs can be distinguished for using test data in school accountability systems. In this paper, these will be called Uniform Target (UT), Successive Cohort (SC), and Individual Growth (IG) designs. Distinctions among the three designs are

[2] Discussion in this paper is limited to designs for *school* accountability. "Value-added" models that estimate individual *teacher* effects are beyond the scope of this paper, and are not addressed.

illustrated in Figures 1, 2, and 3 below.[3] Two of these designs are now being used in California. The SC design is used when schools are required to meet annual growth targets established for the Academic Performance Index (API) and the UT design is used when schools are required to meet the State's Annual Measurable Objectives (AMOs) in order to demonstrate Adequate Yearly Progress (AYP). "Value-added models" are all varieties of IG designs.

Figure 1: Uniform Target (UT) Design

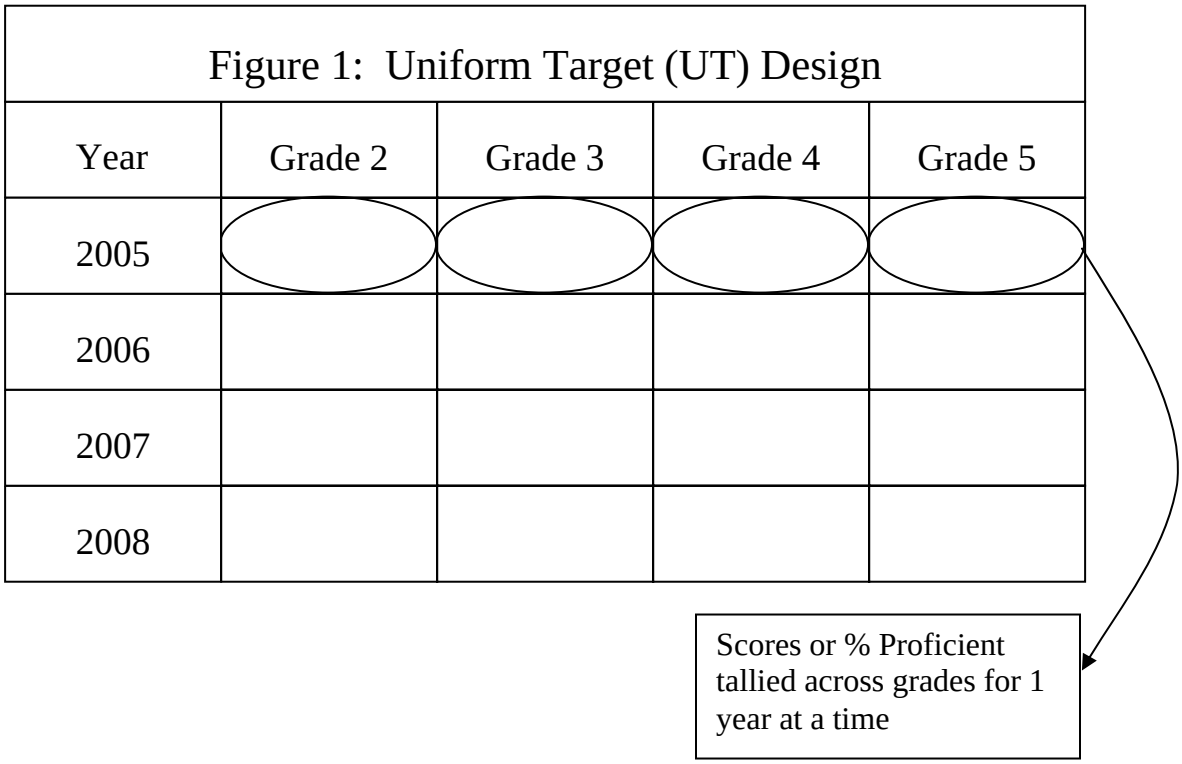


Figure 1 illustrates Uniform Target (UT) designs. To understand this figure, imagine an elementary school with students in grades two through five tested annually. The rows represent test scores from successive years and the columns represent scores from students at successive grades. The ovals show that, for the UT design, data are considered for just one year at a time. Average scores (e.g., reading scores) for each grade (second-grade, third-grade, etc.), or percents of students meeting some standard at each grade (e.g., percents with reading test scores at or above the proficient level) are accumulated over grade levels to arrive at a single number describing the entire school. That number is then compared to some target, which is typically the same (uniform) for all the schools in the state in any given year.[4] The target may rise over time. Thus, schools might be held to a higher performance target in 2008 than in 2007. The UT design might or might not involve combining test scores in different subject areas. The same calculations might or might not also be carried out separately for demographic subgroups within the school. There could be many variations. The *defining feature* of all

[3] Gong (2004) used diagrams like Figures 1, 2, and 3 to explain these designs, although he called them by different names.

[4] There may be separate targets for different subject areas; for elementary, middle, and high schools; etc.

UT designs is that test data for just *one year* at a time are considered. Such designs are also referred to as *cross-sectional* models.

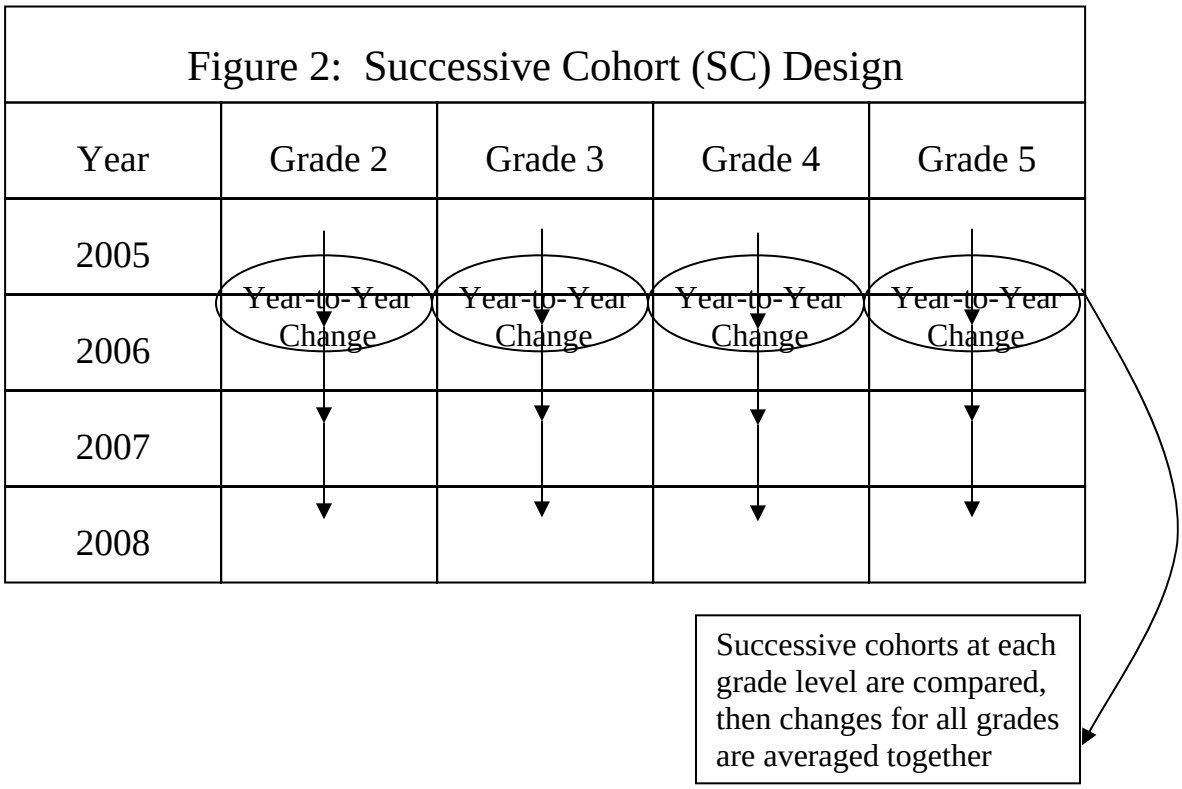


Figure 2 illustrates Successive Cohort (SC) designs. Unlike the UT design, the SC design is a *growth model*, with a focus on year-to-year *change* in school performance. As in Figure 1, the rows and columns represent test scores from children tested annually in each of grades two through five. In Figure 2, however, vertical arrows have been added. These represent comparisons of scores earned by *successive cohorts* of second graders, third graders, etc. Thus, the scores of second graders in 2006 are compared to the scores of second graders in 2005, and similarly for other grades. The average score in 2005 might be subtracted from the average score in 2006 to determine how much higher (or lower) students at that grade level were scoring in 2006. Or, the percent proficient in 2005 might be subtracted from the percent proficient in 2006 to determine whether a higher proportion of students were meeting the standard in 2006. These *school-level change scores*, however they were computed, would be combined across grades to arrive at a single number for the entire school.[5] That number would then be compared to some *growth* target, which might be specific to that school. The school's overall change score might also be compared to the distribution of change scores for some collection of schools. Note that the ovals representing this accumulation of scores across grade levels each cover a span of two years. The SC design might or might not involve combining test scores in different subject areas. The same calculations might or might not also be

[5] Calculations for the API are organized a little differently, but are mathematically equivalent to this description. Scores are first accumulated across grades for each year separately, and then each school's growth is calculated by subtracting the previous year's "Base API" from the current year's "Growth API."

carried out separately for demographic subgroups within the school. There could be many variations. The *defining features* of all SC designs are first, that test data for *two years* at a time are considered, and second, that direct comparisons are made between the scores of *different groups of students*. The children who are second graders in 2006 are a different group from those who were second graders in 2005. Such designs are *longitudinal* at the *school level*, in that each school is observed for two years in a row. However, there is no linkage of individual students' scores over time.

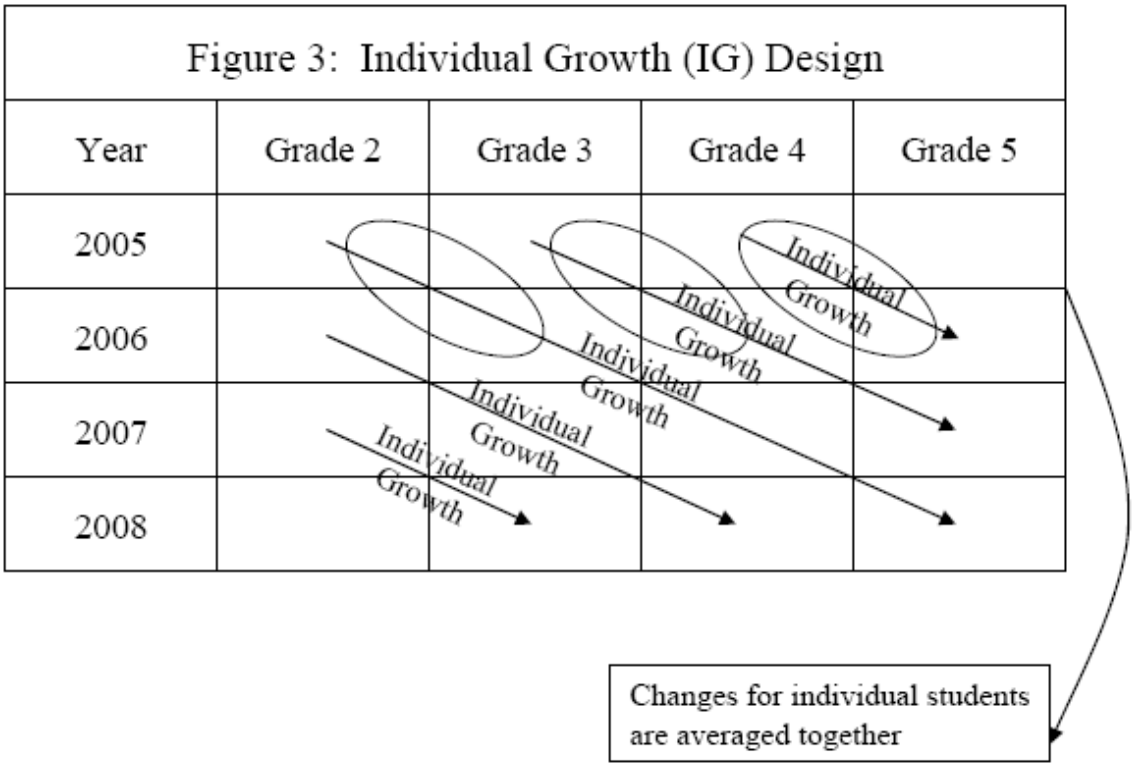


Figure 3 illustrates Individual Growth (IG) designs. Like the SC designs, the IG design is a *growth model*. Whereas the SC design focused on change at the *school* level, the IG design focuses on change at the *individual student* level. As before, rows and columns represent scores from an elementary school, organized by grade and year. In Figure 3, however, diagonal lines have been added. These represent the tracking of *individual students* from one grade to the next. For example, most students in second grade in 2005 would be in third grade in 2006, fourth grade in 2007, and fifth grade in 2008. With IG designs, actual scores would typically be compared, rather than proficient/not proficient classifications. Different IG designs might track students for just one year (previous year to current year) or might reach back further in time to incorporate scores from two, three, or even more years earlier. The ovals in Figure 3 illustrate a simple IG design using data from just two years. Note that there are only three ovals in Figure 3, compared to four ovals in each of Figures 1 and 2. With this design, Grade 2 data from 2006 and Grade 5 data from 2005 would not be used in the year 2006 calculations. Only three groups of students could be tracked, from grades 2 to 3, 3 to 4, and 4 to 5. Individual third-, fourth-, and fifth-graders' gains since the previous year are accumulated to arrive at a

single number for the school. That number might then be compared to a target or to the distribution of overall gains for some collection of schools. The growth target might be specific to the school. The IG design might or might not involve combining test scores in different subject areas. The same calculations might or might not also be carried out separately for demographic subgroups within the school. There could be many variations. The *defining features* of all IG designs are first, that test data for *two or more years* at a time are considered, and second, that growth is measured for *individual students*. Such designs are *longitudinal* at the *student level*. These are the only designs that require the year-to-year matching of individual students' scores that CALPADS is intended to provide.

**Questions addressed by the three designs**. These alternative designs may be expected to identify different schools as high- versus low-performing, because they address fundamentally different questions (see Figure 4). The UT design addresses the question, "How high are students scoring?" The SC design addresses the question, "Are students scoring higher this year than last year?" The IG design addresses the question, "How much are individual students' test scores increasing from year to year?"

| **Figure 4: Questions Addressed by Each Design** | |
|---|---|
| Uniform Target (UT) | "How high are students scoring?" |
| Successive Cohort (SC) | "Are students scoring higher this year than last year?" |
| Individual Growth (IG) | "How much are individual students' test scores increasing from year to year?" |

The basic UT design does not take account in any way of differences in the characteristics of the students various schools serve. It takes a snapshot at one point in time. Because student performance is influenced by out-of-school factors as well as school policies and practices, those schools ranked highest under a UT design are likely to be among those serving the most advantaged students.

The SC design uses each school's prior-year performance as a baseline for the current year's performance. Schools ranked highest will be those that have made the most improvement since the previous year. Note, however, that any school that is in a "steady state" would be expected to show zero growth with a SC design. Students in one school, call it "School A," might be making large year-to-year gains, while those in another school, say "School B," might be making tiny gains. But if School A does the same excellent job two years in a row and School B does the same poor job two years in a row, they will appear identical under the SC design, except for random variations due to imprecision in test scores and year-to-year fluctuations in the distributions of proficiencies among the students each school serves. Each would be expected to show the same year-to-year growth, namely zero.

The IG design uses each individual student's previous performance as the baseline for that student's current-year performance; some IG designs incorporate demographic

variables, as well. Thus, with IG designs, schools serving more- versus less-advantaged student populations should be on a more even footing compared to UT designs. In addition, because the IG design focuses on individual students' year-to-year growth, it would distinguish between School A and School B from the previous paragraph. The schools ranked highest under an IG design should be those in which students' test performances exhibit the largest gains from one year to the next, perhaps adjusting for out-of-school factors. In general terms, the highest-ranked schools would be those in which third graders were furthest along relative to second graders, fourth graders were furthest along relative to third graders, etc.

On close examination, it is not entirely clear which of these designs should be preferred. It might be argued that the UT model should be rejected on the grounds that schools should be held accountable for the effectiveness of their programs and practices, not for the influences of out-of-school factors beyond their control. For example, in their review of the Dallas value-added system, Thum and Bryk (1997, p. 102) concluded that "From a purely technical perspective, the arguments seem very clear: Anything other than a value-added-based approach is simply not defensible." But a counterargument might be made that the adjustments for different starting points inherent in SC and IG models in effect institutionalize lower expectations for traditionally lower-performing schools and students. If all students can reach the "proficient" level, the argument goes, then all schools should be held to the same expectation.

Choices among designs are complicated by the fact that different questions will matter to different audiences (Raudenbush, 2004; Rubin, Stuart, and Zanutto, 2004). A parent might want to know in which school an individual child is likely to do best. To answer that question, the UT design might be optimal. The parent is not concerned with disentangling peer influences, the effects of out-of-school factors, teacher quality, and school policies. Parents tend to seek out the schools with the highest test scores, period. The question for a state policy maker or a state board of education is more complex, because disentangling those influences is important. Consider the following possible questions, for example: "If all the schools in the state were serving the same, average mix of students, at the end of the year, which would have higher or lower scores?" Or, "How well is this school performing relative to other schools that serve the same mix of students as it serves?" Or, "How well is this school performing relative to other schools that have the same mix of teachers and serve the same mix of students?" Or, "If this school, with its current mix of students and teachers, adopted the policies and practices of that other school, would its students' achievement be more likely to rise, or to fall?" Or, "If the entire student body of this school were moved to that other school, would their achievement be expected to be better or worse than if they stayed where they are?" These various questions almost certainly have different answers, and posing one or another will likely result in different schools being singled out for rewards or sanctions. Furthermore, in deciding where to intervene (i.e., where to target resources intended for schools most in need of improvement), policy makers may wish to focus on schools that are not only poorly performing, but are also on a flat or declining trajectory over time. Those that are getting better year by year might best be left alone. Thus, the choice of an

appropriate accountability design requires great care in specifying what questions one wishes to answer.

**Differences in school rankings under the three designs**. As stated in the previous section, the UT, SC, and IG designs might be expected to identify different schools as high- versus low-performing. Such differences are illustrated in this section using data from the Los Angeles Unified School District (LAUSD). The LAUSD maintains a district-level student identification system, and has produced reports of year-to-year school-level gains derived from matched student-level data. This may be thought of as a very simple IG model, implemented at the district level.[6] These data, together with APIs and API component scores for the same schools, permit some illustrative comparisons among crude versions of the three designs.

In 2002, under the direction of Dr. Ted Bartell, the Program Evaluation and Research Branch of the LAUSD correlated two measures of school performance. These were both "growth models." One was calculated following a SC design closely related to schools' changes in their APIs from one year to the next (school-level longitudinal growth model). The other represented an IG design based on individual students' gains (student-level longitudinal growth model). The two measures were both based on data from the spring 2000 and spring 2001 administrations of the Stanford Achievement Tests, 9th edition (SAT-9).[7] Comparisons were carried out separately for 436 elementary schools, for 74 middle schools, and for 57 high schools in the LAUSD.

For the SC design, SAT-9 scores were summarized to the school level exactly as was done at that time to calculate the API, except that results were not combined across subject areas. This yielded separate school-level measures for reading, mathematics, language arts, and spelling (elementary and middle schools) or for reading, mathematics, language arts, social studies, and science (high schools). For the IG design, individual students' SAT-9 scores were converted to Normal Curve Equivalents (NCEs) in each subject area for 2000 and for 2001, and then merged using the LAUSD's student identification system. Individual-student NCE gains were then averaged to obtain school-level measures in each subject area based solely on matched cases (students in the system both years). Students who changed schools were included in the averages for the school they attended at the time of testing in 2001[8] The correlations obtained between these two measures of school growth are shown in Table 1.

---

[6] Aggregating year-to-year gain scores for individual students up to the school level is vastly simpler than the estimation of school (or teacher) effects using mixed models. The comparisons in this section do not use any models approaching the complexity of the "value-added" models implemented in Dallas (Webster & Mendro, 1997) or Tennessee (Sanders, Saxton, & Horn, 1997).

[7] The LAUSD individual-student gain calculations also used data from their district-level administration of the SAT-9 to first-grade students. Statewide testing began at grade 2. Recall that only three ovals were shown in Figure 3, indicating that only three student cohorts could be tracked longitudinally within a four-year grade span. Because the LAUSD had scores for first-grade students, the district was able to track four cohorts of elementary-school students, from grades 1 to 2, 2 to 3, 3 to 4, and 4 to 5, as shown in the calculations below.

[8] Dr. Ted Bartell, personal communication, 2002.

**Table 1. Correlations Between School Performance Measures Calculated Using SC Versus IG Designs**.

| Subject | Elementary (N = 436) | Middle (N = 74) | High School (N = 57) |
|---|---|---|---|
| Reading | .59 | .60 | .63 |
| Mathematics | .70 | .71 | .66 |
| Language Arts | .73 | .76 | .71 |
| Spelling | .65 | .70 | |
| Science | | | .66 |
| Social Studies | | | .45 |

Clearly, although the SC and IG designs are telling similar stories, schools would be ranked quite differently using one versus the other. This is true even though each school's scores under the two designs were derived from much the same data. There are several reasons why the rankings differ. First, only about seven-eighths of the data used for the two designs overlap.[9] Let 00G2 refer to Grade 2 test data collected in 2000 and similarly for other year-grade combinations. For elementary schools with a K-5 grade configuration, the change in API (SC design) can be represented as

$$(01G2 - 00G2) + (01G3 - 00G3) + (01G4 - 00G4) + (01G5 - 00G5)$$

or, equivalently, as

$$(01G2 + 01G3 + 01G4 + 01G5) - (00G2 + 00G3 + 00G4 + 00G5).$$

In other words, the grade 2 through 5 scores from 2000 are subtracted from the grade 2 through 5 scores from 2001. The IG design can be thought of as

$$(01G2-00G1) + (01G3 - 00G2) + (01G4 - 00G3) + (01G5 - 00G4).$$

Rearranging terms, this is the same as

$$(01G2 + 01G3 + 01G4 + 01G5) - (00G1 + 00G2 + 00G3 + 00G4).$$

Note that for the IG design, the lowest grade's scores from the current year and the highest grade's scores from the previous year drop out. Thus, the IG design is based on data from grades 1 through 4 in 2000 and grades 2 through 5 in 2001. The SC design is based on data from grades 2 through 5 in 2000 and grades 2 through 5 in 2001.

In addition, the IG measures are based only on those students who could be tracked and matched within the LAUSD from 2000 to 2001. The SC values would include

[9] As explained in an earlier footnote, the LAUSD tested first graders, and so NCE gains were available for second graders in 2001. If, instead, only grade 2-5 data had been used for the IG design, then the IG design for elementary schools would be based on just 75% of the full data used for the SC design, because the year-2000 grade 2 data and the year-2001 grade 5 data would drop out.

unmatched cases. Finally, correlations between SC and IG values would be lower because the IG numbers are based on NCEs and the SC numbers are based on a different scaling of test scores used for the API (so-called "progressive scoring weights")[10].

A helpful comparison of all three designs may be conducted using LAUSD elementary school data for 2000-2001 merged with elementary school APIs for the same two years. School APIs and matched-student NCEs were assembled for 412 elementary schools.[11] The NCEs were averaged across subject areas (Reading, Mathematics, English Language Arts, and Spelling) using the same weights as are used in the API. For this comparison, only matched-student gains from grades 2 to 3, 3 to 4, and 4 to 5 were used. Thus, these comparisons reflect the fact that gains since the previous year would not be available for second graders under the current statewide testing system.

Three measures were calculated for each school, representing the three designs. The 2001 API was used to represent a (cross-sectional) UT design. The matched-student NCE gain from 2000 to 2001 was used to represent a simple version of a (student-level longitudinal) IG design. The change in API from 2000 to 2001 was used to represent a (school-level longitudinal) SC design. Correlations among these three measures are shown in Table 2.

**Table 2. School-Level Correlations Using Alternative Accountability Designs**

| | UT | SC | IG |
|---|---|---|---|
| UT | 1.00 | -.20 | -.16 |
| SC | -.20 | 1.00 | .66 |
| IG | -.16 | .66 | 1.00 |

Note that there is a *negative* correlation between school rankings according to the cross-sectional UT model versus either the SC or the IG growth model. This is what would be expected if lower-performing schools were making greater test-score gains. The larger negative correlation between the SC and UT designs compared to the correlation between the IG and UT designs probably reflects the "progressive scoring weights" used with the API, which give greater credit for gains by lower-performing students. (The "progressive scoring weights" exaggerate the pattern of initially lower-scoring schools gaining faster.) The correlation of .66 between the IG and SC models is consistent with the corresponding within-subject elementary-school correlations reported in Table 1, which were .59, .65, .70, and .73.

---

[10] Progressive scoring weights and other details of API calculations are explained in the annual *Information Guides*, which can be accessed by following links from http://www.cde.ca.gov/ta/ac/ap/.

[11] LAUSD data may be extracted from publicly available pdf files, which are organized by District (A through K) within the LAUSD. For District A, the report is located at http://www.lausd.k12.ca.us/lausd/offices/perb/files/pau/SAT901/LD_GainReport2001_A.pdf. Replacing "A" with other letters ("B" through "K") gives the urls for the remaining reports. API files may be downloaded following links from http://api.cde.ca.gov/datafiles.asp. Matching is complicated by the fact that the LAUSD pdf files do not include the CDS (county-district-school) codes used by the State of California. The fact that only 412 cases were matched, versus 436 elementary schools reported in Table 1, does not appear to have materially affected the findings presented in Table 2.

**Strengths and weaknesses**. There is much to be said about the strengths and weaknesses of these different kinds of designs; a brief overview may be helpful. More details about IG design choices are presented later in this paper.

Uniform Target (UT) designs are easy to understand and to implement, because they use data from just one year at a time. Also, if properly implemented, these designs may challenge lower-performing schools to make greater efforts (or may encourage allocation of greater resources to lower-performing schools), because those schools must improve more to reach the target. The biggest negative for UT designs may be their failure to account for factors out of the school's control. "How high are students scoring?" may not be the right question for an accountability system to ask, because the answer depends on demographics, peer effects, etc., in addition to the school's policies and practices. Ignoring factors that schools are unable to control may give rise to a perception that schools serving low-performing students are being penalized unfairly. In general, then, UT designs do a poor job of establishing challenging but attainable annual improvement goals for all schools. If the uniform target is regarded as completely unrealistic for a given school, then rather than challenging that school to greater effort, the accountability system may just create discouragement, depressing student morale and giving teachers and principals incentives to quit or move elsewhere.

Successive Cohort (SC) designs place greater demands than UT designs on the stability of the testing system over time, because year-to-year comparisons of successive cohorts' test scores are only sensible if they have taken the same test form or, better, alternate forms for which scores can be expressed on the same scale. SC designs are highly flexible with regard to which school subjects are tested at which grade levels, however, provided only that the same testing schedule is followed in successive years. Because each school serves as its own baseline, demographic differences are incorporated in a natural way, although "control" for demographic differences is still far from perfect. Growth targets for different schools can provide incentives to reduce achievement disparities between schools, as with California's API growth targets, which are more ambitious for lower-performing schools in order to bring all schools up to at least the interim statewide performance target over time.[12] If a school's demographics change significantly from one year to the next, however, growth targets based on a previous year's scores may be inappropriate. SC designs are predicated on the assumption that all schools (or all but the highest-performing) ought to be doing better each year than the year before, and can be crafted to do a reasonably good job of setting challenging but realistic annual improvement goals for schools across the spectrum from low- to high-performing. However, by focusing solely on year-to-year school-level growth, SC designs fail to distinguish among schools that are more versus less effective at any one point in time, like School A and School B mentioned earlier.

Individual Growth (IG) designs place by far the strongest demands on the testing system. Not only do they require matching individual student records across years, but they also

[12] API growth targets are calculated as five percent of the difference from a school's base-year API to the interim statewide performance target of 800.

generally require that students be tested at every grade level in all subject areas included. Unlike SC designs, IG designs also generally require a "vertical scale." Recall that with SC designs, second-grade test scores are compared to second-grade test scores; third-grade test scores are compared to third-grade test scores, and so on. Thus, it doesn't matter if there is any linkage between the score scales used for tests at different grades. With IG designs, second-grade scores are subtracted from third-grade scores, third-grade scores are subtracted from fourth-grade scores, and so on. These subtractions are only meaningful if the two scores involved are on the same scale.[13] Strong and difficult-to-test assumptions must be made about the linearity and dimensionality of such scales over grade levels. In mathematics, for example, early-grade achievement tests may focus on arithmetic problem solving and later-grade tests may involve more abstract mathematical concepts. It is not clear that taking the difference between scores on tests with such different contents yields a meaningful result, even if both are placed on the same "vertical scale" (Reckase, 2004).

If the strong assumptions of IG designs can be met or approximated to a satisfactory degree, the benefits of these designs may be substantial. Because individual students are tracked over time, a school's scores and rankings under an IG model should be less influenced by demographic shifts caused by student mobility. IG models can do the best job of disentangling out-of-school factors from the effects of school policies and practices. For that reason, they may come closest to answering questions of greatest interest to educational policy makers. Because individual students' prior test scores provide some statistical control for their prior learning, IG designs have the potential to provide more accurate information than UT or SC designs about which schools are bringing about larger versus smaller year-to-year achievement gains. There is general agreement, however, that school-to-school comparisons (e.g., rankings) will be most accurate if the schools being compared are highly similar. Even the most sophisticated IG models should not be trusted to provide fair comparisons between schools serving very different student populations.[14]

## The Current Accountability System in California

The present accountability system for California's public schools was created in response to the mandates of California Education Code section 52050, the Public Schools Accountability Act of 1999 (PSAA), and modified to comply with the requirements of Public Law 107-110, the No Child Left Behind Act of 2001 (NCLB). Currently, schools are subject to both sets of requirements. PSAA requirements use a SC design. An Academic Performance Index (API) is calculated annually for each school, and year-to-year changes in the API serve as the basis for determining rewards and sanctions. (Each

[13] It is convenient to think about each student's prior-year score being subtracted from the current-year score, and in some IG models, this is just what is done. "Difference scores" are used as dependent variables in statistical models. In many IG models, however, more complicated statistical procedures are used to adjust current-year performance for prior performance. More complicated models may adjust for performance on tests in other subject areas and/or for scores from multiple prior years. With some of these more complicated models, there is no requirement for a vertical scale.

[14] In Tennessee's Value-Added Assessment System, for example, statistical models are fit at the level of school districts. There are no cross-district comparisons.

year's API is also used to determine the annual statewide decile rankings of schools, overall and relative to schools with similar characteristics, as required by the PSAA legislation.) NCLB requires a UT design. Adequate Yearly Progress (AYP) is determined by a school's attainment of a specified percent at-or-above Proficient. Required percents proficient in reading and in mathematics differ for elementary, middle, and high schools, and are referred to as Annual Measurable Objectives (AMOs). The AMOs must increase to 100 percent proficient by 2014. Both the API and the AYP determinations are derived from the same student test scores, although some tests at some grade levels are used in one system and not the other. The API is a single number obtained as a weighted combination of test scores for different subject areas. NCLB requires that test performance in reading and in mathematics be treated separately.[15]

Both systems place primary reliance on the California Standards Tests (CSTs). These are grade-specific or course-specific tests in English-language arts (including direct assessment of writing at grades 4 and 7), mathematics, science (grade 5 and 9-11), and history-social science (grades 8, 10, and 11). The CSTs were created for the State of California and are aligned with the State's Academic Content Standards for each grade level or course. A judgmental process was used in deciding upon cut scores representing Below Basic, Basic, Proficient, and Advanced performance levels on each test.[16] These judgment-based cut points were then adjusted somewhat to reduce fluctuations from one grade to the next in the percents of students at or above Proficient, and similarly for other performance levels. Finally, scale scores were constructed for each test in such a way that the Basic cut point was mapped to a scale score of 300 and Proficient to 350. (Scale scores representing Below Basic and Advanced vary somewhat on CSTs for different subject areas and grade levels.) Because these tests were constructed and scaled independently of each other, there is presently no vertical scale for CST scores across grade levels. A feasibility study for the vertical linking of the CSTs in English language

[15] Details of both the PSAA and NCLB systems are much more complicated, and distinctions between them are not quite so clear as this summary implies. Both systems include requirements for high levels of student participation in mandatory testing, in addition to requiring specified levels of school-level year-to-year growth (PSAA) or annually specified levels of performance (NCLB). Also, both systems set targets for "numerically significant" subgroups defined by race/ethnicity and socioeconomic disadvantage, as well as for the school as a whole. NCLB also treats English Language Learners and Students with Disabilities as separate subgroups. While NCLB places primary reliance on uniform AMOs for all schools (the UT design), it does take limited account of year-to-year growth (SC design), in that schools where one or more numerically significant subgroups, or the school as a whole, fail to meet an AMO may still show "Adequate Yearly Progress" (AYP) based on sufficient improvement over the previous year (the so-called "Safe Harbor" provision). Conversely, while PSAA relies primarily on year-to-year improvement in the API (SC design), it also incorporates a uniform statewide target (UT design), in that schools with APIs at or above the interim statewide performance target of 800 are required to make zero or minimal growth. In both systems, additional provisions dealing with small schools, schools of special kinds, or schools that have undergone significant year-to-year changes in demographics add further complexities. Rules for inclusion of students with disabilities and the treatment of scores of former English Language Learners who have been redesignated Fluent English Proficient (RFEPs) also differ. The API includes some tests (e.g., history-social science as well as some science tests) that are not included in NCLB requirements. NCLB incorporates graduation rate as an additional criterion. The two systems are not entirely distinct, in that the API serves as an "additional indicator" for NCLB. Further details may be found by following links from http://www.cde.ca.gov/ta/ac/.

[16] Students not reaching the Below-Basic cut score are designated Far Below Basic.

arts (grades 2-11) and mathematics (grades 2-7) was conducted by Educational Testing Service in January 2005. While not definitive, that study strongly suggested that vertical scaling of the CSTs would be problematical. Note the use of course-based CSTs in mathematics at grades 8-11 (General Mathematics, Algebra I, Geometry, Algebra II, Summative High School Mathematics) and in science at grades 9-11 (Earth Sciences, Biology, Chemistry, Physics, and four different Integrated Science tests). The fact that different students are tested on different content might make it nearly impossible to construct a valid and technically sound vertical scale in mathematics beyond grade 7. Vertical scales in science and in history-social science are likewise ruled out by the current design of the CSTs.

In addition to the CSTs, both the PSAA and NCLB accountability systems also make use of the California Alternate Performance Assessment (CAPA) taken by severely disabled students. The API also incorporates scores from a norm-referenced test, the California Achievement Tests, 6th edition (CAT-6), at grades 3 and 7 only. At the high school level, the API also incorporates results from the California High School Exit Examination (CAHSEE), which is taken by all students in tenth grade, and subsequently by those who must retake one or both sections to earn a passing score.[17] While the inclusion of these additional tests in California's current SC design does not pose any difficulties, it does not appear possible to calculate year-to-year student-level growth measures for any of them.

## Implementation Challenges for Accountability Designs Using CALPADS

California's current accountability system has evolved steadily since it was introduced pursuant to the PSAA of 1999. There is considerable value in keeping the system as stable and predictable as possible; change in accountability requirements is disruptive to schooling practices. Schools that enter into a school improvement process based on one set of criteria may be justifiably concerned if the rules then change, so that exiting the improvement process is based on different criteria. For all schools, a shift from the current design to an IG design using CALPADS would be highly disruptive. It would require legislative changes at the State level. At the federal level, it would require change in the law, change in regulations, or waivers of current NCLB requirements to proceed. Because not all subjects are tested at all grade levels, and because vertical scales appear problematical for the CSTs and impossible for the CAPA, the CAHSEE, and norm-referenced tests used only at isolated grade levels, moving to an IG (value-added) accountability system could narrow the range of information sources and subject areas incorporated in the API, and would probably entail significant changes in the Standardized Testing and Reporting (STAR) system underlying California's accountability system.

In this section, some of the technical challenges in creating a "value-added" (IG) accountability model for California are described in general terms. These are summarized in Table 3, which is patterned after a similar table presented by McCaffrey,

[17] Because California's NCLB implementation uses the API as an additional indicator, AYP determinations could also be viewed as using CAT-6 and CAHSEE results indirectly.

**Table 3. Issues in Value-Added (IG) Models for School Effects[a]**

**Basic Issues of Statistical Modeling**

- How should the State of California be stratified (segmented) for purposes of statistical modeling? (E.g., district by district?) If strata include multiple districts, should district-level effects be incorporated?
- Should separate models be fit for each grade level?
- Should the nested structure of students within classes be incorporated in models?
- Should school effects be specified as fixed or random?
- Should gain scores be modeled, should prior test scores be entered as covariates, or should the full multi-year score vector be used as the dependent variable?
- How should school influences be apportioned for students who change schools?
- How many summative indexes of school performance (across subject areas and grade levels) should be created, and how? Should they measure actual performance against expectations? Should they reflect current status or trajectory?

**Issues Involving Confounders, Omitted Variables, and Missing Data**

- How many prior years of data should be included?
- How should missing data be dealt with? Should incomplete records be included?
- Should student-level background variables be included as covariates, or should prior test scores alone be used to adjust for level and trajectory of prior learning?
- Should (school-level) means of student-level background variables be included as covariates to reduce the influence of context effects?
- What school-level background variables, if any, should be included?
- Test scores from longer ago are less informative about current achievement. Should the function representing this decay be specified a priori or should it be modeled?

**Issues Arising From the Use of Achievement Test Scores as Dependent Measures**

- Should the CSTs be replaced to provide defensible vertical scales? How should score inflation unrelated to broad achievement gains be controlled?
- What subject areas should be included?
- Should scores for different school subjects be modeled separately or jointly?
- Should scores from multiple content areas be used in covariance adjustments?
- Should out-of-level testing be permitted? If so, how should test level be determined?
- How should schedules for year-round schools be accommodated?

**Uncertainty About Estimated Effects**

- How should total uncertainty in school effects be estimated?
- Should alternative models be compared?
- What sensitivity analyses are required?
- How should uncertainty be communicated to teachers, principals, policy makers, and the public?

[a]Adapted from McCaffrey, et al. (2004, p. 53)

et al. (2003, p. 53). Questions within and across the four categories of Table 3 are interrelated. Answers to some questions will constrain the range of possible answers to others.

**Basic issues of statistical modeling**. The first category of issues in Table 3 includes basic decisions that would need to be resolved concerning the kind of value-added model to implement, the way the State's schools would be divided into smaller groups for purposes of analysis, and the sorts of variables that would be included.

No value-added model has ever been implemented for a single system of schools approaching the size of the State of California.[18] As noted in footnote 14, the Tennessee Value-Added Assessment System (TVAAS), perhaps the best-known system of this kind, models school and teacher effects within each Tennessee school district separately. If districts were used as units, special rules would probably have to be devised for the State's 300 or so single-school districts. If districts were grouped together, the question would arise as to whether district-level effects should be modeled. Not only would fitting a model for all of California be computationally prohibitive, but in addition, most experts would be very reluctant to trust the statistical assumptions required in creating such a model for a collection of schools as heterogeneous as those across the State of California.[19]

Because the assumptions of "vertical scales" are more tenable across narrower grade spans than broader grade spans, it might be best to fit separate models for each grade level. However, doing so would discard some information that could be used in a cross-grade analysis. Also, as noted later in this subsection, separate analyses would give rise to separate school effect estimates, which would then probably need to be combined in some way to arrive at a school-level index.

A key assumption of *causal* models for school effects is the Stable Unit Treatment Value Assumption (SUTVA) (Rubin, 1986). In the present context, this is an assumption that a school's influence on the test scores of a given student does not depend on what other students are also assigned to that school. Not only does this assumption imply that there are no differential peer effects, but it also implies that teachers would not change their instruction if the students in their classes had a different mix of achievement levels. For this reason (and others), value-added models might better be regarded as *descriptive*

---

[18] A simple IG model like the one illustrated using LAUSD data, which just aggregates individual gain scores for matched students, could be implemented statewide. However, such a system would fall far short of affording the benefits touted for value-added systems employing more complex statistical procedures. Moreover, even the simple IG model illustrated for the LAUSD could be highly problematical. It would require narrowing the range of information sources and subject areas included, and would be more difficult to justify using the current CSTs than using the SAT-9 scores available in 2000-2001, before the CSTs were developed. (Even though NCEs, rather than SAT-9 scale scores, were used in the LAUSD example, at least the SAT-9 was a coherent system of tests designed to support a vertical score scale.)

[19] For example, Rubin, Stuart, & Zanutto (2004, p. 109) summarize one of their concerns with a system implemented in a Florida school district by saying, "If school A has no students who 'look like' students in the other schools, it is impossible to estimate the effect of school A relative to the comparison schools without making heroic assumptions."

rather than *causal*. However, if no causal claim is made, i.e., if schools are not regarded as *responsible* for student achievement patterns, then rewards and sanctions become more difficult to justify. Several things can be done to render the SUTVA assumption somewhat more tenable, but only at the cost of increasing the complexity of the statistical model. SUTVA would be better approximated if the nesting of students within classrooms within schools were modeled explicitly. Note, however, that this would require that CALPADS not only track the assignment of students to schools, but also track the linkage of students to particular classes (or teachers) within schools. Team teaching, changes in teachers in the course of the year, and reassignments of students to different classes within schools might all compromise the accuracy of such a system. Also, incorporating teacher identifiers might be seen as opening the door to additional uses of the accountability system, which might encounter resistance. Another way to render SUTVA somewhat more plausible, again at the expense of complicating the model, would be to average student covariates up to the school level for use as school-level indicators of peer context.

A value-added model would need to specify school effects as either fixed or random. The fixed-effects choice is probably easier for non-experts to understand--An effect is estimated for each specific school, without regard to how that school compares with others. With a random-effects model, each school's estimated effect is "shrunk" toward the mean for all schools. The degree of shrinkage depends on the amount of information available about the school. Thus, smaller schools' estimates are moved further toward the mean than are larger schools' estimates. Unfortunately, fixed-effects models are likely to over-estimate individual schools' effects and random-effects models are likely to under-estimate them. Although on balance a random-effects model is probably more defensible, it is also more difficult to explain.

There are three fundamentally different approaches for value-added models, with different strengths and weaknesses. A full discussion is well beyond the scope of this paper, but the three choices, briefly, are (1) modeling students' gain scores (current year minus prior year) as the dependent variable;[20] (2) using current-year scores as the dependent variable and entering scores from one or more previous years as covariates; or (3) treating the entire collection of scores, across years, as a single dependent variable. None of these three is entirely satisfactory. As might be expected, the tradeoffs involve model complexity, the strength of the assumptions required, and the degree to which bias (distortion) in estimated school effects can be controlled. Options (1) and (2) would probably need to be implemented separately by subject area, again increasing the number of separate school indices that would need to be combined in some way.

A value-added model would also need to incorporate one of several approaches for apportioning the contributions of students who change schools during the year. This has been handled by various value-added models in at least two different ways. One approach is to assign each school a proportion of the responsibility for the student's end-of-year test score according to the proportion of the school year that student was enrolled

[20] The earlier illustration using LAUSD data represents the simplest possible model of this kind, in which a student's gain score is modeled as a fixed effect due to school plus a random error.

in the school (cf. McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004, p. 70). This seems unsatisfactory, as one would not expect the annual gain for a student who spends one-half year in each of two equivalent schools to equal the gain that student would have attained had the entire year been spent in just one school or the other. The alternative of excluding mobile students is also unsatisfactory, however, because mobile students are, on average, unlike those who stay in one school. As explained in the next subsection (concerning missing data), excluded scores for mobile students would violate the assumption that data are "Missing Completely At Random" (MCAR). Thus, excluding mobile students distorts comparisons among schools with different levels of student mobility.

As noted earlier in this subsection, separate grade-by-grade analyses place less reliance on questionable assumptions concerning vertical achievement score scales. However, such analyses would lead to separate grade-by-grade scores for each school. School rankings would vary across grades and (if subject areas were analyzed separately) for different school subjects. An issue would arise as to how these should be combined to construct a school-level index. Also, for accountability purposes, policy makers might be interested in changes over years. The "obvious" option of calculating an annual value-added index and then taking the difference between the current-year index value and the prior-year value may not be appropriate. If each annual index is a complex function of test scores over time, it may be difficult to figure out exactly what the *change* in such an index is actually measuring.

**Issues involving confounders, omitted variables, and missing data**. The second category in Table 3 concerns more specific choices that would need to be made about exactly which test scores and other variables should be included. The answers would involve considerations of data availability, cost, tenability of alternative model assumptions, bias and accuracy of estimates, and computational complexity.

If only two years of data were included (current year vs. previous year), there might be some question as to whether the benefits of an IG model outweighed the cost, disruption, and complexity of changing from the current system. As shown in Table 2, a model based on individual student gains might be expected to produce school rankings correlating about .66 with rankings based on year-to-year API growth under California's current accountability system. The gains in accuracy and comparability of schools under an IG model would increase if three or more years of data were incorporated, but tracking students across more years would exacerbate inevitable missing data problems. Also, with more years of data, estimation would be more complicated. Each additional year of data also reduces by one the number of grade levels at which the full set of prior scores is available. (Data from two years back would only be available beginning at grade 4, for example. Data from three years back would only be available beginning at grade 5, the highest grade level in many elementary schools.)

Missing data are a problem for any model, because the assumption that data are "Missing Completely at Random" (MCAR) is not tenable (Rubin, Stuart, and Zanutto, 2004). In brief, that means that lower-achieving students are more likely to have missing data. Any

analysis option that entails the MCAR assumption, therefore, will produce biased results. But methods that do not rely on that assumption are complex, and bring alternative assumptions of their own, which are also subject to challenge. Note that in a high-stakes accountability context, schools might also have an incentive not to work vigorously to track low-achieving students, depending on how models were implemented. (With some designs, incorporating a low-achieving student's low prior-year scores would improve the effect estimates for the current school.)

Arguments for and against various kinds of covariance adjustments are complex. Proper choices would depend on the precise formulation of the question the model was designed to answer, as well as assumptions as to whether, for example, low socioeconomic status is a student characteristic that is fixed over time or varies from year to year. Every additional covariate, of course, brings its own attendant missing-data problems.

Test scores from longer ago are less informative about current student achievement. In some statistical models (e.g., the TVAAS "layered" model), year-to-year effects are simply summed. Thus, a pupil's schooling two years before is assumed to be as powerful a determinant of current status as the pupil's schooling the previous year. More complex, multivariate models that estimate influences each prior year call that assumption into question (McCaffrey, et al., 2004). However, McCaffrey, et al.'s multivariate model is much more computationally intensive than models relying on simpler assumptions.

**Issues arising from the use of achievement test scores as dependent measures**. Several of the major issues in this category, dealing with the need for and problems with vertical scales, design limitations of the CSTs, and California's present pattern of testing different subjects at different grade levels were described earlier. Some additional issues referred to in Table 3 concern score inflation, separate versus joint modeling of different subject areas, out-of-level testing, and fair comparisons among schools with different numbers of instructional days prior to the date of testing (e.g., due to year-round school schedules).[21]

The score inflation issue refers to the likely consequence of a focus on raising test scores per se as a goal of schooling. Schools may make principled and well intended changes in curriculum and instruction that have the desired effect of raising scores, but also such undesired effects as narrowing the curriculum or reducing the predictive value of the test scores as indicators of achievement gains over some broader domain. One way to safeguard against such inflation is to change test forms frequently, or to replace annually a significant proportion of the items within test forms. This is done now, with California's current CSTs. Ongoing development of new forms and the consequent need for equating studies to maintain score scales are somewhat more complicated if it is also necessary to maintain a vertical scale over time. This is by no means an insurmountable problem, but would increase test development costs.

[21] By law, STAR testing must be conducted within a narrow window centered at the point where students have completed 85% of the school year. As a result, some students in schools with year-round programs are tested months after most statewide testing is finished. Varying dates of testing windows complicate any statewide accountability model, including the current system.

The choice between separate versus joint modeling of different subject areas is bound up with choices as to analysis of gain scores, covariance adjustments for prior-year scores, or full multivariate models. As before, choices involve tradeoffs.

Out-of-level testing arises as a concern once vertical scales are created. Since NCLB implementation, California has eliminated the option of giving low-achieving students lower-level test forms. The logic and rhetoric of current reform initiatives call for each student to demonstrate proficiency with respect to grade-specific learning objectives. Also, as a practical matter, the absence of vertical scales for the CSTs makes it very difficult to incorporate out-of-level data in analyses. If students are given tests that are much to difficult or much too easy for them, however, the resulting data are of poor quality. So-called "floor" and "ceiling" effects cause distortions because, for example, the score assigned to chance-level performance on an eighth-grade test may overstate the proficiency of a student actually performing at the third grade level. Also, using tests of inappropriate difficulty results in large standard errors of measurement. Allowing out-of-level testing is attractive, therefore, but brings its own logistical and policy challenges, e.g., deciding who should receive an easier or harder test, and preventing attempts to "game the system" by giving tests at inappropriate levels.

The issue of days of instruction prior to testing and inclusion of year-round schools under a "value-added" model would be much the same as with California's current accountability system. Care would be required in defining testing windows so as to accommodate all schools and enable fair comparisons.

**Uncertainty about estimated effects**. One consequence of all the choices and compromises required in specifying a value-added model for the State of California would be to complicate the problem of quantifying the precision of the resulting school effect estimates. This is by no means a trivial or unimportant problem. The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999, p. 31) require reporting of score reliability and standard errors as a matter of sound professional practice. Deriving estimates of precision within the context of one specific model would not be sufficient to fully address this concern. In addition, alternative models should be specified, with different possible assumptions, so as to determine the sensitivity of the results obtained to the particular assumptions of the chosen model. In addition to the problem of *quantifying* the accuracy of school effect estimates, there is also the problem of *communicating* findings about accuracy. End users must be given sufficient information to discourage over interpretation of small differences or other misuses of data.

## Conclusions

In a review of the Dallas value-added system, Thum and Bryk (1997, p. 108) concluded, "Substantial public rhetoric today demands initiatives of this sort [i.e., value-added, or IG approaches]. In our view, it it (sic) easy to do this badly but very hard to execute such a program well." McCaffrey, Lockwood, Koretz, and Hamilton (2003, p. 119) concluded

their extensive review of value-added models for *teacher* accountability on a similar note, stating that "The research base is currently insufficient for us to recommend the use of VAM [Value-Added Models] for high-stakes decisions. In particular, the likely biases from the factors we discussed in Chapter Four are unknown, and there are no existing methods to account for either the bias or the uncertainty that the possibility of bias presents for estimates." Rubin, Stuart, and Zanutto (2004, p. 113) sound a similar note, saying "We argue that models such as these should not be seen as estimating causal effects of teachers or schools, but rather as providing descriptive measures."

Even the simplest possible IG model for California, one that simply aggregated one-year individual-student gains to the school level, would pose serious logistical challenges. For the California Standards Tests (CSTs), there is no vertical scale on which such year-to-year gains could be defined. At best, CST gains over the previous year might be calculated for English Language Arts at grades 3-11 and Mathematics at grades 3-7, although the meaning of even these gain scores would be questionable. For CSTs in Mathematics at higher grade levels, in History-Social Science, and in Science, no gain scores would be possible. In addition, no gain scores would be available for the CAPA, the CAHSEE, or the CAT-6 used only at grades 3 and 7. Moreover, to approximate the benefits claimed for value-added models, analyses of vastly greater complexity than this simple aggregated gain-score model would be required. More complex models would probably need to be fitted separately for subsets of schools, greatly complicating any overall statewide rankings.

In this author's opinion, limited experimentation with these models, probably at the district level, and continued theoretical deliberation, are warranted. However, I believe that it would be a very serious error to rush into any large scale implementation of an accountability system relying on student-level longitudinal data at the present time or for at least several years to come.

## References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Doran, H. C., & Izumi, L. T. (2004). *Putting Education to the Test: A Value-Added Model for California*. San Francisco, CA: Pacific Research Institute. (Downloaded August 19, 2005 from http://www.pacificresearch.org/pub/sab/educat/2004/Value_Added.pdf.)

Gong, B. (2004, November 15). *Models for using student growth measures in school accountability* (Paper presented at the Council of Chief State School Officers "Brain Trust" on value-added models). Downloaded August 19, 2005 from http://www.nciea.org/publications/GongGrowthModels111504.pdf.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability*. Santa Monica, CA: The RAND Corporation.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics*, *29*, 67-101.

Raudenbush, S. W. (2004). What are value-added models estimating and what does this imply for statistical practice? *Journal of Educational and Behavioral Statistics, 29*, 121-129.

Reckase, M. D. (2004). The real world is more complicated than we would like. *Journal of Educational and Behavioral Statistics*, *29*, 117-120.

Rubin, D. B. (1986). Which ifs have causal answers? Discussion of Holland's "Statistics and causal inference." *Journal of the American Statistical Association*, *81*, 961-962.

Rubin, D. B., Stuart, E. A., & Zanutto, E. L. (2004). A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavioral Statistics, 29*, 103-116.

Sanders, W. L., Saxton, A. M., & Horn, S. P. (1997). The Tennessee value-added assessment system: a quantitative, outcomes-based approach to educational assessment. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* (pp. 137-162). Thousand Oaks, CA: Corwin Press.

Thum, Y. M., & Bryk, A. S. (1997). Value-added productivity indicators: the Dallas system. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* (pp. 100-109). Thousand Oaks, CA: Corwin Press.

Webster, W. J., & Mendro, R. L. (1997). The Dallas value-added accountability system. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* (pp. 81-99). Thousand Oaks, CA: Corwin Press.