Examining the Reliability of Accountability Systems[1]

Paper presented at the 2002 Annual Conference of
The American Educational Research Association,
April 3, 2002
New Orleans

A. Background

States have been developing high stakes accountability systems for their schools over the past decade. With the passage of the new "No Child Left Behind" (NCLB) legislation, every state will have a system in place by which schools' scores are evaluated, and significant consequences applied to those who have or have not met the standards established by the program.

There are three major issues that need to be addressed by any accountability design: validity, comparability, and reliability. Only the briefest overview of some issues related to the first two topics will be presented in this section; the third issue is the primary concern of this paper, and is discussed in more detail in the next section.

Validity involves drawing correct inferences from the data. One primary step in doing that is to start by asking a question consistent with one's definition of school "excellence." This is not done as often as one might expect, and therefore is a primary source of low validity of many accountability designs. There are at least three ways of defining acceptable scores:

A. Do students in the school score well?
B. Are the scores of students in the school improving?
C. Are the students learning sufficient amounts as they progress through school?

To answer the first question (Model A), one would look at the scores for the current year and compare them to some standard. For Model B, one would need scores from two years for students at the same grades and compare the two sets. For Model C, one would track the changes in scores for a cohort of students moving from one grade to the next and compare the changes to some standard.

Only Model C answers the question of how much students are learning as they progress through school. One would think, therefore, that answering this question would be at the heart of most accountability designs. That is not true, however; most states' accountability designs use Model A or Model B (as does NCLB), and therefore, the validity of these designs is subject to question. The choice of model is not an inconsequential issue; studies done in several states have shown low correlations (or even negative correlations) among how schools are ranked using the different models.

Comparability addresses the issue of whether scores across units, within year and across year, can be appropriately compared. Two major subcategories of this issue are comparability of populations and comparability of scale. Comparability of *populations* is important when comparing one school to

[1] Special thanks go to Ed Haertel of Stanford University, who was the original creator of many of the ideas presented in this paper. In particular, Ed provided the roadmap for the original study we did in California, all the equations for calculating standard errors and reliability, and helped refine the data generation process described in this paper.

another, or when comparing the results of one year to those of another. If, for example, a school has been permitted to exclude all its special education students from accountability one year, but must include them the next, it can be difficult or impossible to determine in which year educational achievement was higher. Comparability of *scale* is equally important. Historically, the general advice has been, "In order to measure change, don't change the measure." However, because there is virtual certainty that people will remember questions from year to year under these high-stakes conditions, most states feel they must change the questions they ask from year to year—thus creating a real Catch-22 situation; the measurement will change if the questions are not changed, but at least some doubt about the equivalence of measures will be introduced if they do change. In addition, most states use questions that require subjective scoring for which standards can change over time, and care must be taken to eliminate changes in scoring as a source of error. In short, making valid comparisons over time under these conditions is quite difficult.

Even if we considered the problems of validity and comparability to be solved, however, we still would have the issue that the group of students tested in a school each year is just a sample of the population of students that might be enrolled in that school. Some classes of students simply outperform other classes, even when seemingly being exposed to the same curriculum and instruction. As a result, one school might outscore another in a particular year, even though its long-term average would be lower than its comparison school. Similarly, a school might show gains or losses from one year to the next, not because of improvements (or lack thereof) in its program, but simply because it was serving a more or less able group of students that particular year. As a result, a school might get one classification one year and a different one the next, even though no real changes had taken place in its program.

If this happened infrequently, we might accept the occasional error involved as a necessary cost of implementing an accountability program. Most states have not studied the likelihood that their accountability system will make a classification error. The few that have have found those rates to be considerably higher than they would have expected. The remainder of this document will discuss what those error rates have looked like, suggest the changes that states can make in their accountability systems to reduce the error rates, and outline methods by which they can determine the classification error rate for their system.

B. Factors Affecting Reliability

When we observe a result for a school, it is a result that could change upon re-evaluation. For example, if 50 percent of the students pass our reading test this year, the result might have been 48 or 52 percent if we tested the same students again, since some students with reading skills near the cut-off point might pass one time and fail another. More importantly to accountability designs, if we tested the next class of students—*even if nothing changed in the instructional program in the school*—we might very well find that 45 or 55 percent of the students were passing the test. The result for a school for one year is just one observation from which to infer a school's *true score*—what the school's average would be if we could test an infinite number of students from the school's catchment area an infinite number of times on all the test questions that might be asked**.**

Even the most seemingly valid accountability design will be flawed if there is so much error in the system that the labels schools receive are largely based on random error. There are many factors that influence the reliability of an accountability design. Again, it is important to note that we are talking about accountability, not assessment; the various factors have different influence on the reliability of

an accountability system than they do on the reliability of assessment. In fact, as will be shown later in this paper, the reliability of the assessment has little relationship to the reliability of the accountability system.

The model chosen. Each of the models outlined in the previous section has a certain amount of unreliability associated with them. Model A designs are considerably more reliable than those for Models B and C, for two reasons: the effect size of Model A designs typically is larger, and the sampling error is smaller, since only one year of data is required in contrast to the minimum of two years for the other two models. While validity is a more important issue to address when selecting an accountability design, reliability cannot be ignored. If we ask the wrong question of the data, it hardly matters how precise that inappropriate answer is; but if we ask the right question and get an error-prone answer, we are hardly further ahead.

Effect size. A major factor in the reliability of an accountability system is the amount of effect the system is trying to detect. For a statistic with a given amount of error, large differences can be detected accurately, while small differences have so much error around them that true changes may be very hard to certify. Suppose, for example, a school's score on a certain variable is 50 and that there are 10 points of uncertainly around that score. If the school is supposed to score 75 on that variable, we can be pretty certain that results for other samples from that school, while they might vary somewhat from 50, would be consistently below 75. On the other hand, if the school's observed score is 72, we might very well find that the result for the next sample from that school was above 75. In that case, a different decision would be made about the school. The extent to which different decisions would be made about a school when the school has changed nothing to warrant the change in decision is the unreliability of the accountability system.

One of the reasons why Model A designs are more reliable than those for Models B and C is that the size of the effect we are trying to estimate is so much larger. Schools may vary by 50 percent or more on the percentage of students passing a test, but often we are trying to detect changes of 5 points or less when looking at gain. Differences this small are hard to detect accurately, and lead to accountability decisions for which the probability of an inaccurate classification is quite high.

Sample size. When the model and the amount of effect we are trying to detect are fixed, the reliability of the accountability system is determined by the amount of error around the observed statistic. The factor that has the most influence, by far, on the size of that statistic is number of students included in the scores for each accountable unit (that is, *sampling error* is a much larger factor than *measurement error*). Scores based on more students are more reliable than scores based on fewer students, and no other factor has as much influence on the reliability of those scores as the number of students tested. One way that states can improve the reliability of their accountability system is simply to test students at more grade levels.

Reporting statistic. The second most influential factor in the reliability of accountability systems is the reporting statistic. There are three basic ways scores typically are reported:

- Pass/fail, or the percentage of student passing
- Index scores, where information is used about the percentage of students in each category. An example of an index would be to have Below Basic scores count as 1 point, Basic as 2, Proficient as 3, and Advanced as 4. A school's score would be the average of these points.

- Mean raw scores or scaled scores, where no collapsing of the scores into categories is done.

The coarser the categories into which scores are placed, the less reliable the statistic, so it is no surprise that pass/fail scores are less reliable than index scores, while index scores are less reliable than mean raw scores. What has been a surprise from the analyses that the Center has done on data from other states is that, while pass/fail scores are substantially less reliable than index scores, index scores based on four or five categories (presuming there is a reasonable portion of the students in each category) are only slightly less reliable than mean raw scores.[2] Since scores can be communicated much more effectively through categories than through mean or scaled scores, our general advice is to use an index and accept the slight reduction that statistic brings to the reliability of the accountability system. Only in the most ideal situations (e.g., when data for extremely large numbers of students are available, as when computing a statewide average) is a pass/fail statistic sufficiently reliable.

On the other hand, one should not blindly trade reliability for validity. If the intent of the program is to identify the percentage of poor readers, then use of an index could mask a poor result. A school's index could increase by having its better students, rather than its poorer students, improve. In practice, this hasn't proved to be the case in most states; even when an index is used, increase in school scores generally are the result of disproportionately greater improvements by the poorer students. Also, a pass/fail statistic can focus instruction on just the students near the borderline, resulting in even less focus on the students who are most needy.

Before leaving this section, we should note that the reliability of a school score is not affected much by the reliability of individual student scores. Sampling error (the variability associated with choosing the particular group of students to be tested) is a far larger factor than measurement error (the error associated with the scores that the particular group of students gets).

To examine this point further, let's start with the following notation:

Let: σ^2_X = the variance of pupil observed scores,
σ^2_T = the variance of pupil true scores,
σ^2_E = the variance of error in pupil scores,
$\sigma^2_{\overline{X}}$ = the variance of school observed mean scores, if the population of students in the school's catchment area were tested
$\sigma^2_{\overline{X_0}}$ = the variance of school observed mean scores if one sample of size N were drawn for each school,
$\sigma^2_{\overline{T}}$ = the variance of school true mean scores, if the population of students in the school's catchment area were tested
$\sigma^2_{\overline{T_0}}$ = the variance of school true mean scores if one sample of size N were drawn for each school,

[2] Paul Holland has pointed out that Cochran found that reporting data in 5 or 6 categories removes most of the bias associated with categorization (W. G Cochran, *The effectiveness of adjustment by subclassification in removing bias in observational studies*, Biometrics, Vol. 24, Issue 2 (June, 1968), pp. 295-313). Cochran's finding is consistent with the recommendation provided in this paragraph.

$\sigma^2_{X|S}$ = the variance of pupil observed scores within school,
$\sigma^2_{T|S}$ = the variance of pupil true scores within school,
r_X = the reliability of pupil scores across all pupils,
$r_{\overline{X}_0}$ = the reliability of one observation of a school mean score, and
N = the number of students in each school.[3]

The following equations allow us to make some important calculations:

(1) $r_X = \sigma^2_T / \sigma^2_X$

(2) $\sigma^2_X = \sigma^2_T + \sigma^2_E$

(3) $\sigma^2_{\overline{X}_0} = \sigma^2_{\overline{T}_0} + \sigma^2_E / N$

(4) $\sigma^2_{\overline{T}_0} = \sigma^2_{\overline{T}} + \sigma^2_{T|S} / N$[4]

(5) $\sigma^2_T = \sigma^2_{\overline{T}} + \sigma^2_{T|S}$

(6) $r_{\overline{X}_0} = \sigma^2_{\overline{T}} / \sigma^2_{\overline{X}_0}$

(7) $\sigma^2_{X|S} = \sigma^2_{T|S} + \sigma^2_E$

(8) $\sigma^2_X = \sigma^2_{\overline{X}} + \sigma^2_{X|S}$

Now, suppose I observe within a state that for schools with 50 students per grade, $\sigma^2_X = 10000$, $\sigma^2_{\overline{X}_0} = 2500$, and $r_X = .90$. These numbers are chosen somewhat arbitrarily, but it is not unusual to find that the variance of school mean scores is one-fourth that of student scores, and that the student-level reliability for a state's test to be around .90. So while these results do not apply to any particular state, it is likely that the results for many states would be close to the ones we are about to calculate.

[3] Robert Brennan has pointed out that this model could be extended to include at least one additional source of error: occasion. The model used in this paper does not take into account the potential that tests used from year to year will not be highly parallel and well equated. If a state's tests across years were not highly parallel, the accuracy of the model would be improved by adding terms that take into account the variance due to the interaction of school and form.

[4] Brennan also pointed out that this equation is not quite correct. The equation is true only if there is an infinite number of schools. If the number of schools is equal to n, then the correct equation is $\sigma^2_{\overline{T}_0} = \sigma^2_{\overline{T}} + ((n-1)/n)\sigma^2_{T|S} / N$. This correction for a finite number of schools will have only a very small effect for most states, and is ignored for practical purposes in subsequent calculations. A proof of Brennan's correction is provided on the NCIEA website.

From Equation 1, σ^2_T = 9000; from Equation 2, σ^2_E = 1000; and therefore, from Equation 3, $\sigma^2_{\bar{T}_0}$ = 2500 – 1000 / 50, or 2480.

Now, from Equation 4, 2480 = $\sigma^2_{\bar{T}}$ + $\sigma^2_{T|S}$ /50, and from Equation 5, 9000 = $\sigma^2_{\bar{T}}$ + $\sigma^2_{T|S}$. Solving these two equations simultaneously tells us that $\sigma^2_{T|S}$ = 6653.0612 and $\sigma^2_{\bar{T}}$ = 2346.9388.

Given those true-score variances, $r_{\bar{X}_0}$ = 2346.9388 / 2500 = .939 (Equation 6); $\sigma^2_{X|S}$ = 6653.0612 + 1000 = 7653.0612 (Equation 7); and, as a final check, $\sigma^2_{\bar{X}}$ = 10000 – 7653.0612 = 2346.9388 (Equation 8), which equals the value for $\sigma^2_{\bar{T}}$ found above.

Now, suppose the state decided to replace its current test with another, less reliable test. Suppose the student-level reliability of the new test was only .80. All the true score variances would remain the same; all the observed variances would increase because of the reduced reliability of the test.

For this new test, σ^2_X would increase to 11250 (Equation 1), σ^2_E would increase to 2250 (Equation 2), and $\sigma^2_{\bar{X}_0}$ would increase to 2480 + 2250 / 50, or 2525. From Equation 6, we can calculate $r_{\bar{X}_0}$ = 2346.9388 / 2525, or .929. As expected, the reliability of school mean scores declines when the reliability of student scores declines, but the amount of decline in the former is quite small compared to the change in the latter. Thus, as noted above, higher reliability of student level scores will increase the reliability of school mean scores, but the effect is relatively small.

Now that we've taken an initial look at the impact of changing the reliability of the test, let's go back to where we were and change the number of students per school rather than the reliability of the test. Suppose now that we maintain the student-level reliability of the test at .90, but divide all the schools in the state randomly into halves, so that N = 25 for all schools. Since only N changes and not r_X, then σ^2_T, σ^2_E, σ^2_X, $\sigma^2_{\bar{T}}$, $\sigma^2_{\bar{X}}$, $\sigma^2_{X|S}$, and $\sigma^2_{T|S}$ will remain the same as they were in the first example.

Given those values, it is straightforward to calculate that $\sigma^2_{\bar{T}_0}$ increases from 2480 to 2613.0612 (using Equation 4), $\sigma^2_{\bar{X}_0}$ increases from 2500 to 2653.0612 (from Equation 3), and $r_{\bar{X}_0}$ drops to .885 (from Equation 6).

Table 1 provides results computed using the procedures outlined above. The conclusions are immediate: When thinking about the reliability of school means, size matters a lot, but student-level reliability matters only a little. The lowest reliability test for each size provides higher school-level reliability than the highest reliability test for the next smaller size of school. When the desire is to produce an accurate estimate of a school mean, increasing the reliability of the test cannot sufficiently compensate for small school size.

Table 1

Reliability of School Mean Scores for Different Combinations of *N* and r_X

N	r_X	$r_{\overline{X}_0}$
25	.60	0.823
	.70	0.848
	.80	0.868
	.90	0.885
50	.60	0.903
	.70	0.918
	.80	0.929
	.90	0.939
100	.60	0.949
	.70	0.957
	.80	0.963
	.90	0.968

The implication of this for the design of accountability systems is substantial. Suppose, for example, one had a limited budget for testing, and had to make the choice of giving a short test to several grades of students or giving one long test to students at just one grade. If the purpose of the program was to estimate the school mean, it is clear that giving the shorter test to more students would be the better choice. Similarly, if one had to choose between a test with high validity but equally high costs, so that its student-level reliability was low, as opposed to an inexpensive test with low validity but high reliability, it is clear that one should choose the test with higher validity and lower reliability.

C. The Reliability of One State's Accountability System

Just to get a sense of the magnitude of the probability of classification errors that might come about in an accountability system, this section will summarize the findings of a reliability study done for California. For readers interested in more detail, the full report is available at http://www.nciea.org/publications/ReliabliltyCA_API_Hill00.pdf.

California's accountability system involves the calculation of an Academic Performance Index (API) for each school, including students at all grades 2 through 11. The API is calculated by comparing each student's performance on the Stanford Achievement Test to national norms and translating that performance to an index on a scale from 200 to 1000. The school's API is a weighted average of all that information.

The API for each school is reported in several ways: (1) by school decile rank, (2) by similar school decile rank, (3) by gain relative to Growth Target, and (4) by gain of subgroups. Schools must test a minimum of 100 students to receive an API report. The results for each type of reporting are provided in the following sections.

Reporting by School Decile Rank. The APIs of all schools in California are ranked from high to low and divided into 10 groups with an equal number of schools in each group. When a school

receives its API, it is also told the decile rank within which its API falls. The first study that was done was to determine the probability that a school's reported decile rank was its true decile rank.

As would be expected, the probability that a school's reported decile rank was its true decile rank depended on the size of the school and where in the distribution it fell. Schools that were larger and in the tails of the distribution (where the distance between the deciles was larger) had a higher probability than smaller schools and those in the middle of the distribution. But even for the smallest schools (between 100 and 200 students tested) in the middle of the distribution, the probability was .98 that their reported decile rank was within one of their true decile rank.

Reporting by Similar Schools Decile Rank. Each school in California is assigned a School Characteristics Index (SCI) based on the characteristics of the enrolled students. Each school's API then is compared to the APIs of the 100 schools closest to it in SCI (50 above and 50 below). The APIs of those 100 schools are ranked and divided into 10 equal groups, which provides a "Similar Schools Decile Rank" for each school. Since the variance of the similar schools is less than the variance of all schools in the state, the probability of accurate classification is lower for the Similar Schools Decile Rank. Even so, for the smallest schools (between 100 and 200 students tested) in the middle of the distribution, 69 percent received a Similar Schools Decile Rank that was within one of their true Similar Schools Decile Rank.

Gains relative to Growth Target. Each school is assigned a Growth Target equal to the larger of (1) the distance between their API and 800, divided by 20, and (2) one point. Schools with a lower API have a larger Growth Target. Schools are classified by whether their gains in API from one year to the next equal their Growth Target. Several factors influence the probability that a school will meet its Growth Target, including (1) the size of the school, (2) the percentage of students returning from one year to the next (having a high percentage of students remaining in the school over the two measurements reduces the sampling error, which in turn increases the probability of correct classification), (3) the starting point of the school (schools with lower starting APIs have larger Growth Targets, which means that it is easier to detect whether their true growth equals their Growth Target), and (4) the amount of improvement the school actually makes (if the actual improvement of the school is far greater or smaller than its Growth Target, the probability that it will be accurately classified is higher than if its true growth is close to its Growth Target.

Schools that had a starting API between 400 and 599, for example, had Growth Targets of between 10 and 20 points, depending on their starting API. When such schools had 15 points of true growth, the probability that their API would increase by the amount of their Growth Target was fairly close to 50/50. What is interesting is that if they had between 100 and 200 students and made *no true growth at all*, they still had a one in four chance of having their API increase from one year to the next by at least the amount of the Growth Target. That is, if all the schools in that group did nothing to improve, approximately one-fourth of them would reach their Growth Target simply due to the luck of the draw of having a more able group of students the second year than the first. And conversely, schools of that size that made 30 points of true growth (on average, about double the amount of improvement expected for schools in that group) had a chance of almost one in four of having their observed change in API be less than

their Growth Target. So for schools of that size, making considerable improvement in their school's educational program was no guarantee that their observed results would result in the reward they had earned—on the basis of one year's results, at least.

For schools with higher starting APIs (and therefore smaller Growth Targets), the results were predictably worse. For schools with a starting API between 600 and 799 and 100-200 students tested, there was a .40 probability that they would meet their Growth Target even if they did nothing to improve their program; the odds were one in four even for schools with 800 to 1600 students. Thus, large numbers of schools that were reported as having met their Growth Targets the second year of the program easily could have done so without making any real improvement in their program, but just had, due to the luck of the draw, a more able group of students taking the test the second year than they did in the first.

Taking Gains of Subgroups into Account. The teachers in a school that met its Growth Target were eligible for substantial cash awards. In order to receive the award, however, each "numerically significant" subgroup in the school had to make improvement equal to at least 80 percent of the school's Growth Target. A subgroup was "numerically significant" if there were at least 100 students in the subgroup in the school who provide test scores, or if there were at least 30 such students and the subgroup comprised at least 15 percent of the students in the school who provided test scores. The subgroups included all the major minority racial and ethnic groups, as well as those participating in the free or reduced price lunch program. Naturally, having this many tests to run for all these subgroups reduced the probability that a school that actually had made substantial gain would receive a reward. Table 2 is a reproduction of selected sections of Table 25 from the original report.

To generate the data in Table 2, we assumed that a school's *true* improvement was equal to *twice* the amount required on average for schools in its group. That is, if the system were perfectly reliable, every school would have received a reward, since they all improved at least as much (and, on average, twice as much) as they needed in order to be eligible. However, the system is not perfectly reliable. Sometimes, a school's change in *observed* score would not be enough to warrant the reward, even though its change in *true* score met or exceeded the target. This, of course, would happen when a school would have a good class (or relatively good subgroup of students) the first year and then a poorer group of students the second year. Even though the school's educational program might have improved, the improvement would not be great enough to overcome the differences in the two groups of students.

Table 2

Average Probabilities for Schools Whose True Gain Is Approximately Twice Their Growth Target, Summarized from Tables 17-24

Starting API (and Amount of True Gain)	Number of Students Tested	Probability of Meeting Growth Target	Probability of Being Eligible for Reward
Less than 400 (50 points)	100-200	0.934	0.828
	200-400	0.980	0.907
	400-800	0.997	0.971
	800-1600	1.000	0.982
	1600 or more	1.000	0.996
400-599 (30 points)	100-200	0.764	0.540
	200-400	0.840	0.616
	400-800	0.901	0.705
	800-1600	0.971	0.754
	1600 or more	0.994	0.819
600-799 (10 points)	100-200	0.589	0.358
	200-400	0.622	0.335
	400-800	0.653	0.362
	800-1600	0.697	0.312
	1600 or more	0.736	0.243
800 or more (5 points)	100-200	0.596	0.446
	200-400	0.637	0.440
	400-800	0.666	0.492
	800-1600	0.707	0.421
	1600 or more	0.761	0.365

A couple of observations are immediately obvious. The more a school's true gain, the higher the probability that it will meet its Growth Target, and the larger the school, given a large amount of gain, the higher the probability that it will meet its Growth Target. But notice that for schools in the top two categories of starting API, *the larger the school the lower the probability that it will be eligible for a reward*. This seeming incongruous result occurs because larger schools have more numerically significant subgroups than do smaller schools, and therefore more tests to pass. And the probability of passing all those tests becomes low, even when (or perhaps more accurately, especially when) the school is large. The most glaring example is for schools with a starting API of 600 to 799 points. These schools needed to make between 1 and 10 points of growth, depending on their starting API. Their subgroups needed to make between 1 and 8 points of growth. If the school and every one of its subgroups had a true gain of 10 points in their API, the probability that the observed gain would be at least 80 percent of the school's Growth Target for every one of the subgroups ranged from just over one in three (for the smallest schools) down to less than one in four for the largest schools. The decreased reliability was not due to the amount of gain required. We reran the

study setting the required gain for subgroups at 60 percent and 100 percent of the school's Growth Target, and that change had only a minor change on the classification accuracy. The primary reason for the sharp decrement in classification accuracy was the increased number of subgroups that had to pass the test in the larger schools, and therefore, the increased likelihood that the observed result for at least one of those subgroups would not be as great as the Growth Target (even though the true scores for all subgroups had improved the same amount as the school as a whole).

D. Determining the Reliability of an Accountability System

Simply computing the reliability of a school mean score usually does not answer the question about the reliability of the accountability system, since the rules associated with accountability systems are usually far more complex than simply requiring a school mean to be at or above a certain level. For example, there may be an additional requirement that subgroups also score at or above a certain level, or there may be a process by which student scores are converted to an index before being aggregated. Thus, the procedures outlined in Section B are usually inadequate to answer the question of how likely it is the decision made about a school would change if another sample of students were drawn for the school. As a result, there are at least four distinct methods for computing the reliability of an accountability system: direct computation, split-half, random draws with replacement, and Monte Carlo. Each of those methods will be discussed in turn below.

Direct computation. "Direct computation" involves computing the errors around estimates and using areas under the normal curve to determine the probability of correct classification. This was the approach taken in the study outlined in Section C.

If the process for classifying schools is straightforward enough to estimate the standard errors directly, this is the most appropriate process to use. One significant hurdle to overcome in using this system is to estimate the variance of student scores within school. The observed variance for any particular school in a given year may not be a close estimate to what the observed variance would be for a larger sample. To overcome this problem, it makes sense to pool the observed variance within school for schools that likely have similar values for this statistic. The key to using this strategy effectively is determining which schools to pool. For the California study, we divided schools into many categories and looked at the average variance of students within school for all the schools in a particular category. When we found the values were similar across categories, we pooled those categories. We found that the average variance of students within school were substantially different for different grade levels and School API Decile Groups. Therefore, we calculated the variance of students within school for thirty different cells (three grade levels by ten School Decile Groups) and used the estimate for that cell for all schools within that cell. More specifically, we found that the variance of students within school was larger for schools with a high API, and the effect was larger for high schools than for middle schools and elementary schools.

The advantage of the direct computation method is its accuracy. It is the method of choice when the decision rules are simple enough that the error variances can be calculated.

Split-half. The split-half method involves dividing all the students in a school into two halves and applying the rules of the accountability system to each half. If the system is reliable, the same decision should be made on both halves of the students in the school.

While this method is simple to execute, there are a large number of places to go wrong. Obviously, a critical assumption behind the method is that the students are placed into the two groups randomly,

thereby simulating the process of creating new classes of students each year. Perhaps the simplest way to execute this method is to place all the students who have an odd number in the data file into one group and the even-numbered students in the other. Often, however, students are not randomly placed in the data files provided to researchers. For example, they may be sorted into classes, and some classes may be more able than others. If the file is sorted in some way, the system will over-estimate the reliability of the system. One state implemented this method using a file that was sorted by students' scores on the test—they came to the (mistaken) conclusion that their system was highly reliable.

One way around this problem is to use compare results of the direct computation method with the split-half method for some simple decision rules. If those results are close to each other, one can have more confidence in the randomness of the student assignments to group.

Another issue to be concerned about is that the "schools" in the split-half method are all of half-size. Therefore, one must stratify the results and then reweight the estimates created for each stratum to reflect the actual distribution of school sizes in the state. One problem with this method is that the estimates of classification accuracy may be weak for the largest schools.

Thus, the advantage of the split-half method is its simplicity. However, since one must be certain that students have been randomly assigned to groups for the results to be accurate, it is important to check the results carefully. Also, since the results are for half-sized schools, they must be interpreted with caution.

Random draws with replacement. Another possibility, used by David Rogosa, is to draw repeated samples for a school from the given sample. It would appear that a disadvantage of this method is that the variance of means for the replications would be larger than the original variance of school means—and it would be a greater issue for smaller schools than for larger schools. If that indeed is the case, then the estimates of classification consistency would be affected by that artifact. Other concerns about this method include the fact that the possible range of observations for a given school are limited to the values generated by the original distribution, and that the variance of students within each school in the one observed sample is taken to be the variance of students within the school for all draws for that school. However, in a personal correspondence, Haertel reported that the results using this method were very similar to those found using direct computation on the California data .

Monte Carlo. A fourth possibility is to carefully estimate all the parameters for a school and then, using a random number generator, make repeated draws of "students" for a school. Once that is done, one can apply the decision rules for the accountability system to each draw, and then determine the proportion of time the decision for the school was consistent with the original decision. This approach works particularly well when the decision rules are complex.

As with the direct computation method, making an accurate initial estimate of the variance of students within school is critical. Again, pooling the variance estimate across schools makes sense if it is reasonable (and data support the conclusion) that similar schools have similar variances of students within school.

Once that is done, all the variance components can be estimated using the equations in Section B. If the students are normally distributed within school, one can produce a set of scores that will mirror the original distribution through the following steps:

1. Starting with the original observed mean score for a school, estimate the true score mean for that school (using the notation in Section B, $\overline{T}$).
2. Knowing the standard deviation of student true scores within school and the n-count for that school, make a random draw of the true score for that school, , by selecting a random normal variable and computing $\overline{T_0} = \overline{T} + z * \sigma_{T|S} / N$.
3. Given the chosen true score for that school and the standard deviation of student observed scores within school, make N draws of a random normal variable and compute $X = \overline{T_0} + z * \sigma_{X|S}$. This provides you with a set of randomly generated observed scores for each school in your original distribution.

These generated observed scores will have the same student mean and standard deviation, as well as same school mean and standard deviation, as the original data set (within random error). You can check your work to this point by computing those statistics (and any other statistics that you deem appropriate) on both the original data set and the generated data set. They should provide highly similar results.

What you have done to this point, essentially, is create a data set of one year's results that is drawn from the same distribution as the original set. This can be done as many times as necessary to create a distribution of scores that will mirror the original data set. Given the power of one's computer and the amount of uncertainty one is willing to deal with, one might decide to generate a hundred, or even several hundred, random draws for each school.

Now, one can model the changes in scores over time. Suppose, for starters, that we wanted to know what the decision accuracy of our system would be if no schools made any improvement. In that case, we would draw data for as many years as necessary under the assumption that $\overline{T}$ remained constant over time. One then would apply the decision rules in the system to each replication of the data and simply count up the number of times a school was correctly and incorrectly classified. If we wanted to know what the decision accuracy of our system would be if all schools made some amount of improvement, we would specify that amount of improvement for each year, add that amount to $\overline{T}$ for each school from the first year, and then generate the distribution of observed student scores around that estimate. Again, once having done that, you would count the number of correct and incorrect decisions.

The above method works well so long as just one score is being generated for each student. Often, however, the system assumes that students take two or more tests, and then processes each test separately. The problem is that a student's performance on different tests is likely to be correlated. Thus, if, by the luck of the draw, a school gets a more able group of students in reading in any particular year, it is likely that those students will also score higher in mathematics (although, in general, not as much higher in mathematics as they are in reading). When each student takes two tests, the data for a school needs to be generated using the following additional steps:

1. Compute the correlation between the means of the school true scores on the two tests, $r_{\overline{R_T},\overline{M_T}}$.
2. Compute the correlation between the two tests on observed student scores within school, $r_{R,M}$.

3. Draw two random normal variates, z_1 and z_2, for each school. To compute a school's $\overline{T_0}$ for reading, do the same thing as given above; that is, $\overline{T_0} = \overline{T} + z_1 * \sigma_{T|S} / N$. But for the other content area, compute $z = z_1 * r_{\overline{R_T},\overline{M_T}} + z_2 * (1 - r_{\overline{R_T},\overline{M_T}})$. Use that value of z to compute the school's score in math.
4. Use a similar process to generate the observed student scores within each school. Draw two random normal variates, z_3 and z_4, for each student. Compute the observed score for the student in reading by the method outlined above. For math, however, substitute $z = z_3 * r_{R,M} + z_4 * (1 - r_{R,M})$ as the z-score.

This process will yield scores that are correlated to the appropriate degree. To demonstrate the process, I took the data from one grade for a state and followed these steps, generating five years' worth of data under the assumption of no improvement, with 100 replications of each school within each year. Table 3 provides some descriptive statistics from the original data set and the comparable statistics for the generated data.

Table 3

Descriptive Statistics from Original and Generated Data Sets

Level	Statistic	Original Data Set	Generated Data Set
Student	Reading Mean	251.8	251.8
	Reading SD	21.5	21.2
	Math Mean	254.4	254.5
	Math SD	18.8	18.7
	Correlation between reading and math	.73	.73
School	Reading Mean	250.8	250.8
	Reading SD	7.8	8.0
	Math Mean	253.6	253.7
	Math SD	7.4	7.3
	Correlation between reading and math	.80	.81

As can be seen, the means in the generated data set were almost exactly the same as those in the original. The standard deviations were close, but not as close as the means. The correlations between reading and math were very similar in both data sets at both the student and school levels.

As one final check on the data, it was observed that if scores in the state had not, in fact, changed very much from one year to the next, it might be inferred that little change had taken place in schools' instructional programs. If that were true, the correlations of random draws within year

should equal the correlation of school means across years[5]. The results for two recent years are shown in Table 4.

Table 4

Descriptive Statistics for Two Consecutive Years on Actual School Performance

Content Area	Statistic	Year	
		2000	2001
Reading	Mean	251.6	250.8
	Standard Deviation	7.8	7.8
Math	Mean	252.7	253.6
	Standard Deviation	7.3	7.4

Thus, performance in reading between 2000 and 2001 declined, while performance in math improved by a little over .1 school standard deviations. The data described earlier in this paper were generated from the 2001 scores, and were based on the assumption of no improvement across years. Given the actual results, where it appears that there was little, if any change in programs in reading (while there seems to be some substantial improvement in mathematics), it seemed reasonable that if we correlated the actual scores from schools for two consecutive years, we ought to get the same results when we correlated the scores from two years on the generated data in reading, and a slightly higher correlation in mathematics. That would be an acid test; if scores for schools really do resemble random draws from a larger population, the correlations across years should be identical when there is no improvement statewide.

Indeed, that proved to be the case. As shown in Table 5, the correlation between 2000 and 2001 for the actual school mean scores was .62 for reading. As shown in Table 6, the correlation between Year 1 and Year 2 of the scores generated from the 2001 data set under the assumption of no improvement was .62—exactly the same. For mathematics, the correlation between 2000 and 2001 on the actual data was .60; for the generated data, it was .64—a slightly higher value, as expected. Note also that the correlations across content areas across years also match up well. For example, the correlation between math in 2000 with reading in 2001 was .60, while the correlation between reading in 2000 with math in 2001 was .49. It isn't apparent why these two correlations should be any different from each other; the correlation from the generated data is directly in the middle of these two results: .54. These results give us confidence that the model followed—that scores from a school operate as if they are a random draw from the population in the school's catchment area—was a reasonable one.

[5] This is true only if one is looking at the scores of one grade. If school means were based on the averages of schools across adjacent grades, then the scores for one year would be dependent on those of the previous year; if a "good class" contributed to an atypically high score for a school in one year, it will also contribute the next year when that "good class" is promoted to the next grade, if the two adjacent grades both count in the school average. If that were the case, then that lack of independence would need to be taken into account when creating the samples from year to year.

Table 5

Correlations among Actual School Means for Two Consecutive Years

	Math, 2000	Reading, 2001	Math, 2001
Reading, 2000	.83	.62	.49
Math, 2000		.60	.60
Reading, 2001			.80

Table 6

Correlations among Two Draws of Generated School Means, with Assumption of No Improvement between Draws

	Math, Year 1	Reading, Year 2	Math, Year 2
Reading, Year 1	.81	.62	.54
Math, Year 1		.54	.64
Reading, Year 2			.80