Calculating And Reducing Errors Associated With The Evaluation Of Adequate Yearly Progress

Richard Hill
Advanced Systems

Paper Presented at the 1997 CCSSO Annual Assessment Conference
Colorado Springs
June 16, 1997

Introduction

In the Spring, 1996 issue of *CRESST Line*, Eva Baker and Bob Linn, pointed out that, in efforts to measure the progress of schools, "the fluctuations due to differences in the students themselves could swamp differences in instructional effects." The primary purposes of this paper are two-fold: Taking a typical system of measuring yearly progress, to illustrate, with specific examples, the truth of Baker and Linn's comment; and to describe alternatives that help distinguish schools that are truly improving from those that are not. Secondary purposes of this paper include showing how typical values of sources of error can be determined, displaying typical sources from certain states for which Advanced Systems is the assessment contractor, providing sufficient examples so that other states can estimate their own sources of error, and describing the amount of increase in error created by reporting data in terms of the percentage of students at various proficiency levels rather than a mean score.

One Proposed System of Evaluating Yearly Progress

There have been two recent trends in statewide assessments: reporting the performance of students in terms of performance levels, and evaluating schools on the basis of improvement, rather than their status within any particular year. Both of these trends are positive, since they provide results in a clearer and fairer fashion than was done previously. However, these improvements are not without their cost—and one of the costs is increased variances of error. In order to provide a framework for this paper, I describe a system for evaluation that is not atypical of the kind of models that states currently are using and proposing to use for Title I evaluation of adequate yearly progress.

The reporting strategy for many statewide assessment programs is to place students into one of three or four categories, for student-level reporting, and then to create an index from those categories to create a school average. Many systems look somewhat like the following: categorize its students into Novice, Partially Proficient, and Proficient categories, and assign a value of 0 to students at the Novice level, 50 for Partially Proficient, and 100 for Proficient. Scores thus would range from 0 to 100.

To implement a Title I evaluation system, states typically are taking a school's index from one year, adding a value to that, and using the sum to be a target of expectation for the school for the next year's testing. In part because there are no rewards in the Title I system, but simply negative consequences for those who fail to meet the criterion of success, states have selected small values for improvement—typically one-half or one percent. Thus, for example, a school that starts with an index of 40 the first year of the program might be expected to score at least 41 the second year, 42 the third year, and so on.

There are many variations on this theme being proposed by states. However, most of the differences being used (e.g., different numbers of categories into which students are placed and different score points assigned to each of the categories) have only a minor impact on the points being made in this paper, and the general discussion will be of value to states using such systems.

Concerns about the System

The primary point of this paper will be that statistical analysis shows that sampling error will make it difficult to determine which schools truly are making adequate yearly progress and which are not. However, before we get into that point, some discussion about the amount of expected improvement is warranted.

A school with an index of 40 could have 60 percent of its students scoring at the Novice level[1]. Suppose a state decided that schools should make progress of 1 point a year. Most advocates of educational reform believe that a reasonable goal for an effective school is to have most of its student scoring at the Proficient level; they might be hard-pressed to accept a long-term goal that schools should have no more than 20 percent of their students at the Novice level. This means then, however, that we are providing a school with *40 years* to transform itself from its current level to one of acceptable performance. While it is reasonable that states would not want to set unreasonably high expectations for growth, this seems to be an exceedingly modest expectation. Schools seeking significant improvements in their educational programs likely would have shorter timelines than several decades to see these kinds of increased test scores.

Computing Standard Errors

Whether there is agreement or not about the sufficiency of this target for improvement, it is a model being widely proposed by several states. Using this as an example, therefore, let's take a look at the sampling error that surrounds these estimates, and compute the probabilities of detecting such improvement.

.

Model 1: Testing One Grade Level Each Year

Model 1A: One year each in baseline and posttest. To start, let's take the simplest example: Comparing the index of a school in one year with its index the previous year. We will assume that the standard deviation of student indices is 30, and the standard deviation of school mean indices is 12. Of course, the size of the standard deviations will vary from state to state, depending on the variability of students and schools within the state, and the statistic chosen as the index. Also, the standard deviation of school means will vary depending on the size of the school—smaller schools will tend to have larger standard deviations. This issue is discussed later in this paper. But these numbers are not just drawn from thin air; they are not dissimilar from the data in many states.

[1] This is not to say that all schools with an index of 40 will have 60 percent of their students scoring at the Novice level. A school with an index of 40 might well have 40 percent of their students at the Novice level, 40 percent at Partially Proficient, and 20 percent at Proficient. However, one *possible* result would be for them to have 60 percent of their students at the Novice level, and this is chosen as the simplest example to follow.

To compute the variance of the error for the difference between the mean of a school in one year versus that same school's mean in another year, we will assume that the distribution of school mean scores in any one year is simply a function of the random draw of students, and that the school has not changed its "true score" (what its mean would be if the school could have tested an infinite number of students from its catchment area each year) from one year to the next. That is, we are supposing that, conceptually, there is a population of students from which a school could draw its students each year, and any particular class of students is simply a random sample from that population. In fact, observed data indicate that this is a viable model. Under such a scenario, the error variance—the variance of the difference of school mean scores—is computed as follows:

$$\sigma^2_{\overline{Year1} - \overline{Year2}} = \frac{\sigma^2_{STUDENTS,\,Year1}}{N} + \frac{\sigma^2_{STUDENTS,\,Year2}}{N}\,, \qquad (1)$$

where $\sigma^2_{\overline{Year1} - \overline{Year2}}$ is the variance of the difference between two mean scores,

$\sigma^2_{STUDENTS,\,Yearx}$ is the population variance of student scores *within* a school in Year *x,* and

N is the number of students in the school in any given year.

For our case, we will presume that the population variance of student scores within schools is the same in all years of the program, so we can simplify Equation 1 to be:

$$\sigma^2_{\overline{Year1} - \overline{Year2}} = \frac{2\sigma^2_{STUDENTS|SCHOOL}}{N} \qquad (2)$$

In this case, $\sigma^2_{STUDENTS|SCHOOL}$ is *not* 900. Nine hundred is the variance of students across all schools. If all schools are of equal size, the variance of students within schools is equal to the variance of students across schools, minus the variance of school means. That is:

$$\sigma^2_{STUDENTS|SCHOOL} = \sigma^2_{STUDENTS} - \sigma^2_{SCHOOLS} \qquad (3)$$

Thus, in this case the correct value for $\sigma^2_{STUDENTS|SCHOOL}$ is 30^2- 12^2, or 756. So, for example, if a school has an enrollment of 20 students per grade per year, then the variance of the difference of mean scores for that school is 75.6; the standard error of the mean difference is the square root of 75.6, or 8.7. This means that if a state had the above statistics for student- and school-level standard deviations, and all its schools had 20 students per grade, we could expect that 32 percent of school mean scores would change by more than 8.7 points per year, *presuming that no school changed its educational program from one year to the next*; stated in more statistical terms, under the null hypothesis (no change in the true mean of school scores), the standard deviation of difference scores would be 8.7 points. Using similar logic, the standard deviation of difference scores for a school of 50 students is 5.5; for a school of 80 students, 4.3. The variance of the errors and standard errors of difference scores for these three sizes of schools are displayed in Table 1, along with those values for other evaluation models that will be discussed later in this paper.

Now, let's take a look at what that means in practical terms. Let's start with the low end of improvement goals--half a point a year. Suppose no schools in the state *actually* improved—that all we observed in changes in scores from one year to the next were fluctuations due to different classes of students. For schools of size 20, the standard error is 8.7, so a change of 0.5 is a *z*-score of .06 (0.5/8.7). Tables of the normal curve tell us that 48 percent of observations will have a *z*-score of .06 or larger. That is, if all schools had 20 students per grade, and our criterion for reward was improvement of 0.5 points from one year to the next, 48 percent of all schools would meet that standard, *assuming that no school had changed at all*. Similarly, for schools with 50 students per grade, 46 percent would meet the criterion; for schools with 80 students per grade, 45 percent would meet the criterion. In contrast, we note this statistic: If all schools *had* improved 0.5 points, 50 percent would have met the criterion, and 50 percent would have failed, regardless of how much error there was in the system.

Said another way, suppose there were 100 schools in the state, and each had 20 students per grade. Suppose further that 50 of the schools had not changed at all, and the other 50 had *true* improvement that met the state's criterion of 0.5 points. At the end of the second year's testing, our most likely expectation would be that 49 of the schools had met the criterion, and 51 had not. But of the 49 who met the criterion, 24 would be schools that, in fact, had not improved at all. And of the 51 who failed to meet the criterion for improvement, 25 would be those who had, in fact, improved to the point of meeting the state's standard. That is, the probability of being identified as a successful school, given that a school had actually met the criterion for improvement, is only marginally higher than the probability of receiving such designation even if the school had made no change at all. When there is such a high probability of misclassification, the credibility of an assessment and accountability system will be brought into serious question. Of course, if most schools improve a great deal, then this issue is moot: most schools will meet the criterion for success, and they will be accurately classified. But if the criterion for improvement is small, and only a little more than half meet the criterion, there is great likelihood that many of the schools will have been misclassified if we use a model as simple as Model 1A.

If the above statistics sound surprising, it might be worthwhile to reflect for a moment on what an improvement of 0.5 means. Suppose a school has 50 students per grade. An improvement of 0.5 means that *one* student moves from Novice to Partially Proficient, *or* from Partially Proficient to Proficient, *once every two years*. Given the sampling fluctuations that occur on a yearly basis (the "good class, bad class" syndrome), it should not be surprising that it is virtually impossible to detect a change in the true performance of one student out of 100. Put another way, suppose I toss a coin 200 times and get 100 heads; then I take another coin, toss it 200 times as well, but now get 101 heads. Would you believe that the increased number of heads was the result of just the luck of the flips, or would you feel you had enough evidence to believe that the second coin was more biased in favor of heads?

Similar statistics have been computed for the criterion of improvement by one point, as well as alternative criteria of improvement by 5 and 10 points, and are displayed in Table 2. While the numbers for improvement by one point change somewhat because of the doubling of the criterion, the general summary remains the same. It is difficult to detect changes of this magnitude under such a simple evaluation model. Even if all schools had 80 students per grade and we used this larger criterion for improvement, 41 percent of the schools with no real improvement would be identified as "false positives;" that is, schools that were reported as improving, even when no improvement actually had taken place.

Now, however, suppose we changed the criterion for improvement to, say, 5 points. This would mean that a school would be given 6 years to effectuate a 30 point improvement in the achievement level of their students. This may or may not be considered a more reasonable timeline for improvement than the currently proposed model. However, there is no question that increasing the criterion sharply reduces the number of schools that would be identified as improving when there was, in fact, no real change. As can be seen from Table 2, the probability that an unchanged school with 50 students per grade will improve by 0.5 points, by chance alone, is .48; and the probability that it will improve by 1.0 points is .43. But the probability that it will improve by 5 points by chance alone is .18. Whether that is still too high a probability or not is a policy decision; what is clear is that the error rate for schools with no *true* change ("Type I" error) is reduced dramatically by establishing a criterion of 5 points. A criterion of 10 points reduces the probability of false positives even further; schools of 50 would have a probability of just .03 of increasing that much by chance alone—but that very well may be an unrealistic expectation for school improvement.

Model 1B: Two years each in baseline and posttest. The above example—determining improvement on the basis of comparing one year's result to the next—was the simplest case. Standard errors can be reduced, and the probability of rewarding those schools that have truly improved can be increased, by combining more years of data into the baseline score and more years into the posttest score. Thus, for example, we might improve the precision of the system by using two years' data for the baseline, and another two years' to determine improvement.

Under such a model, the variance of the error of the difference scores would be as follows:

$$\sigma^2_{\frac{\overline{Y1}+\overline{Y2}}{2}-\frac{\overline{Y3}+\overline{Y4}}{2}} = \frac{\sigma^2_{STUD,\,Yr1}}{4N} + \frac{\sigma^2_{STUD,\,Yr2}}{4N} + \frac{\sigma^2_{STUD,\,Yr3}}{4N} + \frac{\sigma^2_{STUD,\,Yr4}}{4N}, \qquad (4)$$

where $\sigma^2_{\frac{\overline{Y1}+\overline{Y2}}{2}-\frac{\overline{Y3}+\overline{Y4}}{2}}$ is the variance of the difference between two years' averages.

As was true for Equation 1, we can assume that the variance of student scores within schools is constant across all years of the program. Under that assumption,

$$\sigma^2_{\frac{\overline{Y1}+\overline{Y2}}{2}-\frac{\overline{Y3}+\overline{Y4}}{2}} = \frac{\sigma^2_{STUDENT|SCHOOL}}{N} \qquad (5)$$

Thus, when we double the number of years of information, we cut the variance of difference scores in half. This is an intuitively sensible finding. Essentially, the consequences of combining two years' worth of data are the same as doubling the number of students in a school. Therefore, for example, the results of an error analysis that we would do on a school of 100 students in one year would be the same as those for a school of 50 students, if we used two years' worth of data from the latter school (which would consist of 100 students).

As can be seen from Table 2, doubling the number of years of data has a small but consistent effect on the number of false positives that would occur. The improvement is not as much as one might expect because the change in the *standard* error (the statistic used to calculate the *z*-scores) is only the square root of the change in the *variance* of the error—so doubling the number of students (or years) only reduces the standard error by the square root of 2. Note that even with two years' worth of data in a school of 50 students, 40 percent of the schools that had no real change would be classified as improving if our criterion for improvement was 1 point. On the other hand, that percentage drops to 10 percent if the criterion is 5 points—and 10 percent might be a much more tolerable error rate than 40 percent.

Model 1C: Three years each in baseline and posttest. The results above can be extended by adding a third year to each of the baseline and posttest scores. For this example, the variance of the error is 2/3 of the variance of the error for Model 1B. The calculations for this model also are shown in Table 1 and Table 2. Once again, for schools with 50 students per grade, 38 percent of those with no real change would be erroneously classified as improving if our criterion for improvement was 1 point. This still seems to be a high error rate, especially when one considers it would take six years' worth of data to make the first judgments about schools.

Model 2: Testing More Than One Grade Level Each Year

No matter how the problem is approached, the solution lies in including more students in the assessments. The answer is *not* testing the same number of students in greater depth—it is including more students in the sample. It is sampling error, not measurement error, that is the primary source of uncertainty in this evaluation model.

Some have suggested using a longitudinal model to evaluate schools, rather than the cross-sectional model discussed here. That is, rather than testing the fifth graders two years in a row, for example, testing the students in the fourth grade and then again when they are in the fifth grade. While that approach *would* significantly reduce (but not eliminate) sampling error, it is not practical. How would one establish the standard for accomplishment in grade 5 on the basis of grade 4 testing? That is, suppose we set standards at the end of the fourth grade and determine that 40 percent of the students statewide are Proficient. The problem, then, is to set standards for the end of grade 5 that make sense. Should we arbitrarily say that if 40 percent of the fourth graders are proficient, then so are 40 percent of the fifth graders? What happens then if those standards don't appear to make sense? Or, on the other hand, should we proceed to set standards for the fifth grade independent of the standards set at grade 4? Suppose then we find that only 30 percent of the students are classified as Proficient. Should we then presume that many of the schools statewide did not make adequate progress with their students? No matter how one approaches this issue, the question of fair and equivalent standards will always be a problem. With a cross-sectional design, this is not an issue because, no matter how high or low we set the bar, we keep the bar at the same height for the next class of students—the work required to be Proficient at the fifth grade remains the same year after year. As a result, the equivalency of the standards from year to year is not an issue. We might debate how much *improvement* should be expected from schools, but there would be no argument about whether a school had held its own from one year to the next. Note also that a longitudinal design requires testing at two grade levels (potentially doubling assessment costs) to get evaluation information for one class of students.

Model 2A: Testing two successive grade levels each year. The above discussion, however, does *not* mean that a good design would not include students from more grades. Consider the following design. Test students at two grades (say, grades 4 and 5) each year, using standards that appear to be appropriate for each grade. Then, the next year, test students *at those same two grades.* The assessment information available to us then would be the following: One cohort of students would have been tested as fifth graders only in the first year of the cycle; a second cohort would have been tested at grade 4 in the first year of the cycle and at grade 5 in the second year; and a third cohort would have been tested in grade 4 in the second year only. Thus, we would have data on twice as many students as we would in the Model 1A, and we already have shown that testing more students reduces error variance. We also would have data on students in grade 4 and grade 5 in both years of the cycle, meaning that the comparability problems identified in the previous paragraph, which are the result of a longitudinal design, would be eliminated. Finally, we would have information on one common cohort in the design, which would reduce variance of the errors even more.

Let's take a look at this last point in more detail. The variance in school mean scores that we observe consists of both sampling and measurement error. That is, suppose we took the students in one grade in a school and divided them randomly into two classes. If we were to test those two classes, we likely would observe that those means are somewhat different from each other. One of the reasons for that is that the students are different; when classes are randomly assigned, each class is a random sample from the population of students. But even if we had tested the same students a second time on a parallel form of the test, we likely would have found some differences in observed means. These differences would be due to the measurement error in the test—the questions asked, the scoring of the questions, the different answers that students might give on different occasions, etc. While the relative magnitude of measurement error versus sampling error would vary depending on such things as the variability of students, the length of tests and the reliability of the scoring, the ratio of sampling error to measurement error will usually be quite high, especially when the results for several content areas will be combined into one overall index for a school, as is the case in this system. When Jonathan Dings computed this ratio for the Kentucky Instructional Results Information System, he found that the variance of the error due to sampling was about *four* times that of *all other sources of measurement error combined*. This likely would be a typical finding.

Suppose now that we were to test students within a school in two consecutive years, and we observed that the correlation between student scores in the two years' scores was .7. As with all the other assumptions made in this paper, it would not be difficult to actually conduct such testing and compute the correlation. Changes in the assumption being made would have some effect on the calculations in this paper, but unless these assumptions were way off, would not affect the conclusions. Under most conditions, it seems reasonable that one would expect to find a correlation of at least .7 across two consecutive years' of testing.

Whatever the correlation might be, we would find that it is not 1.0. That would be because there was two years' worth of measurement error in the results, as well as another source of error that has not been discussed up to this point: The changes in rankings that would naturally take place over two years as some students progressed faster than others and passed them in achievement between the end of one grade and the end of the next. That almost certainly would be a much smaller source of error than that due to random sampling, but a source of error nonetheless. To summarize, the error we are dealing with in this model is two years' worth of measurement error and one year's worth of error due to differential learning.

The variance of the error is computed as follows:

$$\sigma^2_{YEAR2-YEAR1} = \sigma^2_{YEAR1} + \sigma^2_{YEAR2} - 2r\sigma_{YEAR1}\sigma_{YEAR2} \tag{6}$$

Thus, if the variance of students within schools for each year is 756, and the correlation between scores of two consecutive years of students within schools is .7, then the variance of student-level difference scores within schools is 453.6, and the variance of school means, when students are held constant, is 453.6/*N*.

Therefore, if our model is the one described above—i.e., testing two consecutive grades for two consecutive years—the variance of the error of school means can be computed as follows, using the notation convention that Gi,Yj refers to the scores for students in Grade *i* in Year *j*:

$$\sigma^2_{\frac{\overline{G1,Y2}+\overline{G2,Y2}}{2}-\frac{\overline{G1,Y1}+\overline{G2,Y1}}{2}} = \sigma^2_{\frac{\overline{G1,Y2}-\overline{G2,Y1}}{2}+\frac{\overline{G2,Y2}-\overline{G1,Y1}}{2}} \tag{7}$$

$$= \frac{\sigma^2_{G1,Y2}}{4N} + \frac{\sigma^2_{G2,Y1}}{4N} + \frac{\sigma^2_{G2,Y2-G1,Y1}}{4N} \tag{8}$$

$$= \frac{\sigma^2_{STUD,Yr1}}{2N} + \frac{\sigma^2_{G2,Y2-G1,Y1}}{4N}, \tag{9}$$

where $\sigma^2_{STUD,Yr1}$ is the variance of students within schools for any year, and $\sigma^2_{G2,Y2-G1,Y1}$ is the variance of the error of the difference scores from Equation 6.

As can be seen from Table 1, there is dramatic improvement in the standard errors of this evaluation model versus the one in which independent samples are drawn. The standard errors that one derives from testing two successive grades each year are smaller than those for testing one grade for *three* years. The reason for this is clear from looking closely at Equation 9. The first term of that equation is one-half the value of Equation 5; that is, part of the error variance we are dealing with is the sampling and measurement error created by testing two different groups of students, each of which counts as half the total difference we are calculating. But the second term is smaller than the other half of Equation 5, because it involves only one group of students, not two, and because the error term we are using involves the *change* in students, not the *selection* of students.

Model 2B: Testing three successive grade levels each year. In a similar vein, we can compute the error variance if we tested three grades each year and compared performance of the second year to that of the first. Now, there would be two groups of students that would be in common from one year to the next rather than just one, so the error would be further reduced. The variance of the error can be computed from the following equation:

$$\sigma^2_{\frac{\overline{G1,Y2}+\overline{G2,Y2}+\overline{G3,Y2}}{3}-\frac{\overline{G1,Y1}+\overline{G2,Y1}+\overline{G3,Y1}}{3}} = \sigma^2_{\frac{\overline{G1,Y2}-\overline{G3,Y1}}{3}+\frac{\overline{G2,Y2}-\overline{G1,Y1}}{3}+\frac{\overline{G3,Y2}-\overline{G2,Y1}}{3}} \qquad (10)$$

$$= \frac{\sigma^2_{G1,Y2}}{9N} + \frac{\sigma^2_{G3,Y1}}{9N} + \frac{\sigma^2_{G2,Y2-G1,Y1}}{9N} + \frac{\sigma^2_{G3,Y2-G2,Y1}}{9N} \qquad (11)$$

$$= \frac{2\sigma^2_{STUD,Yr1}}{9N} + \frac{2\sigma^2_{G2,Y2-G1,Y1}}{9N} \qquad (12)$$

The variance of the error due to sampling in Equation 12 is still due to the drawing of the two "end" classes of students (Grade 1 in Year 2 and Grade 3 in Year 1) as it was in Equation 9, but now those two classes compose only one-third of the difference score, rather than one-half. The other two classes, for which there are data in both years 1 and 2, account for the other two-thirds of the difference score. Since those two classes have substantially smaller errors associated with them, the net result is another substantial decrease in error variance.

Exactly how much smaller the error is can be seen from Table 1. In Table 2, we see that the probability of misclassification is sharply reduced from what it was for the unmatched model, even when three years' of data were included in both the baseline and post-test. Note that the probability of misclassification for schools with no true change is quite small when the criterion for improvement is 5 points even for schools of modest size. Note also, however, that even with this model, the probability of misclassification remains substantial when the criterion for improvement is 0.5 or 1 point. Even with three classes of students included in each year's testing, changes in true scores of that magnitude are hard to accurately detect.

Table 1

Error Variance and Standard Errors for Three Selected School Sizes

	Error Variance			Standard Error		
	School Size			School Size		
Model	20	50	80	20	50	80
1A: One year each in baseline and posttest	75.6	30.2	18.9	8.7	5.5	4.3
1B: Two years each in baseline and posttest	37.8	15.1	9.5	6.1	3.9	3.1
1C: Three years each in baseline and posttest	25.2	10.1	6.3	5.0	3.2	2.5
2A: Testing two successive grade levels each year	24.6	9.8	6.1	5.0	3.1	2.5
2B: Testing three successive grade levels each year	13.4	5.4	3.4	3.7	2.3	1.8

Table 2

Probability that a School Will Have a Gain in Index of Varying Amounts, Given No True Change in the Achievement Level of Students in the Schools, for Three Selected School Sizes

	Gain = 0.5			Gain = 1			Gain = 5			Gain = 10		
	School Size			School Size			School Size			School Size		
Model	20	50	80	20	50	80	20	50	80	20	50	80
1A: One year each in baseline and posttest	48	46	45	46	43	41	28	18	12	13	3	1
1B: Two years each in baseline and posttest	47	45	44	44	40	37	21	10	5	5	1	*
1C: Three years each in baseline and posttest	46	44	42	42	38	34	16	6	2	2	*	*
2A: Testing two successive grade levels each year	46	44	42	42	37	34	16	5	2	2	*	*
2B: Testing three successive grade levels each year	45	42	39	39	33	29	9	1	*	*	*	*

*Less than 0.5 percent

Discussion of the Models

There is no question that if economics and politics will support the solution, the best way to minimize standard errors is to increase the number of adjacent grade levels being tested. Even if just two grades are tested each year (Model 2A), the reduction in standard errors matches the model in which one grade per year is tested, but three years' data are used in both the baseline and post-test (Model 1C).

There are many disadvantages to using three years' data in both baseline and post-test. First is the obvious need to wait six years before the first evaluation can be done. Another, and probably equally important, concern is the need to maintain a constant testing program over that period of time. As a new testing program is implemented, many lessons are learned that encourage one to modify the program. If one is using an evaluation model that requires comparable data over six years, there is great pressure to keep the testing program unchanged (or, more accurately, unimproved). There also is likely to be pressure to modify both content and performance standards as the testing program is implemented; with such long evaluation cycles, one either must keep the standards constant or report results differently for different reporting cycles. This can be very confusing to the public. Finally, if the assessment cycle is so long, it is likely that there will be changes in both the staff and students at the school. If new administrators and teachers arrive in the school in the middle of the evaluation cycle, they quite legitimately would be concerned about being held accountable for the work of their predecessors. Also, if the student population changed, it could very well be true that changes in test scores had nothing to do with changes in the educational program, but were simply a function of changes in the students being tested. Thus, it is clear that an evaluation system that keeps the time short between baseline and posttest has many advantages over a system that requires several years' worth of data.

It also is possible to design an assessment program so that testing at three grade levels is not three times the cost of testing at one grade level. One of the factors that makes assessment programs expensive is making the test reliable enough to be reported at the student level. One can design a program that provides student-level results at one grade level, but makes use of matrix-sampling at the other grade levels. Such a change would have only a nominal effect on the calculations in this paper, and could greatly reduce costs at those other grades. In a similar vein, most of the benefits of this model would accrue if one tested some content areas at one grade level and tested the remaining content areas at an adjacent grade level, since the correlation of students from one year to the next in two different content areas is usually almost as high as the correlation when the same content area is tested. The important concept in this model is removing some students as a source of error and substituting small errors in its place.

One relatively minor point to note is that Model 2 assumes that all the same students are tested from one year to the next in the grades where the same cohort is tested. That is an ideal that will not be realized in most schools. If more students transfer in and out of a district, the error variances for Model 2 will increase, with an upper limit being the error variances for Model 1. So, for example, if all the students in a school transferred in and out of the school from one year to the next, the error variance for Model 2A (testing two grades) would be the same as the error variance for Model 1B (testing two years), and the error variance for Model 2B (testing three grades) would be the same as the error variance for Model 1C (testing three years). That is, to the extent that the students are not the same from one year to the next, the improvement in precision that comes from matching students will be lost.

Additional Issues

There are four additional issues that are not directly related to this paper that need to be at least touched for completeness. Each issue affects the interpretation of the results provided earlier.

Cost of a Type I Error; Dichotomous Decision Rules

This paper addresses the probability of making Type I errors, but cannot evaluate the cost of such errors. That is, an important question to ask whenever looking at the probability of misclassification is, "What is the cost of making an error?" If the cost is small, one might be unconcerned, even when there is a high probability that an error is being made. In Kentucky, for example, there are substantial consequences for schools being placed in one category or another—cash rewards are provided to schools that are determined to be "successful," whereas serious negative consequences can accrue to a school that is determined to be "in decline." In another state, the primary concern might be simply not labeling a school as "declining" unless it truly was. In that case, the cost of a Type I or Type II error might be much smaller. Other things being equal, it always is desirable to minimize the standard error of school scores; but if the costs of misclassification are very small, it might cost more to minimize error than it would be worth to a state.

With this much said, however, it should be noted that the stakes for schools often are far higher than states will readily acknowledge. That is, while the state might not issue sanctions or rewards, the local public reading the results in the newspaper might very well be influenced by the results, and this can be a significant consequence for schools to bear.

It also needs to be recognized that the problem of categorization errors is greatest when one actually categorizes. That is, one problem with placing schools into a few categories is that there can be more similarily between schools in two different categories than there are between two schools in the same category. For example, we have shown that the probability of a Type I error can be reduced to a fairly low level, even in the simplest model, if one simply accepts a margin of five points of error around a school's score instead of one point. But that means that if we treat schools that gain five points differently from those who have not, we haven't solved the problem, only moved it to another point on the scale. That is, exactly the same probabilities of making classification errors exist in trying to separate out schools that have gained 4 points from those that have gained 5 as when one tried to separate out those that have gained 1 point from those that have gained none. The difficulty of detecting differences of one point are pretty much the same for most of the scale.

One way of minimizing this problem is to simply report the gain without categorizing it, or if one is forced to categorize, to make as many categories as possible. That is why Kentucky, for example, created many categories of reward schools. In that state, a school needs to make a specified amount of gain before it is eligible for rewards. But once eligible, the amount of the reward is determined by the amount of the gain. Not all schools receive the same amount of reward—more gain leads to more reward. As a result, a school that makes an extra point of gain, and therefore has a higher probability of having been more successful, get a larger reward than another school without that extra point, but the difference is small, commensurate with the probability.

Assessing in Multiple Cycles

This entire paper has been based on the presumption that the reliability and effectiveness of the evaluation system will be based on one cycle. That is, it is important to accurately assess the improvement that each school has made *each* time testing is done. In fact, that is the way that the politics of evaluation have worked out in many states employing such systems.

However, that doesn't necessarily make sense, and if the results are communicated with this thought in mind, it doesn't necessarily have to play out that way. Suppose we decide to reward schools on the basis of improvement by one point each year. As the main part of this paper has shown, that would mean that, regardless of the evaluation system chosen and the size of schools in the state, a large number of misclassifications will occur. We will reward many schools that have not improved, and fail to reward many that have in fact improved as much as the standard called for. Therefore, suppose we further prepared the public for this evaluation system by stating that we are not going to judge schools on the basis of one evaluation cycle, but that we are going to take a long-term view—that we are going to keep track of which schools continue to make progress over time and which ones don't. We would create a system that accumulates information over time, and evaluate schools on their success in meeting the criteria repeatedly, rather than reporting each year's results as if previous years' results did not exist.

Such a procedure would provide far more reliable and informative results. If such a system were in place for ten years, and we expected a point of improvement each year, the schools that would be at the top of such a system would be those that had improved by at least ten points—and we know from this paper that it is reasonable to expect to reliably detect changes of that magnitude. Of course, such a system would also have the same problems as one that used several years of both baseline and posttest scores: There inevitably would be changes in the assessment system, the school staff, and the students in the school that would make an historical picture misleading.

The Loss of Reliability Due to Recoding

In the past, it was typical for an assessment to use the mean as the primary descriptive statistic of a school's achievement. Today, however, it is becoming more common to report the percentage of students who have met certain performance standards. Often, this information is then recoded to create an index for a school. A good example of that recoding was used in this paper: Novices earned a score of 0, Partially Proficients a 50, and so on.

What has not been generally discussed is the impact that this recoding has on reliability. When continuous data (original test scores) are reduced to a limited number of categories, there is a loss of information. This necessarily leads to lower reliability, even if the recoded scores are used to create a new mean. To illustrate this fact, a small study was undertaken, modeling Kentucky's evaluation system.

In this study, simulated statewide data were generated twice by using the assumptions that the variance of school means was 16 percent of the variance of student means and that all schools in the state were of the same size. This simulated two years of data with each school maintaining its same true score in Year 2 as it was in Year 1. After generating these data using a continuous score scale, a second, recoded data set was created by recoding each score in the first data set, such that 40 percent of the students were Novice (recoded into a score of 0 on Kentucky's scale), 20 percent were Apprentice (a score of 40), 35 percent were Proficient (a score of 100), and 5 percent were Distinguished (a score of 140). The data for 1,000 schools were generated. Finally, two pairs of means were calculated for each simulated school—a pair of means of continuous scores and a pair of means of recoded scores. Each pair of school means was correlated, and the correlations compared. This provided a Monte Carlo estimate of the reliability of school scores calculated under each conditions. For schools of size 20, the reliability of the means from the continuous scores was .81; for the recoded means, .76. For schools of size 50, the comparable statistics were .90 and .89, and for schools of size 80, .96 and .94. Thus, there was some loss as expected. However, two questions remained: How accurate was this Monte Carlo study? (Even with a sample size of 1,000, there is some sampling error.) What was the best way to describe the amount of information loss?

Ed Haertel was of great help in answering both these questions. He produced a more mathematical treatment of the problem. Using the percentages of Novice, Apprentice, Proficient and Distinguished (NAPDs) provided above, the numerical values assigned to these respective levels, and the fact that school variance is 16 percent of the variance of student means, he computed the joint probability distribution of NAPDs that would occur from a bivariate normal distribution with a rho = .16. Then he computed *rho* after recoding to the NAPD scale. The value was reduced to .12776. He also showed that this value could be regarded as the reliability of a school score based on testing just one student. The reliability of school means based on more students can be derived from this "one-student" reliability by using the Spearman-Brown formula, exactly as if adding more student scores to each school mean was comparable to adding more items to a test; that is, the Spearman-Brown prophecy formula predicts the reliability due to increased sample size. He then showed that the Monte Carlo estimates we had computed were within reasonable sampling error of his predicted values, thereby lending credibility to both his and our calculations.

The bottom line to all these calculations was that he was able to show that the loss of information due to recoding, given the assumptions above, is equivalent to reducing each school's size by 23 percent. That is, if one computed the reliability of recoded means for a school of 50 students, one could achieve the same level of reliability from continuous data for a school of 38 students.

What the actual figures would be for any particular state would have to be determined by knowing the ratio of student variance to school variance, the percentages of students in each category, and the particular recoding system for the state. Note also that this is an *average* loss of precision. The impact of recoding for any particular school will be a function of the student score distribution in that school in relation to the cut points. Nonetheless, the example provided above using Kentucky's recoding system probably would not be atypical. Whether the loss of efficiency is worth the improvement in reporting clarity is a judgment for each state to make, and partially affected by the evaluation system and its stakes. But it is not an issue that should be ignored.

The Typical Range of the Ratio of Student Variance to School Mean Variance

Throughout this paper, we have used the assumption that the variance of student scores is 6.25 times that of school means (or, conversely, that the variance of school means is 16 percent that of student scores). That ratio was used as an example throughout this paper because it closely matches actual data in our some of the states for which Advanced Systems is the contractor. In fact, we looked at these data for several states: Kentucky, Maine, New Hampshire, Massachusetts, and Arkansas.

The ratio of 16 percent was almost exactly the figured observed in Kentucky and New Hampshire, and with some explanation to come later, in Maine. The ratio in Massachusetts, which has a much more diverse student population than these other three states, was considerably lower, ranging from .08 in reading to .13 in science. Interestingly, the ratio in Arkansas also was lower (.12 in literacy, .14 in mathematics). Perhaps the reason in this case is not that the student population is more diverse, but that the population tends to distribute itself more uniformly than it does in the Northern states, leading to smaller variability between schools—or perhaps it is due to an absence of significant numbers of enclaves with affluent, high-achieving schools. Thus, while it would be necessary to check the variability of students to schools before implementing the findings in this paper in any state, it would not be surprising to find those values to be close to the ones used in this paper.

The findings for Maine were especially interesting because, in contrast to the other states, Maine uses 10 questions to assess students in reading and mathematics, while in the other four content areas, students take just two questions. Therefore, the reliability of student level scores is considerably lower in the other four content areas, and one might expect that the ratio of the variance of school means to the variance of student means is lower in those content areas. That, in fact, is the case. For reading and mathematics, the ratios were .19 and .16 respectively, while for science, social studies, humanities and health, the ratios were .11, .11, .13, and .12.[2] Thus, states that use matrix-sampling to estimate school scores, and therefore use considerably shorter tests than other states, should expect this critical ratio to be smaller than it would be if they used longer tests.

[2] Ed Haertel has pointed out a method to show that these ratios appear to be plausible. Assume that the reliability of all the subject areas is the same, given equal numbers of items, and therefore, the differences among those six values are due to random error and the different numbers of items. Then, a two-item test would produce a ratio of .12 (the average of the four values, rounded up), and a 10-item test would produce a ratio of .18 (the average of the two values, rounded up). Each ratio is equal to the variance of schools, divided by the sum of the variance of schools, the variance of students within schools, and the variance due to items. If we arbitrarily say that the variance due to items in the two-item test is 1, then the variance due to items in the 10-item test is 1/5. Therefore, if we use the notation that *S* represents schools and *P* represents students, we can write that $\frac{\sigma^2{}_S}{\sigma^2{}_S + \sigma^2{}_{P|S} + \sigma^2{}_e} = .12$, and $\frac{\sigma^2{}_S}{\sigma^2{}_S + \sigma^2{}_{P|S} + \frac{\sigma^2{}_e}{5}} = .18$. If $\sigma^2{}_e$, the variance due to a pair of items, is 1, then $\sigma^2{}_{P|S}$ = 1.11, and the reliability of a 2-item test, which equals $\frac{\sigma^2{}_{P|S}}{\sigma^2{}_{P|S} + \sigma^2{}_e}$, is $\frac{1.11}{1.11+1}$, or .53, and the reliability of a 10-item test is .85. Both results seem plausible.