# Issues Related to the Reliability of School Accountability Scores

Richard Hill
The National Center for the Improvement of Educational Assessment, Inc.
August 10, 2001
Revised September 5, 2001

## Background

The National Center for the Improvement of Educational Assessment is a non-profit organization dedicated to the improvement of student achievement through improved practices in assessment and accountability. As part of the Center's mission, it annually hosts the Edward F. Reidy Interactive Lecture Series (RILS). In 2000, the RILS focused on schools accountability, and featured three major lectures on this topic: reliability, validity, and comparability of scores across years. This paper is a report on the first lecture—reliability. This lecture was intended for people familiar with many of the terms used in assessment and reliability, but not necessarily familiar with the issues of reliability when used in the context of accountability.

## Overview

This paper addresses three major points:

- Most states with accountability systems have not calculated the reliability of that system. While most have looked at the reliability of their assessments, few have made similar calculations on their accountability systems—and as will be shown in this paper, there can be wide differences between the two reliabilities.

- If the rules for an accountability system are simple, calculation of the reliability of the system is not difficult. This paper will demonstrate two methods for making those calculations—the "direct calculation" method and the "split half" method. Both methods give similar results in most cases; when the two values are different, it is important to understand the causes of those differences.

- The reliability of an accountability system is affected to different degrees by different factors. When designing an accountability system, it is important to understand which factors make a big difference and which ones don't—and to have some idea of what the reliability of a particular design will be—so that consequences of the accountability system can be proportional to the reliability of the system.

## Reliability, not Validity

Perhaps the first issue that needs to be made clear is that this paper is addressing issues of reliability, not validity. Validity addresses the concern of whether we are measuring what we intend to measure—or in the case of accountability systems, that the inferences people are drawing about the results of the system are consistent with actual results. The following example of an accountability design with low validity illustrates this point: Suppose an accountability system is supposed to identify the schools where teachers are doing the best job, but instead identifies the schools where scores are the highest. To the extent that high scores are not associated with good

teaching (but are, instead, highly correlated with other factors, such as high socio-economic status, regardless of the quality of teaching), such an accountability system will have low validity.

Reliability, on the other hand, addresses the extent to which the measurements we make are repeatable. A system has high reliability if a school that has been placed into a particular category one year would likely be placed in that same category in subsequent years if no change were made in the quality of teaching and learning in the school. It is possible for a system to be reliable without being valid, so it might seem logical to discuss validity before reliability. However, reliability is a *sine quo non*. An accountability system without reliability cannot tell us anything about the quality of teaching at a school, regardless of any other features the accountability system might have. Therefore, checking the reliability of an accountability system should be a necessary first step in its evaluation.

**How the Reliability of Student Scores Affects the Reliability of School Mean Scores**

One of the factors to consider in the design of an accountability system is that the reliability of a school score can be considerably higher than the reliability of the student-level measures. This is a critical factor that often is overlooked in the design of an accountability system. Exploring this issue will form a good background for the remainder of this paper.

First of all, it is worthwhile noting that an accountability system cannot be more valid than validity of its individual measures. If the tests are not measuring the intended outcomes, nothing in the further design of the accountability system can correct this flaw. Of course, it is possible to design an invalid accountability system even when the tests used in the system are highly valid, but it is not possible to have a valid accountability system without valid measures.

On the other hand, this section will show that it is possible to have a highly reliable accountability system even when the reliability of the individual measures is modest. This seeming contradiction occurs because the reliability of individual measures is mostly a function of the reliability of the test, while the reliability of school scores is mostly a function of the number of students tested. This difference between the two reliabilities is an issue that should be considered when designing an accountability system, as will be discussed below.

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

Let: $\sigma^2_X$ = the variance of pupil observed scores,
$\sigma^2_T$ = the variance of pupil true scores,
$\sigma^2_{\bar{X}}$ = the variance of school observed mean scores,
$\sigma^2_{\bar{T}}$ = the variance of school true mean scores,
$\sigma^2_{T|S}$ = the variance of pupil true scores within school,
$\sigma^2_E$ = the variance of error in pupil scores,
$r_X$ = the reliability of pupil scores across all pupils,
$r_{\bar{X}}$ = the reliability of school mean scores, and
N = the number of students in each school.

Of course, there are many sources of error in a school mean score. The sample of students chosen is a major one, as are the various elements that lead to variation in student scores. A

complete explication of error might separate these sources out, but it is unnecessary to do so in this case. We ultimately are interested in the total variation in observed scores that we might draw in a sample from a school. Some of that variation will be due to sampling, and some will be due to measurement error, given the students chosen. However, the variance of school observed mean scores can be calculated without separating the errors into these different components, since the distribution of students within a school already has these variance components built in. As a result, the following equations are all true:

$$\sigma^2_X = \sigma^2_{\bar{T}} + \sigma^2_{T|S} + \sigma^2_E,$$

$$\sigma^2_{\bar{X}} = \sigma^2_{\bar{T}} + \frac{\sigma^2_E}{N},$$

$$r_X = \frac{\sigma^2_{\bar{T}} + \sigma^2_{T|S}}{\sigma^2_{\bar{T}} + \sigma^2_{T|S} + \sigma^2_E}, \text{ and}$$

$$r_{\bar{X}} = \frac{\sigma^2_{\bar{T}}}{\sigma^2_{\bar{T}} + \frac{\sigma^2_E}{N}}$$

States generally know (or can readily compute from existing data) the first three items on the list—the variance of student scores, the variance of school means (for schools of a given size) and the correlation of student scores. Given that information, one can solve for all the unknowns in the above equations. For example, suppose I have observed that in my state, $\sigma_X = 100$, $\sigma_{\bar{X}} = 50$ when there are approximately 50 students per school, and $r_X = .90$. Note that this example isn't completely arbitrary—it is not unusual for states to find that the ratio of student variance to school mean variance is approximately 2:1, and for the student-level reliability of a state-developed test to be in the ballpark of .90.

Then, for my state it would be true that

$$10000 = \sigma^2_{\bar{T}} + \sigma^2_{T|S} + \sigma^2_E,$$

$$2500 = \sigma^2_{\bar{T}} + \frac{\sigma^2_E}{N}, \text{ and}$$

$$.90 = \frac{\sigma^2_{\bar{T}} + \sigma^2_{T|S}}{\sigma^2_{\bar{T}} + \sigma^2_{T|S} + \sigma^2_E}, \quad \text{or} \quad 9000 = \sigma^2_{\bar{T}} + \sigma^2_{T|S}.$$

One can solve the three equations above to get the following:

$$\sigma^2_E = 1000,$$
$$\sigma^2_{\bar{T}} = 2480, \text{ and}$$
$$\sigma^2_{T|S} = 6520$$

One of the givens above was that the reliability of the individual student measures was .90. The reliability of school mean scores is:

$r_{\bar{X}}$ = 2480/2500, or .992.

Now, let's take a look at how the reliability of school mean scores will change as the reliability of the test varies. We have already calculated $\sigma^2_{\bar{T}}$ and $\sigma^2_{T|S}$ —those values are fixed, regardless of reliability of my student-level measure. The variance of true scores will be the same no matter what the reliability of scores. The variance of *observed* scores will increase when reliability is lower, but the variance of true scores remains constant when reliability changes.

So suppose, for example, that the reliability of my student-level measure decreased from .90 to .80. If we accept $\sigma^2_{\bar{T}}$ and $\sigma^2_{T|S}$ as 2480 and 6520, respectively, then

$$.80 = \frac{9000}{9000 + \sigma^2_E}, \text{ or } \sigma^2_E = 2250.$$

If $\sigma^2_E$ = 2250, then

$\sigma^2_{\bar{X}}$ = 2480 + 2250 / 50, or 2525, and
$\sigma^2_X$ = 2480 + 6520 + 2250, or 11250.

When $r_X$ = .90, the ratio of pupil variance to school mean variance is 4:1. When the reliability of pupil level scores is reduced to .80, the ratio of pupil variance to school mean variance increases to 4.46:1, and the reliability of school mean scores drops from .992 to 2480 / 2525, or .982.

The following chart is calculated using the principles established above, and summarizes the relationship between the reliability of pupil scores, the ratio of pupil variances to school mean variances, and the reliability of school mean scores for various reliabilities of pupil scores.

**Table 1**

**The Impact of the Reliability of Pupil Score on the Reliability of School Mean Scores, Given the Example State's Variance of True Scores**

| Reliability of Pupil Scores | Variance of Observed Pupil Scores | Variance of Observed School Means | Ratio of Variance of Observed Pupil Scores to Variance of Observed School Means | Reliability of School Mean Scores |
|---|---|---|---|---|
| .90 | 10000 | 2500 | 4.00:1 | .992 |
| .80 | 11250 | 2525 | 4.46:1 | .982 |
| .70 | 12857 | 2557 | 5.03:1 | .970 |
| .60 | 15000 | 2600 | 5.77:1 | .954 |

Note that the values in Table 1 are not those for any state, but those for the example I have provided (although, as noted above, the values I chose are not atypical of the results many states report). So far, we know the following:

- The reliability of school mean scores decreases as the reliability of pupil scores decreases.
- The reliability of school mean scores is much higher than the reliability of pupil scores.
- The variance of observed pupil scores increases at a much faster rate than the variance of observed school means as reliability of pupil scores decreases, which means that the ratio of the two variances increases as reliability goes down. In fact, one useful measure of relative reliability is the ratio of these two variances. If true variances are fixed (as they would be if all the analyses were done within a given state), the ratio of the two observed variances would tell us the relative reliability of the tests and the systems. We will use this fact in the next section of this paper to judge various systems..
- The reliability of school mean scores remains fairly high even when the reliability of individual student scores becomes fairly low.

This last point is the most important one for this section of the paper. Note that even when the reliability of student scores has fallen so low that we wouldn't report them, the aggregation of the information to the school level (for schools with 50 or more students) provides a fairly accurate estimate.[1]

It is important to be aware that the reliability of school means can be considerably higher than the reliability of student scores because there usually is a trade-off between validity and reliability, time, and cost in the design of a test. Tests that are more valid often are much less reliable, and much more time-consuming and costly to administer and score. As a result, for student-level assessment, practical considerations require us to accept a test that is less valid than one we would ideally like to administer. However, as has been shown above, the trade-offs are dramatically different when the highest stakes for test use are going to be at the school level, rather than the student level.

This point is worthwhile noting specifically because lack of understanding of these principles has already resulted in changes for the worse in at least one state. When an outside panel evaluated Kentucky's assessment and accountability programs in 1995, the panel noted that the reliability of school mean scores was likely to be low (although they did not actually compute the reliabilities, they reasoned that they must have been low). They also noted that the student-level reliability of some of Kentucky's more innovative assessments was low. Because the panel assumed that the low reliability of school mean scores was due to the low reliability of the student-level scores, those innovative assessments were eliminated from Kentucky's assessment program. That was unfortunate, because, as we now know, the reliability of those assessments had a minimal negative effect on the reliability of the *accountability* system, but did have a considerable positive effect on its validity.

[1] These analyses assume that the number of items administered in each *school* remains constant. The reliability of student-level scores is strongly affected by the numbers of questions students take. If all students take the same questions (i.e., if there is just one form to the test), and the number of questions in the assessment is reduced, this will affect both the reliability of both school- and student-level scores. However, if the tests are matrix-sampled, so that the number of questions administered in each *school* is the same even though the number of questions taken by each *student* is reduced, then the conclusions drawn in this section are valid.

**Sampling Error vs. Measurement Error**

The test results for any given school will vary across measurements. That would be true even if we repeatedly measured a school within a single year; even with the students held constant, the measurement error associated with the observations for each student would result in changing scores for a school. However, measurements are typically taken across years, which means that not only has the occasion of testing changed, but so has the collection of students taking the test. Even if a school did nothing to change its educational program, its scores would vary from one year to the next. This variation would be due to both measurement error (the error associated with the observation of each student's test score) and sampling error (the group of students tested in any particular year).

An issue that is often misunderstood is that although we might test all the students in a grade each year, there still is sampling error associated with those scores. This year's class of third graders is a sample of the population of third graders that might attend the school from its catchment area. Experience with the variances of school means from year to year has shown that scores vary in a way that one would expect if a random sample of students were drawn repeatedly from the catchment area.

Thus, the measurement error associated with scores is a function of $\sigma^2_E$, while the sampling error is a function of the variance of students within school, or $\sigma^2_{T|S}$. One can compute the standard error of school mean scores with the knowledge of these two variance components.

Work through an example will serve to illustrate these calculations. We will use the same values for variance components that we used in the previous section. For this example, we will assume that the student-level reliability of our statewide test is .90, the variance of observed student scores is 10000, and the variance of school mean scores is 2500 when there are 50 students per school. With those given, we can compute that the variance of error in pupil scores is 1000, and that the variance of true scores of pupils within school is 6520. This would mean that the variance of observed scores of pupils within school is 7520 (the sum of the variance of true scores plus the error around those true scores). The standard error of the mean for schools then simply would be the square root of the variance error (7520) divided by the number of students. Thus, if a school had 20 students, the standard error of its mean would be 19.4 (the square root of {7520/20}).

Table 2 provides similar calculations for a range of test reliabilities and school sizes, using the assumptions that $\sigma^2_{\bar{X}} = 2500$ for schools with 50 students, and $\sigma^2_X = 10000$.

**Table 2**

**Standard Error of the Mean for Schools of Different Sizes and Tests of Different Reliability, Given the Example State's Variance of Observed Student and School Means**

| $r_X$ | $\sigma^2_E$ | $\sigma^2_{T\|S}$ | $\sigma^2_{X\|S}$ | $\sigma_{\bar{X}}$ if N = 20 | $\sigma_{\bar{X}}$ if N = 50 | $\sigma_{\bar{X}}$ if N = 100 |
|---|---|---|---|---|---|---|
| .90 | 1000 | 6520 | 7520 | 19.4 | 12.3 | 8.7 |
| .80 | 2480 | 6520 | 9000 | 21.2 | 13.4 | 9.5 |
| .70 | 3857 | 6520 | 10377 | 22.8 | 14.4 | 10.2 |
| .60 | 6000 | 6520 | 12520 | 25.0 | 15.8 | 11.1 |

Table 2 provides some very important information. While a lower level of test reliability increases the standard error of the mean of school scores for schools of a given size, the number of students tested is far more influential on that statistic. For this example, the standard error of the mean is larger when 50 students are given a test with a reliability of .90 than it is if 100 students take a test with a reliability of .60. Thus, for example, if one had a choice of testing two grade levels of students within a school on a very short, relatively inexpensive test or testing just one grade level of students on a much longer test, one would estimate the school true mean better with the shorter test over two grades than with the longer test at one grade. In short, the reliability of school accountability systems is maximized by getting a little bit of information about a large number of students than it is getting a lot of information about a few students.

**A Closer Look at the Reliability of School Classifications**

The data source. In the remainder of this paper, we will be looking at ways to compute the reliability of school classifications, the factors that influence that statistic, and some sample results. The source of the data used in these calculations was the assessment program in Louisiana for 1999. We are grateful to Louisiana for supplying us with these data and their permission to compute and publish the results. This leads, however, to two very important cautionary notes:

1. Our results are specific to these tests at this point in Louisiana's history. We describe the testing program in some detail below so that the reader might gain a sense of whether these data would generalize to another state. Louisiana's tests are typical of what many states are currently using, but each application of these data to another situation would have to be carefully considered. In short, it would be better to consider these analyses a model of those that a state might do on its own data, rather than accepting these results at face value and assuming they would be transfer to another situation.

2. These results in no way should be considered an evaluation of Louisiana's assessment and/or accountability system. In fact, we have made of point of using accountability examples that are NOT those being used in Louisiana.

Louisiana started its new assessment program in 1999. The data we are using are from its grade 4 Louisiana Educational Assessment Program for the 21st Century (LEAP 21) and its grade 5 administration of the Iowa Tests of Basic Skills (ITBS) in 2000. For most of the analyses, we used the English language arts portion of LEAP 21, which assesses reading, writing, proofreading and information resources. This portion of the test includes 33 multiple-choice and 10 two-point

constructed response questions, as well as an essay (worth 12 points). Thus, there are 65 possible points on the entire test. In addition to ELA, students also took tests in mathematics, science and social studies. Each of these tests was approximately the length (and reliability) of the ELA test.

To show the impact of test reliability (or lack thereof), we provide results for four possible tests:

1. An aggregate total over all four content areas,
2. The total ELA test,
3. The four separate parts of the ELA test in combination, and
4. The proofreading portion of the test only.

We chose the proofreading portion of the ELA test to provide an extreme example of what might happen with a very short test. The proofreading section consisted of just eight multiple-choice questions. There is no suggestion that it would be appropriate to design an accountability system around a test of that length—we simply wanted to show, for purposes of contrast, what the statistics were like for such a test when compared to much longer, more reliable tests.

Converting scores into indices. For purposes of creating student scores, we used Louisiana's system for classifying students into levels. However, Louisiana reports student results in terms of five levels; we combined the top two levels, giving us a 4-level scale which we then converted to an index of 1, 2, 3 and 4 to compute means. Therefore, once again, we did not use the system the way Louisiana does (and consequently the results cannot be applied back directly to Louisiana's accountability system)

In the analyses to follow, we consider three different types of reporting statistics—pass/fail, index, and mean scaled score--for each of the four tests. For pass/fail statistics, we considered scores of Basic or above to be passing (which, again, is inconsistent with the way Louisiana handles its data internally). About half the students (55 percent) were "passing" in ELA according to the definition used for this study. When we looked at students passing all four content areas of the LEAP 21, we used a conjunctive rule (students had to score "Basic" or above on all four content areas to be considered "passing" for this study—that was about 34 percent of the population). We defined "passing" on the Proofreading subtest to be 5 questions right out of 8, which meant almost half (47 percent) passed; for a conjunctive decision on the four parts, we determined raw scores that would have about half the students (43 percent) passing. It was important to look at the percentage passing for each of the tests, since it is known that the percentage passing will have an impact on the reliability of the statistic. Pass/fail statistics will classify schools most consistently when about half the students are passing. Thus, other things being equal, we would expect the pass/fail results for all four content areas combined to be somewhat less reliable than for the other three, since only a third of the students passed that standard, whereas the result was close to half for the other three.

For the analyses using 4-point index and mean standard score, we did not look at the combination of the four separate parts of the ELA test, since neither levels nor standard scores are reported for the subparts. However, we did create an index for the proofreading test, just so that we would have something to compare to the longer tests. We gave a raw score of 0-3 an index of "1," 4-5 a "2," 6 a "3," and 7 or 8 a "4." We also used raw score in place of standard score for that subtest.

Measures of Consistency. One could look at the reliability of school results in a large number of ways.. For example, one could compute reliability coefficients. While that statistic has

the advantage of being scale-independent, it is not readily interpretable. An alternative would be to report the standard error of means, or some other measure of the number of points of uncertainly associated with a school's observed result. That would be more interpretable, but would be scale-dependent. That is, if one reporting scale had a total of 50 possible points while another had a range of 2 (pass/fail), how would one meaningfully compare a result of 10 points of uncertainty on one scale vs. .5 points on the other? For these reasons, more and more people are using "the probability of correct classification" as a reporting statistic. It is scale-independent, and it is readily interpretable. However, it has its own limitations as a device to compare the reliability of different models, as we shall see shortly. Also, while such an analysis reflects the reality of the current situation statewide, it might not be an accurate reflection of what the statistics for the state would be once student performance begins to improve. For example, suppose a state uses a pass/fail statistic, and currently about half the students are passing the test. That would be the most reliable status for the state's accountability system. However, if after several years of improvement, 80 percent of the students were passing the state's test, the percentages of consistent classifications likely would change dramatically. School scores would be more closely bunched together, meaning that the likelihood a school's relative position would change upon retesting could increase dramatically.

An initial look. Let's start by taking a look at some results. Suppose we divided all the schools in the state into four groups according to their score on the test. It would be interesting to know what percentage of schools would score in that same quarter of the statewide distribution if another sample of students from that school (another year's class, say) took the same test.[2] There are at least two ways of doing this—"direct calculation" and "split-half." For the first method, one computes the standard error of the mean for schools and applies that uncertainty to the observed mean for the school. Using areas under the normal curve, one can compute the probability that a school's subsequent score might fall in any of several possible ranges. For the second method, one randomly divides the students in a school into two groups, applies the decision rules of the accountability system independently to each half, and determines whether the outcomes for both halves are the same. Since the students received the same instructional program and came from the same community, any differences between the halves can be assumed to be random error.[3]

To make these calculations, one first must calculate the variance of students within school. A separate estimate can be made for each school based on one year's results, but it probably is better to pool estimates across similar schools or to make one estimate for the state, since the observed variance for any particular school might be a poor estimate of what the true variance would be for that school. One way of obtaining better estimates of variances within schools is to stratify schools on some logical basis, compute the average variance within school for all the schools in each stratum,

[2] As was noted earlier, we are not considering changes in the test to be a source of error. If changes in the test were taken into account, it would lower all the probabilities in Table 3, but for tests of reasonable length drawn from the same content framework with attention to comparability across year, the decreases in those probabilities would be quite small.

[3] Since the presentation of this lecture in October, 2000, we have come to realize that there is a third method that works especially well with complex accountability systems. It is very similar to the "direct computation" method, but instead of calculating probabilities based on a normal curve, one simply writes a computer program to make random draws for each student from the infinite number of possibilities, given the school's mean and standard deviation of students within school. The accountability system is applied to that random draw and the school is classified. This is done many hundreds (or thousands) of times, allowing one to determine how often a school is consistently classified under the various draws. This "Monte Carlo" method provides very similar results to the direct computation method, but is simpler to employ when the decision rules for an accountability system are complex.

and see whether there are systematic differences between strata that should be taken into account. If there are no systematic differences, it is better to use a pooled estimate across all schools; if there are, it is better to use the result for each stratum for all the schools within it. More detail about this procedure is in the Center's publication of this research, entitled "The Reliability of California's API," available at http://www.nciea.org/publications.html.

Once the standard error of the mean for each school is known, one simply multiplies that statistic by the square root of 2, then computes the distribution of differences between the original observed mean for the school and all possible other observed means. (The standard error of the mean must be multiplied by the square root of 2 because both the original observed mean and all other draws have error variance associated with them, and the variance of the difference scores is the sum of the variances of both distributions).

The following is an example of the calculations used to determine the percentages reported for "pass/fail—Total ELA" in Table 3. The variance of students within schools was, on average across all schools in the state, 2041. There were no systematic differences between types of schools, so we used that estimate for all schools in the state. We computed the distribution of school means statewide, and found one-quarter were below 37.14, one-half were below 55.88, and one-quarter were above 70.10. One school had a mean of 74.42 and an N of 43 (that is, 32 of their 43 students {or 74.42 percent} scored Basic or above on the ELA test). The standard error for this school was 9.74 (the square root of 2 * 2041 / 43). The school's observed score is in the fourth quarter (i.e., was above 70.10). The probability that another observed score for this school would also be in the fourth quarter is the probability of drawing a z-score of -.44 or higher ({70.10 – 74.42} / 9.74 = -.44). That probability is .67. Using similar logic, the probability of drawing a random sample that remained in the top quarter was computed for all schools that started in the fourth quarter, and then averaged. That final result provides the probability that a randomly drawn school that has an observed score in the fourth quarter will get a second observed score in that same quarter, which is the probability of a consistent classification.

Table 3 provides the results for our sample data set. There are several results of note in the table. First of all, it makes sense that the highest reliability would be for all four content areas, next highest for the total ELA score, and lowest for the score on proofreading only. Similarly, it is logical that the most reliable results would be for standard scores, then the index, and then pass/fail, since reducing a scale to fewer points reduces its reliability. It was noted earlier that the ratio of the variance of students to the variance of schools is a good measure of relative reliability when other factors are held constant, as they are in this case. The ratios are smallest for all four content areas, then total ELA, then proofreading only, when one holds the reporting statistic constant. Similarly, if one holds the test constant, the ratio is smallest for standard score, then four-point index, then pass/fail. Note also that the difference in the ratios between four content areas vs. total ELA is less than the gap between total ELA and proofreading only. That is consistent with expectations, since once there is a fairly reliable test in place (the total ELA), providing additional reliability in the test doesn't yield as much improvement. But perhaps the biggest surprise, and most important result, for the chart is this: The ratio for proofreading only using a standard score is not much higher than for all four content areas when using pass/fail as the reporting statistic. That is, the amount of information lost by moving from standard scores to pass/fail is almost equivalent to changing one's accountability system from an entire battery to an 8-item test. That result alone should give pause to anyone thinking of designing a statewide accountability using pass/fail data.

Note, however, that the percentages of agreements within each quarter do not reflect many of these concerns. Look, for example, at the results for pass/fail on all four content areas vs. the mean standard score for those same tests. Although the ratio of student variance to school variance is far higher for pass/fail (5.26 vs. 3.07), the percentage of agreements is only slightly higher for the mean standard score (76 vs. 71). And indeed, the percentage of consistent classifications is even higher for pass/fail for the fourth quarter. How could this be? Remember that the percentage of students passing all four tests was considerably lower than the percentage passing any one particular test (about one-third compared to one-half). This means that school means bunch up in the lower end of the distribution for pass/fail more than they do for the mean standard score. In turn, this means that the range of scores in the first and second quarters is smaller for pass/fail than for the mean standard score (but then also, almost by definition, is larger for the third and fourth quarters). Thus, there is a wider range of scores in the first and second quarter for the mean standard score, making it easier for a school to remain in the same quarter upon a second sample; the opposite happens in the upper quarters.

**Table 3**

**Percentage of Consistent Classifications, Using Q1, Q2, and Q3 as Cuts**

| Reporting Statistic | Tests | Ratio of Variance of Students to Variance of Schools | Percentage of Consistent Classifications, by Quarter | | | | |
|---|---|---|---|---|---|---|---|
| | | | First | Second | Third | Fourth | Overall |
| Pass/Fail | Four content areas | 5.26 | 72 | 64 | 60 | 86 | 71 |
| | Total ELA | 5.68 | 83 | 62 | 52 | 78 | 69 |
| | Four ELA subparts | 6.58 | 83 | 62 | 59 | 82 | 72 |
| | Proofreading only | 8.54 | 79 | 50 | 50 | 78 | 64 |
| Four Point Index | Four content areas | 3.30 | 87 | 68 | 63 | 85 | 76 |
| | Total ELA | 4.47 | 86 | 59 | 58 | 80 | 71 |
| | Proofreading only | 6.60 | 82 | 56 | 52 | 80 | 68 |
| Mean Standard Score | Four content areas | 3.07 | 90 | 71 | 60 | 83 | 76 |
| | Total ELA | 4.00 | 87 | 59 | 51 | 81 | 70 |
| | Proofreading only | 5.83 | 84 | 59 | 49 | 77 | 67 |

This result made it clear that one could not trust the percentage of agreements by quarter to be an accurate reflection of the relative reliability of a reporting statistic. There are too many idiosyncrasies that can arise from the placement of the quarters. Also, when reporting statistics are unreliable, the variance of observed scores increases. As a result, school means are spread further out, providing a wider range into which a school can fall to be consistently classified. This gives unreliable statistics an advantage that can mislead the unwary or casual reader of the results.

Note also in Table 3 that the probability of consistent classification is lower for schools in the middle of the distribution than it is for those in either of the two extreme quarters. That occurs because the range in the extreme quarters is so much larger than it is for the middle quarters. Also, a school in the first quarter, for example, cannot be classified any lower, so at least half the possible outcomes (those lower than the observed outcome) are always automatically consistent classifications for those schools, regardless of how close the school was to the cut-score or how unreliable the statistic might be. Therefore, the first question we tackled was whether there might be a way of dividing schools into groups that might overcome the problems we noted with the quarters. Our first attempt was to calculate the percentage of consistent classifications by decile, rather than quartile. We computed the nine deciles (D1-D9), then the probability that a school would classified within one decile of its observed score.

Table 4 provides those results for the two extreme groups as well as one in the middle. The same problems observed with the quartiles were as bad or worse here: the differences between the percentage of consistent classifications were hardly any larger for the more reliable statistics than for the less reliable ones (in fact, there was a reversal—there was greater consistency for the four-point index than there was for the mean standard score); and the percentage of consistent agreements was far greater in the tails than in the middle of the distribution. Thus, it was clear that simply dividing the distribution into more points was not the solution to the problem.

**Table 4**

**Percentage of Consistent Classifications, Using D1-D9 as Cuts**

| Reporting Statistic | Tests | Percentage of Consistent Classifications within One Decile | | | | | |
|---|---|---|---|---|---|---|---|
| | | First | Second | Fifth | Ninth | Tenth | Overall |
| Pass/Fail | Four content areas | 87 | 86 | 71 | 94 | 97 | 82 |
| | Total ELA | 97 | 93 | 70 | 84 | 93 | 81 |
| | Four ELA subparts | 96 | 93 | 72 | 90 | 96 | 83 |
| | Proofreading only | 93 | 87 | 60 | 85 | 91 | 74 |
| Four Point Index | Four content areas | 99 | 97 | 83 | 91 | 97 | 89 |
| | Total ELA | 99 | 94 | 74 | 85 | 96 | 83 |
| | Proofreading only | 96 | 92 | 65 | 86 | 93 | 78 |
| Mean Standard Score | Four content areas | 100 | 98 | 78 | 90 | 96 | 86 |
| | Total ELA | 99 | 94 | 70 | 85 | 96 | 82 |
| | Proofreading only | 97 | 92 | 69 | 83 | 92 | 78 |

Table 5 provides an overview of four different possible methods for showing consistency of decisions. In addition to the two already considered in Tables 3 and 4, we looked at the possibility of dividing the distribution into four groups, but with cuts at the $10^{th}$, $50^{th}$, and $90^{th}$ percentiles instead of

the 25th, 50th and 75th, as we did for Table 3. That didn't work; on the very first analysis (pass/fail), we had a reversal: the ELA test alone had higher consistency than the four content areas together. Finally, we looked at the percentage of times a school's second observed score would be within one-half a standard deviation of the first score. That criterion listed all the cases in the correct order; the longer the test, and the finer the reporting statistic, the greater the percentage of consistent classifications. So while we wouldn't recommend this criterion as a good one to use to communicate to the public (because it is not as simple to understand), we decided this would be the best to use to compare the various issues we were about to investigate.

**Table 5**

**Comparison of Four Different Methods of Computing Consistency On State-Developed Test**

| Reporting Statistic | Tests | Q1, Q2, Q3 | P10, P50, P90 | Deciles (within 1) | ½ Standard Deviation |
|---|---|---|---|---|---|
| Pass/Fail | Four content areas | 71 | 72 | 82 | 80 |
| | Total ELA | 69 | 73 | 81 | 78 |
| | Four ELA subparts | 72 | 76 | 83 | 75 |
| | Proofreading only | 64 | 69 | 74 | 68 |
| Four Point Index | Four content areas | 76 | | 89 | 90 |
| | Total ELA | 71 | | 83 | 84 |
| | Proofreading only | 68 | | 78 | 75 |
| Mean Standard Score | Four content areas | 76 | | 86 | 92 |
| | Total ELA | 70 | | 82 | 89 |
| | Proofreading only | 67 | | 78 | 78 |

The split-half method. All the above calculations were done using "direct computation." One of the areas to study was whether the split-half method would give similar results to direct computation. These are two different ways of estimating the reliability of a school's score. If they gave dramatically different results, one would question the accuracy of both.

For the split-half method, we divided all the students in a school into two groups by assigning a record number to all the students in the file (that is, just numbering them all from 1 to about 55,000), and then placing all the students with an odd number into Group A and all the students with an even number into Group B. One then simply sees whether the means for the two groups fall within the acceptable range of each other to be considered consistent.

While this method has the advantage of great simplicity (presuming one has a computer on which to conduct the analyses), it has at least three disadvantages when compared to direct computation:

1. The results one gets are for half-sized schools. Obviously, larger schools have more stable means than smaller schools, so one must classify these schools on the size of *each half*, not on the size of the original school.

2. This means, then, that there will be fewer data points to estimate the scores for larger schools. If the largest school in the state had 100 students, for example, the split-half method would only allow for estimation of reliability for schools of size 50.

3. One must be very careful to divide the students into random halves. Any stratification in the data file will lead to overestimates of the reliability of the system.

Table 6 provides the results for both methods for all the combinations of tests and reporting statistics that we have looked at to date, but also by size.

**Table 6**

**Percentage of Times Two Halves Are within One-Half of a Weighted School Mean Standard Deviation of Each Other**

| Tests | School Size | Percent Passing | | Index | | Mean | | N | |
|---|---|---|---|---|---|---|---|---|---|
| | | D.C. | Split ½ | D.C. | Split ½ | D.C. | Split ½ | D.C. | Split ½ |
| Four content areas | 11-20 | 50 | 63 | 64 | 66 | 67 | 75 | 40 | 158 |
| | 21-40 | 65 | 67 | 80 | 85 | 82 | 91 | 158 | 330 |
| | 41-80 | 81 | 81 | 92 | 93 | 94 | 95 | 329 | 251 |
| | 81 or more | 92 | 95 | 98 | 100 | 99 | 100 | 277 | 21 |
| | All schools | 80 | 82 | 90 | 92 | 92 | 95 | 804 | 760 |
| Total ELA | 11-20 | 48 | 53 | 54 | 54 | 58 | 64 | | |
| | 21-40 | 63 | 62 | 70 | 74 | 74 | 85 | | |
| | 41-80 | 79 | 82 | 85 | 91 | 88 | 92 | | |
| | 81 or more | 91 | 100 | 95 | 100 | 96 | 100 | | |
| | All schools | 78 | 83 | 84 | 89 | 87 | 92 | | |
| Four ELA subparts | 11-20 | 44 | 54 | | | | | | |
| | 21-40 | 59 | 65 | | | | | | |
| | 41-80 | 75 | 75 | | | | | | |
| | 81 or more | 88 | 90 | | | | | | |
| | All schools | 75 | 77 | | | | | | |
| Proofreading only | 11-20 | 39 | 40 | 44 | 38 | 47 | 41 | | |
| | 21-40 | 52 | 54 | 59 | 62 | 62 | 66 | | |
| | 41-80 | 67 | 66 | 75 | 76 | 78 | 82 | | |
| | 81 or more | 82 | 81 | 88 | 95 | 90 | 95 | | |
| | All schools | 68 | 68 | 75 | 78 | 78 | 81 | | |

Note that, in order to make the results as comparable as possible, the results for "All schools" for the split half method were calculated by weighting each of the cells for school size by the original size of the school, not the split-half size of the schools.

There are several observations worth noting in Table 6. First, the split-half method fairly consistently gave higher estimates of the reliability of the school scores than did direct computation. Upon reflection, we felt there were two possible reasons for this. First, it is likely that there was

some unanticipated stratification in the data file. It is likely that the order of the file within school was the order in which the contractor received students' tests. If, for example, all the better students sat in the front of the room and all the poorer students sat in the back, it might be true that there was some stratification in the file because of that. Another possible reason might be that classes within school might have generally better or weaker students, and the file was organized by class. In short, if we were to do this study again, it would be better to randomly sort the students before placing them into the two groups. Systematic assignment appeared to have some unintended stratification to it. Also note that the schools placed in each size category are not the same for the two methods. The 40 schools of size 11-20 in the direct calculation method truly had that number of students tested. However, the 158 schools listed in that size category for the split-half method actually had 21-40 (note that that is the number listed for that size for the direct computation method). If there was something systematically different about those larger schools (for example, suppose they actually had a larger variance of students within school than the smaller schools), the finding for "schools of 11-20 students" actually wouldn't be that. It would be a finding for schools of 21-40 students, divided in half—which wouldn't necessarily be the same thing. However, that is likely to be a trivial issue here.

The good news is that both methods provided conclusions that are highly similar. Results are more reliable when the reporting statistic is mean scores than it is when the reporting statistic is the index, which in turn is more reliable than pass/fail (percentage passing). Longer tests are more reliable than shorter tests.

Impact of size and reporting statistic on reliability. By breaking the results out by school size, it becomes clear what a dramatic effect the number of students tested has on the reliability of school scores. In all cases, the percentage of times a school will be consistently classified increases dramatically when the size of the school increases. With even the least reliable test—the eight-question proofreading test—schools with 81 or more students had two observed means within one-half of a standard deviation of each other 90 percent of the time. Schools with 21-40 students had two observed means that close only 82 percent of the time, even when the test length was increased to an entire battery.

Another piece of valuable information comes from this table by comparing widely across cells. Note the impact of the choice of reporting statistic—the reliability of mean proofreading scores is as high as pass/fail total ELA test scores. That is, the loss of information by reducing one's reporting statistic to pass/fail is equivalent to the loss that would come from reducing a full length test to one of eight questions. That is something for people to seriously consider in the design of accountability systems.

Another way of looking at this same issue is to compare the consistency statistics for "Four content areas." The results for reporting the mean on schools of 21-40 students are more consistent than those for pass/fail on schools of 41-80 students. That pattern holds up across all sizes of schools. Thus, changing from means to pass/fail is equivalent to giving up data on half the students in a school when tests are quite reliable. The impact is reduced when tests are less reliable, but there still is a considerable loss in reliability with the use of that statistic.

Impact of combining years. As is clear from the results above, the number of students tested is a major determinant in the stability of a school's score. There are two ways that a system could increase the number of students tested in a school: aggregate results across years, or test students at

more than one grade. Table 7 looks at the results for the first option. In this table, "School Size" is the total number of students tested, whether that is for one year or two.

**Table 7**

**Percentage of Times Two Observed Results Are within One-Half of a Weighted School Mean Standard Deviation of Each Other**

**Comparing One Year to Two Years**

**Direct Computation Method**

| Tests | School Size | Percent Passing | | Index | | Mean | | N | |
|---|---|---|---|---|---|---|---|---|---|
| | | One Year | Two Years | One Year | Two Years | One Year | Two Years | One Year | Two Years |
| Total ELA | 11-20 | 48 | 45 | 54 | 52 | 58 | 55 | 40 | 17 |
| | 21-40 | 63 | 65 | 70 | 72 | 74 | 76 | 158 | 38 |
| | 41-80 | 79 | 79 | 85 | 86 | 88 | 88 | 329 | 150 |
| | 81 or more | 91 | 95 | 95 | 97 | 96 | 98 | 277 | 622 |
| | All schools | 78 | 90 | 84 | 93 | 87 | 94 | 804 | 827 |

It is clear from the table that the result for two years' data is just about the same as doubling the number of students within a year. When stratified by "school size," the results for one year and two years are almost identical. The exception is schools of "81 or more," because even though that category is the same, there were many more schools of much larger size in that category with two years' worth of data than one.

Note that because of the doubling of the number of students tested, there is a sharp increase in the number of "large" schools. As a result, the percentage of times a school's result will be within a half a standard deviation upon resampling is dramatically increased when aggregated over all schools. For example, with one year's worth of data, 87 percent of the schools had observed means that close; but the percentage rose to 94 when two years' data were used. Note again the tremendous impact school size has on the reliability of scores. Even though "percent passing" is a much less reliable reporting statistic than the mean, as we have noted repeatedly in previous discussion, more schools will be accurately classified in a system that uses pass/fail and two years' worth of data than a system that uses means but just one year.

Table 8 provides another look at the impact of combining data for two years. One of the problems of using the split-half method is that the N for each school is half the number is should be. If two years were combined before splitting in half, however, that might provide the same results as one year using the direct computation method.

That, in fact, turned out to be the case for the Total ELA test. Table 8 reports results for the direct computation and split half methods. These results are identical to those reported in Table 6. When we combined the data for 1999 and 2000, and then split the schools in half, we came up with average results across all schools for the split-half method as we had for direct computation.

We also ran one other analysis, just to demonstrate that it is an incorrect approach to exploring the issue of reliability. We compared the results of 1999 to those of 2000—not by splitting the results of each year in half, but by comparing a school's score in 1999 to that of 2000. Note that the percentage of times that a school's 2000 score fell within one-half of a standard deviation of its 1999 score was dramatically lower for schools of every size. This is because we are no longer looking just at random error, but also systematic effects, such as changes in teachers and teaching. Thus, it is clear from Table 8 that one cannot estimate the reliability of a school's score by comparing one year's results to those of another. One will consistently (and often substantially) underestimate the reliability of the accountability system by such an analysis.

**Table 8**

**Percentage of Times Two Halves Are within One-Half of a Weighted School Mean Standard Deviation of Each Other on Total ELA Score**

| Group and Method of Computation | School Size | Percent Passing | Index | Mean | N |
|---|---|---|---|---|---|
| 2000 only, direct computation | 11-20 | 48 | 54 | 58 | 40 |
| | 21-40 | 63 | 70 | 74 | 158 |
| | 41-80 | 79 | 85 | 88 | 329 |
| | 81 or more | 91 | 95 | 96 | 277 |
| | All schools | 78 | 84 | 87 | 804 |
| 2000 only, split half | 11-20 | 53 | 54 | 64 | 158 |
| | 21-40 | 62 | 74 | 85 | 330 |
| | 41-80 | 82 | 91 | 92 | 251 |
| | 81 or more | 100 | 100 | 100 | 21 |
| | All schools (weighting by number of schools originally in category) | 83 | 89 | 92 | 760 |
| 1999 and 2000 combined, split half | 11-20 | 51 | 59 | 68 | 37 |
| | 21-40 | 58 | 71 | 76 | 153 |
| | 41-80 | 80 | 86 | 92 | 343 |
| | 81 or more | 89 | 95 | 97 | 276 |
| | All schools | 78 | 85 | 90 | 809 |
| 1999 compared to 2000 | 11-20 | 40 | | | |
| | 21-40 | 45 | | | |
| | 41-80 | 66 | | | |
| | 81 or more | 76 | | | |
| | All schools | 65 | | | |

There is another caution that must be noted here: combining two years' worth of data must be done carefully (or perhaps, not done at all) if there has been substantial improvement in performance statewide between two years. Table 9 shows the same results as Table 8, but for the fifth grade ITBS rather than the fourth grade LEAP 21 test. First, note that while the split-half method produced estimates of reliability that were somewhat higher than direct computation on

LEAP 21, the opposite was true for the ITBS. However, that was primarily due to substantially lower results for the 26 largest schools (those with 81 or more students in each half), so that result might not be generalizable.

A more important finding comes from comparing the 2000 only split-half to the 1999 and 2000 combined split-half. Statewide scores increased between 1999 and 2000 more on the ITBS than they did on LEAP 21, so one would anticipate more of a stratification effect on that test. Indeed, that proved to be the case. The estimates of reliability were substantially higher for the estimates generated from two years' data than those for one year. It is clear that one must be very careful when combining two years' worth of data, and that significant improvements in scores statewide should be a warning flag. In such a case, it might be that the best way to combine the data would be to create one large data file across the two years, and then randomly assign students to their halves.

**Table 9**

**Percentage of Times Two Halves Are within One-Half of a Weighted School Mean Standard Deviation of Each Other on Total Math NRT Score**

| Group and Method of Computation | School Size | Percent Passing | Index | Mean | N |
|---|---|---|---|---|---|
| 2000 only, direct computation | 11-20 | 49 | 58 | 59 | 40 |
| | 21-40 | 62 | 72 | 73 | 158 |
| | 41-80 | 78 | 86 | 87 | 329 |
| | 81 or more | 90 | 96 | 96 | 277 |
| | All schools | 78 | 85 | 86 | 804 |
| 2000 only, split half | 11-20 | 48 | 63 | 66 | 155 |
| | 21-40 | 67 | 74 | 77 | 331 |
| | 41-80 | 77 | 85 | 88 | 225 |
| | 81 or more | 85 | 85 | 88 | 26 |
| | All schools (weighting by number of schools originally in category) | 76 | 82 | 85 | 737 |
| 1999 and 2000 combined, split half | 11-20 | 63 | 65 | 70 | 43 |
| | 21-40 | 58 | 74 | 81 | 160 |
| | 41-80 | 82 | 87 | 93 | 347 |
| | 81 or more | 93 | 96 | 97 | 250 |
| | All schools | 80 | 86 | 91 | 800 |

Impact of combining grades. Another way that one might increase the N in a school is by testing students in more than one grade. For the data set available to us, different tests were given to students at different grades, so we could not investigate the impact of combining grades when the tests are designed to be comparable across grades. However, this data set gave us a chance to see what might happen when the tests are, in fact, quite different from each other.

The results are provided in Table 10. One very interesting result is that, for a given number of students tested, the reliability of the system went down slightly when the tests were long (all four content areas on LEAP 21, the entire battery on the ITBS), but up somewhat when the tests were short. As in the previous analyses, one must be careful interpreting the data for "81 or more students," since the schools in that size category are much larger, on average, when two grades are combined, and the data for "All schools," since the number of students tested in schools is obviously much larger when two grades are combined than when looking at one grade by itself. However, perhaps the most important observation to make is that doubling the number of grades tested dramatically improves the reliability of school scores. While it isn't quite as much gain as one gets from testing two years at the same grade (compare Table 10 to Table 7), it still provides a great improvement. Again, while there are other variables that impact the reliability of a school's score, the most important one is the number of students tested—and it doesn't matter much whether that is because the school is large, we add more years, or we add more grades. Nothing else has as much impact on increasing the reliability of school scores as increasing the number of students tested.

**Table 10**

**Percentage of Times Two Halves Are within One-Half of a Weighted School Mean Standard Deviation of Each Other**

**Comparing One Grade to Two Grades, Using Direct Computation Method (By definition, a combination of NRT and CRT results)**

| Tests | School Size | Percent Passing | | Index | | Mean | | N | |
|---|---|---|---|---|---|---|---|---|---|
| | | One Grade | Two Grades | One Grade | Two Grades | One Grade | Two Grades | One Grade | Two Grades |
| All Content Areas | 11-20 | 50 | 47 | 64 | 60 | 67 | 61 | 40 | 14 |
| | 21-40 | 65 | 65 | 80 | 78 | 82 | 79 | 158 | 50 |
| | 41-80 | 81 | 79 | 92 | 90 | 94 | 91 | 329 | 169 |
| | 81 or more | 92 | 94 | 98 | 98 | 99 | 99 | 277 | 632 |
| | All schools | 80 | 89 | 90 | 95 | 92 | 96 | 804 | 865 |
| One Content Area | 11-20 | 48 | 45 | 54 | 51 | 58 | 55 | | |
| | 21-40 | 63 | 62 | 70 | 69 | 74 | 73 | | |
| | 41-80 | 79 | 76 | 85 | 83 | 88 | 86 | | |
| | 81 or more | 91 | 93 | 95 | 96 | 96 | 97 | | |
| | All schools | 78 | 87 | 84 | 91 | 87 | 93 | | |
| One Subtest | 11-20 | 39 | 39 | 44 | 45 | 47 | 47 | | |
| | 21-40 | 52 | 55 | 59 | 61 | 62 | 65 | | |
| | 41-80 | 67 | 69 | 75 | 76 | 78 | 79 | | |
| | 81 or more | 82 | 88 | 88 | 93 | 90 | 94 | | |
| | All schools | 68 | 81 | 75 | 87 | 78 | 89 | | |

Summary. In this section, we employed two methods for computing the reliability of school scores. The two methods generally give similar results, although it is clear that the split-half method can give misleading results if one is not careful about the assignment of students into halves. The split-half method also requires some cautions in the interpretation of results that the direct computation method does not.

Some of the more important findings about the reliability of school scores are as follows:

1. The most important factor, by far, in the reliability of a school score is the number of students that went into the computation.

2. One can increase the number of students in a school's score by aggregating data across grades or across years.

3. The reporting statistic used has an effect on the reliability of scores, although not nearly as much as the number of students tested. A four-point index, at least in these analyses, was almost as reliable as a mean score; but reducing data to pass/fail reduced reliability substantially.

4. Most of that reduction in reliability was undetectable when the criterion statistic was consistent classification within quarter or decile, because the unreliability of the statistic itself widened the boundaries of those cut-points.

5. When tests are more reliable, school scores are more reliable—but the effect is fairly small, even for great differences in the reliability of the tests. Thus, school scores will be more reliable if more students take less reliable tests than if fewer students take more reliable tests.

**The Reliability of School Gain Scores**

Background. While some states' accountability systems simply look at the performance of a school within a given year, most use *gain* for their criterion of school success. Therefore, we want to present methods for investigating the reliability of such systems, as well as present sample data to provide an idea on how reliable these systems are.

First, it is important to note that states simply have observed results when they look at changes in schools' scores from year to year. Thus, there is no way of telling *with certainty* whether a school has improved or not. However, there are methods for calculating the probability that a school really has changed, given its observed scores. Or more accurately, if one posits how much a school has truly changed from year to year, one can readily calculate the distribution of observed change scores one will encounter.

A second quick observation is that there is considerably more error in the estimate of gain than in that of performance in one year, for at least two reasons. First, when we compare one class of fourth graders to the next year's class at the same grade, we have two samples—each with its own independent error. The variance of the gain scores is the sum of the errors of the two years. Second, the variance of gain scores is far less than the variance of performance in any given year. To illustrate the issue, we took two years' worth of data in our sample state, and divided each year's sample in each school into random halves. The correlation between the odd and even half *within* a year was .96. But the correlation of the gain made by the even half versus the gain made by the odd half was .70.

As will be seen shortly, if a school is expected to make a large amount of gain, it is easy to detect from its observed scores whether change has truly taken place; but if the expected amount of gain is small, accurate detection can be problematic. The issue is signal-to-noise. Each observed

score for a school has a certain amount of uncertainty around it. If the amount of expected improvement is large relative to that uncertainty, it is detectable; but if the expected improvement is small, it likely will not be detectable whether the school has truly changed or not. Any changes in observed scores that are that small may be due to random error only, and not due to any real change in the achievement level of the students in the school.

To study this issue, we needed to create an arbitrary accountability system. We chose one in which the goal for a school would be to move from its current position to one of "near perfection" in 20 years. The definition of "near perfection" varied for the different tests and reporting statistics. For "percent passing," that was defined as 95 percent of the students passing the test; for the 4-point index, it was a value of 3.8 (95 percent of a perfect 4.0); for the mean scaled score, it was a mean value of 400 (or about two standard deviations higher than the current state average). The goals used in these analyses are higher than those that most states are using, but the timeline in our system is longer than most. In sum, the goals we set probably are somewhat higher than those that states are typically setting, but not much larger. To the extent that these goals are larger than the ones a state might set, these results will be *more* reliable than the ones a state might get if it did such a study on its own system. We strongly recommend that be done, rather than assuming these results are sufficiently generalizable. What we are hoping to do with these results is make people aware that (1) error rates are probably considerably higher than they think they are, and (2) they will not know what their actual error rates are until they do their own study.

Results using Direct Computation. The first results are presented in Table 11. To get these results, we asked three questions:

1. If every school in the state made no real improvement, what percentage of them would show observed gains that equal or exceed their goal, simply based on the random fluctuations that occur in schools' observed scores from year to year?

2. If every school in the state made improvement exactly equal to its goal, what percentage of them would show observed gains that equal or exceed their goal? (The answer to this question is easy—exactly half. This is true because the mean of the distribution of gain scores would equal their goal).

3. If every school in the state made real improvement equal to *twice* their goal, what percentage of them would show observed gains that equal or exceed their goal? (Because these distributions are presumed to be symmetrical, the answer to this question is 100 minus the answer to Question 1).

The following is an example of how to read Table 11: The result in the first cell tells you that a school with 11-20 students whose true score does not improve has a 42 percent chance of having its observed percentage of students passing all four content areas improve by as much as its improvement target. A similar school whose true score improves by twice as much as its improvement target has a 58 percent chance of having its percentage of students passing all four content areas improve by as much as its improvement target. In short, for schools that small, the likelihood that a school will reach its improvement target by chance alone is quite high; and even if makes substantial improvement, there is a good probability that its *observed* change score will not be much more than that of other schools that did nothing to improve.

Some of the results might be surprising to people looking at data like these for the first time. For example, even for a system administering a fairly complete battery of tests, if the system uses "percent passing" as its criterion statistic, 34 percent of schools of average size (41-80 students) that make *no* improvement will have observed score changes that get them labeled as a school meeting its improvement goal—and 34 percent of the schools that make *twice* their expected gain will get labeled as having made insufficient gain. The results are better for larger schools, and worse for smaller schools, of course; but that one finding alone might be of surprise to those unfamiliar with the amount of error surrounding school observed scores.

There are two points worth noting. As mentioned above, the gains expected from the system we started are somewhat larger than those expected by most states. That means that the results for most states would be somewhat *worse* than this. On the other hand, if the state's system incorporated two years' worth of data into its pre- and post-test results, one could look at the results for schools of twice the size in the table. Averaging two years' data for pre- and post-test results improves the likelihood of correct classification considerably.

**Table 11**

**Percentage of Times A School's Observed Score Will Improve as Much as Its Improvement Target, Depending on the Amount of True Gain for the School**

**Estimated Through Direct Computation**

| Tests | School Size | Percent Passing | | | Index | | | Mean | | | N |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | No Gain | Goal | 2 x Goal | No Gain | Goal | 2 x Goal | No Gain | Goal | 2 x Goal | |
| Four content areas | 11-20 | 42 | 50 | 58 | 39 | 50 | 61 | 37 | 50 | 63 | 40 |
| | 21-40 | 38 | 50 | 62 | 34 | 50 | 66 | 32 | 50 | 68 | 158 |
| | 41-80 | 34 | 50 | 66 | 29 | 50 | 71 | 27 | 50 | 73 | 329 |
| | 81 or more | 30 | 50 | 70 | 24 | 50 | 76 | 21 | 50 | 79 | 277 |
| | All schools | 34 | 50 | 66 | 29 | 50 | 71 | 26 | 50 | 74 | 804 |
| Total ELA | 11-20 | 45 | 50 | 55 | 41 | 50 | 59 | 39 | 50 | 61 | |
| | 21-40 | 43 | 50 | 57 | 38 | 50 | 62 | 35 | 50 | 65 | |
| | 41-80 | 40 | 50 | 60 | 33 | 50 | 67 | 30 | 50 | 70 | |
| | 81 or more | 38 | 50 | 62 | 29 | 50 | 71 | 24 | 50 | 76 | |
| | All schools | 40 | 50 | 60 | 33 | 50 | 67 | 29 | 50 | 71 | |
| Proofreading only | 11-20 | 44 | 50 | 56 | 42 | 50 | 58 | 40 | 50 | 60 | |
| | 21-40 | 41 | 50 | 59 | 39 | 50 | 61 | 37 | 50 | 63 | |
| | 41-80 | 38 | 50 | 62 | 35 | 50 | 65 | 32 | 50 | 68 | |
| | 81 or more | 34 | 50 | 66 | 31 | 50 | 69 | 27 | 50 | 73 | |
| | All schools | 37 | 50 | 63 | 35 | 50 | 65 | 32 | 50 | 68 | |

Table 12 looks at this same issue from another angle. Suppose no school in the state actually improved at all. The probability that a school's observed score would increase from one year to the next would be 50 percent. Now, suppose every school in the state made *two times* the amount of improvement expected of it in our accountability model. What percentage of them would have an observed score the second year that is higher than their observed score the first year? The answers to that question are in Table 12.

**Table 12**

**Percentage of Times A School's Observed Score Will Not Increase, Depending on the Amount of True Gain for the School**

**Estimated Through Direct Computation Reported by Amount of True Gain**

| Tests | School Size | Percent Passing | | Index | | Mean | | N |
|---|---|---|---|---|---|---|---|---|
| | | No Gain | 2 x Goal | No Gain | 2 x Goal | No Gain | 2 x Goal | |
| Four content areas | 11-20 | 50 | 34 | 50 | 29 | 50 | 26 | 40 |
| | 21-40 | 50 | 27 | 50 | 21 | 50 | 18 | 158 |
| | 41-80 | 50 | 22 | 50 | 15 | 50 | 12 | 329 |
| | 81 or more | 50 | 16 | 50 | 10 | 50 | 7 | 277 |
| | All schools | 50 | 21 | 50 | 15 | 50 | 12 | 804 |
| Total ELA | 11-20 | 50 | 40 | 50 | 33 | 50 | 29 | |
| | 21-40 | 50 | 36 | 50 | 27 | 50 | 22 | |
| | 41-80 | 50 | 31 | 50 | 20 | 50 | 16 | |
| | 81 or more | 50 | 27 | 50 | 15 | 50 | 10 | |
| | All schools | 50 | 31 | 50 | 20 | 50 | 16 | |
| Proofreading only | 11-20 | 50 | 37 | 50 | 35 | 50 | 32 | |
| | 21-40 | 50 | 33 | 50 | 29 | 50 | 26 | |
| | 41-80 | 50 | 27 | 50 | 23 | 50 | 19 | |
| | 81 or more | 50 | 21 | 50 | 17 | 50 | 12 | |
| | All schools | 50 | 27 | 50 | 22 | 50 | 18 | |

As can be seen by the results in Table 12, even if all the schools in the state made twice the amount of gain expected of them, a significant portion of them would actually see their scores decrease. Obviously, this would be a fairly rare event for larger schools when more reliable reporting statistics are used. However, if "percent passing" is the reporting statistic, a fairly high percentage of schools would be misclassified. A specific example follows: Suppose all schools in the state tested between 41 and 80 students, and we reported the percentage of students passing all four content areas. Suppose further that there are 200 schools in the state, with 100 making no

improvement in their educational program at all, while the other 100 make twice the improvement expected of them under our model. On average, 128 schools would show a gain, and 72 a loss. Of the 128 showing a gain, 50 would, in fact, have done nothing to improve. Of the 72 showing a loss, 22 would, in fact, have made educational improvements equal to *twice* our target. If we decided to give awards to any school that improved, we would give awards to 128 schools; but almost 2 in 5 would be getting awards because of random error, not real improvement. And of the 72 schools denied a reward, almost a third of them would actually have made improvements twice what we had asked for.

Results using split-half method. This process of hypothesizing changes in schools in the state and then looking at the changes in observed scores can be done by the split half method. The procedure is simple: randomly divide the students in each school into halves, add the amount of "true gain" to one of the halves, and then compare the two "observed" scores. Table 13 provides the results of such a calculation done for the same question as was answered by Table 12.

**Table 13**

**Percentage of Times A School's Observed Score Will Not Increase, Depending on the Amount of True Gain for the School**

**Estimated Through Split Half**

| Tests | School Size | Percent Passing | | Index | | Mean | | N |
|---|---|---|---|---|---|---|---|---|
| | | No Gain | 2 x Goal | No Gain | 2 x Goal | No Gain | 2 x Goal | |
| Four content areas | 11-20 | 50 | 28 | 48 | 28 | 46 | 22 | 158 |
| | 21-40 | 51 | 27 | 51 | 19 | 52 | 14 | 330 |
| | 41-80 | 45 | 24 | 47 | 15 | 49 | 9 | 251 |
| | 81 or more | 24 | 14 | 43 | 5 | 38 | 0 | 21 |
| Total ELA | 11-20 | 48 | 37 | 50 | 36 | 48 | 28 | |
| | 21-40 | 51 | 38 | 47 | 26 | 47 | 18 | |
| | 41-80 | 46 | 27 | 47 | 18 | 47 | 14 | |
| | 81 or more | 45 | 24 | 48 | 14 | 48 | 5 | |
| Proofreading only | 11-20 | 46 | 37 | 46 | 37 | 44 | 33 | |
| | 21-40 | 52 | 31 | 50 | 29 | 48 | 28 | |
| | 41-80 | 50 | 31 | 48 | 27 | 46 | 20 | |
| | 81 or more | 45 | 19 | 43 | 10 | 47 | 10 | |

As was true for the earlier analyses, one can see that the split-half method gives results that are similar to those of the direct-computation method, but that it indicates that the system is slightly more reliable than it actually is. Again, this likely is due to some small amount of stratification in the original data file which we divided into even and odd halves instead of truly randomizing the data.

Summary. In this section, we examined the reliability of gain scores. The observations made with performance scores are virtually identical: the split-half method gives similar, but not identical, results to the direct computation method; larger schools provide more reliable results than smaller schools; means provide more reliable results than indices, which in turn provide more reliable results than percent passing; and longer tests provide more reliable results than shorter tests. The importance of the factors is in that order: number tested, choice of reporting statistic, reliability of test.

However, the most important result of this section is that the reliability of gain scores is considerably less than that of performance in a given year. Even for schools of average size, the likelihood that they will be misclassified is considerable, even if we simply compare schools that have made no improvement at all to those that have improved substantially beyond the goals set by the state. This means that it is very important for states to do a study similar to the ones done here before assuming that the reliability of their accountability system is sufficiently high to warrant rewards or punishments.