Determining the Reliability of School Scores[1]

Richard Hill and Charles DePascale
The Center for Assessment

November, 2002

Introduction

States are creating procedures for determining whether schools have made "Adequate Yearly Progress" (AYP), as required by the No Child Left Behind Act (NCLB). While NCLB is highly prescriptive in regards to accountability design, states do, in fact, have choices to make. One consideration in choosing a method should be the likelihood of misclassifying whether schools have made the amount of improvement the state is requiring of them.

The purpose of this document is to provide states with the tools to determine the probability of misclassifying schools under varying accountability designs. Because each state will choose different combinations of the alternatives in front of them, and have different distributions of students across their schools, a data analysis done for one state might not pertain to another. Therefore, rather than supplying answers to questions people might not have, this paper is designed to help states answer the questions they have about their own proposed design.

The paper is divided into three chapters. In the first chapter, we explain why studies of the type outlined in this paper should be done. The second chapter discusses a variety of methods that can be used to determine the reliability of an accountability system. We recommend two that are used extensively in Chapter 3. In that chapter, we show how to make the calculations in a series of increasingly complex designs. For each set of calculations, we discuss the implications of the design on the probability of correct classification.

[1] We want to publicly acknowledge the many contributions Ed Haertel of Stanford University has made to this effort over the past three years. Ed outlined the first models for us to study reliability, provided numerous suggestions for us to pursue this work further, and provided all the equations used in pages 9-10.

Chapter 1—Fluctuation in School Scores

School scores vary from year to year. Even if the curriculum and instruction provided to students across years is identical and even if the general achievement levels of students enrolled in the school remain constant across, results will vary. The major sources of this variation (referred to as "volatility" by some authors) are sampling error (testing a different group of students each year) and measurement error (the variation associated with testing students on a particular occasion).

As will be shown later in this paper, sampling error contributes far more to the volatility of school scores than does measurement error[2]. Some classes of students simply outperform other classes, even when being exposed to the same curriculum and instruction. As a result, one school might outscore another in a particular year, even though its long-term average would be lower than its comparison school. Similarly, a school might show gains or losses from one year to the next, not because of improvements (or lack thereof) in its program, but simply because it was serving a more or less able group of students that particular year. As a result, a school might get one classification one year and a different one the next, even though no real changes had taken place in its program.

If this happened infrequently, we might accept the occasional error involved as a necessary cost of implementing an accountability program. But as will be seen from the analyses provided in Chapter 3 of this document, the volatility in school scores can be quite substantial. Some inferences about school scores can be made with a high degree of precision, but others cannot. When choosing an accountability design, one consideration should be whether volatility in the method chosen will permit correct classification of schools. Even the most seemingly valid accountability design will be flawed if there is so much volatility in the system that the labels schools receive are largely based on random error.

Some authors have questioned whether sampling error should be taken into account in considering student scores. Their logic is that all the students within the school are being tested, so why should those students be considered a sample? The following points have been raised to explain why sampling error must be taken into account:

1. When the results are reported, they are not attributed to a particular group of students, but to the school as a whole. As Dale Carlson has pointed out (personal communication), "the inference is about the school, based on the population of students the school serves." Since the inference is about the school, not a particular group of students, it is important to take into account the fact that the group tested in any particular year might not be representative of students in that school across years. Carlson further notes that if people were to insist that a particular group of students in, say, 2001, fully represents the school—is the sufficient definition of that school—then when a new group of students comes in 2002, they actually represent a new school. Under such a belief system, it would be impossible to have any school ever fail to meet AYP two consecutive years, since the group to which the inference was being made would never be the same across those years.

[2] That is not an original finding. Cronbach, Linn, Brennan and Haertel reported this in "Generalizability analysis for performance assessments of student achievement or school effectiveness" (*Educational and Psychological Measurement*, 57, 373-399.

2. In a conversation, Bob Linn noted that it would be inconsistent logic to take measurement error into account and not sampling error. That is, if one is going to include measurement error into account in determining a school's score, it is in recognition of that fact that upon another occasion or faced with another sample of test questions from the universe of possible items, a student who failed once might pass the test a second time. But the student tested in the fourth grade one year is just one possible student to represent that school; another student in the first student's place might also have a different result from the first student. If one believes that the students tested one year should be an error-free representation of a school, how can one believe that the items chosen for the test, the scoring of that test, and the occasion on which the student took the test also are not fixed?

3. Perhaps the strongest argument for taking sampling error into account is not logical, however, but simply observing the fact that sampling error provides an *underestimate* of the volatility of school scores from year to year. Kane and Staiger (2002) separate the sources of fluctuation in school scores from year to year into three categories: "sampling variance," "other non-persistent changes," and "persistent effects." Measurement error is one of the "other non-persistent changes." They proceed to show that (1) the amount of non-persistent variation in scores from year to year for their set of data is greater than would be expected from sampling variance alone but (2) if the design of the system compares different cohorts of students, the variance due to sampling of students is far larger than the variance due to other non-persistent changes (including measurement error). Thus, while we might debate whether to take sampling error into account from a philosophical point of view, the data are conclusive that one will grossly underestimate the volatility of school scores if only measurement error is taken into account and sampling error is ignored.

That issue leads us to the final point of this first chapter. Because the effectiveness of instruction *does* vary in schools from year to year, and it varies in different, *unknown* amounts, the studies we are describing in this publication cannot be done by simply looking at the actual changes in test scores across years. The purpose of these studies is to determine the likelihood that a school will be classified as "improving" if it really is improving, and "not improving" if it really is not improving. Since the data we obtain from schools across years include schools from both improvement categories, and since we cannot tell from the observed results which schools belong in which category, observed data cannot answer the questions we have. While we need to start with one year's worth of real data from a state to do these analyses, we cannot answer the research questions posed by simply comparing observed school results across time. To do the analyses correctly, one must create a model of school performance that includes sampling error and then determine the percentage of times schools are correctly classified.

Chapter 2—Determining Sampling Error

Four Available Methods

There are at least four distinct methods for computing the reliability of an accountability system: direct computation, split-half, random draws with replacement, and Monte Carlo. Each of those methods will be discussed in turn below.

Direct computation. "Direct computation" involves computing the errors around estimates and using areas under the normal curve to determine the probability of correct classification. This method is described in some detail in "The Reliability of California's API," a publication available on the Center's website (www.nciea.org).

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

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

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

While this method is simple to execute, there are a large number of places where one can go wrong. Obviously, a critical assumption behind the method is that the students are placed into the two groups randomly, thereby simulating the process of creating new classes of students each year. Perhaps the simplest way to execute this method is to place all the students who have an odd number in the data file into one group and the even-numbered students in the other. Often, however, students are not randomly placed in the data files provided to researchers. If the file is sorted in some way, the system will over-estimate the reliability of the system. One state implemented this method using a file that was sorted by students' scores on the test—as a result, they came to the (mistaken) conclusion that their system was highly reliable.

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

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

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

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

This method has the advantage of simplicity and appears to provide generally accurate results. We will use this method extensively in Chapter 3 of this report, and thus will not discuss its application further here.

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

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

Once that is done, all the variance components can be estimated using the equations provided in Chapter 3. If the students are normally distributed within school, one can produce a set of scores that will mirror the original distribution through the following steps:

1. Starting with the original observed mean score for a school, estimate the true score mean for that school (using the notation introduced in Chapter 3, $\overline{T}$).

2. Knowing the standard deviation of student true scores within school and the n-count for that school, make a random draw of the true score for that school, by selecting a random normal variate and computing $\overline{T_0} = \overline{T} + z * \sigma_{T|S} / N$.
3. Given the chosen true score for that school and the standard deviation of student observed scores within school, make N draws of a random normal variate and compute $X = \overline{T_0} + z * \sigma_{X|S}$.

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

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

The General Procedure for Computing Decision Accuracy

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

Issues of Concern

Each of the methods contains places where even a careful researcher can make incorrect assumptions, leading to erroneous results. For example, carefully determining an appropriate value for the variance of students within school for each school is a critical issue, since the variance of students within school, along with sample size, determines the variance error of the school mean. A second area to be concerned about is whether the errors of successive draws are independent. The errors will not be independent in at least two cases: (1) when students are tested in more than one content area, and (2) when some of the students tested in one year are the same students tested the next year in a school.

NCLB, for example, requires that schools be evaluated on the performance of students in reading and mathematics. An incorrect way of proceeding would be to draw one random sample of students to obtain the reading scores for a school and then another random sample to get the mathematics scores.

That would make the sample for reading locally independent of the sample for mathematics. However, if a particular draw of students for a school is good in reading, it is likely that the same draw would be at least better than average in mathematics; and conversely, if the draw were weak in reading, it would be poor in mathematics. Similarly, if an accountability system included assessment results from consecutive grades, a school that had a lucky draw for, say, Grade 3 in Year 1, many of those same students would still be in the same school for Grade 4 in Year 2, and therefore produce a higher-than-average result for the school in both years. The correlation of these results is particularly important if schools are going to be evaluated on their progress from year to year. Assuming the draws for each year are independent likely will yield estimates of the reliability of the system that are too low. In Chapter 3, we will outline how to address the first of these issues in applying the sampling techniques; we expect to provide guidance on the second issue in a future revision of this paper.

Because it is possible to make errors in logic along the way that will lead to errors in computing the classification errors of various systems, we recommend that results be obtained by at least two of these methods and compared. The results should be independent of the method, and therefore very much the same across method. Any time there are significant differences across methods, the accuracy of the results is brought into question, so when differences are found, they need to be resolved before proceeding further. In Chapter 3, we will use two different methods for several of the studies, and show the similarity of the results across the methods. We believe this should be standard practice for anyone conducting studies of this nature.

Chapter 3—Applications

In this chapter, we will pose a series of increasingly complex questions, and in the process of answering them, hopefully gradually build the reader's skill in conducting reliability studies. For each substudy, we will pose the issue, describe the process used, provide the SAS program used to generate the results in the appendix and discuss the implications of the results. For the first several studies, we will use the Monte Carlo method to generate data; we will duplicate the last few (most complex) of the Monte Carlo studies using the "random draws" method, and show the similarity of the results using the two different methods. Finally, we will use the random draws methods to determine the reliability of designs that are too complex to do easily with the generated method: those employing subgroups and those measuring school change over time when the same students are assessed in different years.

Study 1—Generating two sets of scores for each school

This first study is designed to show the steps that one needs to take to estimate the various variance components needed to determine the reliability of school scores and to show that the method allows one to produce randomly generated sets of data that actually do behave just like the original data.

To begin, you must know the following, all of which can be calculated from your own data files[3]:

1. The mean of student scores
2. The standard deviation of student scores
3. The standard deviation of school mean scores
4. The reliability of the test
5. The shape of the distribution of student scores

Throughout this paper, we will be working with the underlying distribution of student scores, and then converting those results to performance levels or pass/fail judgments as appropriate. We strongly recommend this approach (in contrast to working with distributions of levels, for example), in that it avoids many potential areas for error. If scaled scores are not available, these analyses can

[3] It will be assumed throughout these analyses that the student scores are distributed fairly normally. If they are not, some adjustment must be made. There are three ways of making the adjustment:

1. Apply a transformation to the scores that will make them normally distributed. For example, if you are using percentage correct scores from a fairly easy test, the distribution of scores will be negatively skewed. A "log-odds" transformation usually will result in a distribution that is far more like a normal distribution. To make a log-odds transformation, divide the percentage the student got right by the percentage the student got wrong (the odds of a right answer) and then take the natural logarithm of that result. So, for example, the log-odds transformation of 90 percent correct is ln (90/10), or 2.197. If a student receives a score of 100 percent on the test, the division by zero will produce an error; the standard approach to this problem is to assume the student got 99.5 percent correct, rather than 100, which leads to a log-odds score of 5.29 for these students.
2. Compute a percentile rank for each score and convert that into a z-score from a normal distribution. So, for example, if a student's score was the 84th percentile (of the state's own percentile rank distribution, the student would get a z-score of 1.00.
3. Rather than draw from a normal distribution (as we do in the examples we provide), make random draws from a distribution that more closely matches your distribution of student scores.

be done using raw scores, provided there is assurance that the raw scores comprise a reasonably interval scale.

To proceed, we need to establish the following notation:

Let: $\sigma^2{}_X$ = the variance of pupil observed scores,
$\sigma^2{}_T$ = the variance of pupil true scores,
$\sigma^2{}_E$ = the variance of error in pupil scores,
$\sigma^2{}_{\overline{X}}$ = the variance of school observed mean scores, if the population of students in the school's catchment area were tested
$\sigma^2{}_{\overline{X_0}}$ = the variance of school observed mean scores if one sample of size *N* were drawn for each school,
$\sigma^2{}_{\overline{T}}$ = the variance of school true mean scores, if the population of students in the school's catchment area were tested
$\sigma^2{}_{\overline{T_0}}$ = the variance of school true mean scores if one sample of size *N* were drawn for each school,
$\sigma^2{}_{X|S}$ = the variance of pupil observed scores within school,
$\sigma^2{}_{T|S}$ = the variance of pupil true scores within school,
r_X = the reliability of pupil scores across all pupils,
$r_{\overline{X_0}}$ = the reliability of one observation of a school mean score, and
N = the number of students in each school.

Note especially the distinction between variances of school means that have a subscript and those that do not. When the subscript is present, the reference is to school means that are derived from a single sample of size N; when the subscript is not present, the reference is to school means that would be obtained if the entire population of students were tested.

The following equations allow us to make two critical calculations: the variance of school true means ($\sigma^2{}_{\overline{T}}$) and the variance of student observed scores within school ($\sigma^2{}_{X|S}$). Those familiar with classical measurement theory will know each of these equations. A short explanation of each equation is provided.

(1) $r_X = \sigma^2{}_T / \sigma^2{}_X$—The reliability of student scores equals the variance of true scores divided by the variance of observed scores

(2) $\sigma^2{}_X = \sigma^2{}_T + \sigma^2{}_E$—The variance of observed scores equals the variance of true scores plus the error variance

(3) $\sigma^2{}_{\overline{X_0}} = \sigma^2{}_{\overline{T_0}} + \sigma^2{}_E \;/\; N$—The variance of school mean observed scores (one sample) equals the variance of school mean true scores (one sample) plus error variance divided by the number of observations

(4) $\sigma^2_{\overline{T_0}} = \sigma^2_{\overline{T}} + \sigma^2_{T|S}/N$—The variance of school mean true scores (one sample) equals the variance of school mean true scores plus the variance of true scores within school divided by the number of observations[4]

(5) $\sigma^2_T = \sigma^2_{\overline{T}} + \sigma^2_{T|S}$—The variance of true scores equals the variance of school means true scores plus the variance of true scores within school.

(6) $r_{\overline{X_0}} = \sigma^2_{\overline{T}} / \sigma^2_{\overline{X_0}}$—The reliability of school means equals the variance of school mean true scores divided by the variance of observed means (one sample)

(7) $\sigma^2_{X|S} = \sigma^2_{T|S} + \sigma^2_E$—The variance of observed student scores within school equals the variance of true scores within school plus error variance

(8) $\sigma^2_X = \sigma^2_{\overline{X}} + \sigma^2_{X|S}$—The variance of student observed scores equals the variance of school mean scores plus the variance of students within school (this is the same as Equation 5, but for observed scores instead of true scores).

Let's work through an example to see how these equations work together to allow us to make the necessary computations. Suppose we know the following statistics for the distribution of scores for our state:

1. The mean of student scores = 300
2. The standard deviation of student scores (σ_X) = 100
3. The standard deviation of school mean scores ($\sigma^2_{\overline{X_0}}$) = 40
4. The reliability of the test (r_X) = .90

Let us suppose further that the number of students in a school is 50.

Knowing that the reliability of the test (at the student level) is .90, we can compute the variance of student true scores from Equation 1:

$$\sigma^2_T = .9\ (10000) = 9000$$

Knowing that the variance of student true scores is 9000, we can compute the error variance of student scores from Equation 2:

$$\sigma^2_E = 10000 - 9000 = 1000$$

[4] Bob Brennan pointed out that this equation is not quite correct. The equation is true only if there is an infinite number of schools. If the number of schools is equal to *n*, then the correct equation is $\sigma^2_{\overline{T_0}} = \sigma^2_{\overline{T}} + ((n-1)/n)\,\sigma^2_{T|S}/N$. This correction for a finite number of schools will have only a very small effect for most states, and is ignored for practical purposes in subsequent calculations. A proof of Brennan's correction is provided on the NCIEA website.

Knowing that the error variance of student scores is 1000, we can compute what the variance of school mean true scores (taking one sample of students) would be from Equation 3:

$$\sigma^2_{\overline{T_0}} = 1600 - 1000/50 = 1580$$

Given that value, we can use Equations 4 and 5 to create two equations that have two unknowns—the variance of school true scores and the variance of student true scores within school:

$$\sigma^2_{\overline{T}} = 1580 - \sigma^2_{T|S} /50$$

$$\sigma^2_{\overline{T}} = 9000 - \sigma^2_{T|S}$$

Solving 4 and 5 simultaneously, we compute that $\sigma^2_{\overline{T}}$ = 1428.5715 and $\sigma^2_{T|S}$ = 7571.4285

As an interesting aside, we can now compute what the reliability of school mean scores will be for our state for schools with 50 students, using Equation 6:

$$r_{\overline{X_0}} = 1428.5715 / 1600 = .893$$

Note that the reliability of school mean scores is almost exactly the same as the reliability of student scores (.893 vs. .90). This is complete coincidence. In Hill (2002), under "Section B: Factors Affecting Reliability," there is a discussion of the relationship between the reliability of student scores and the reliability of school scores. Table 1 of that publication provides some examples that show the reliability of school mean scores is only minimally affected by the reliability of the student tests used in the accountability system; the primary factor driving the reliability of school mean scores is the number of students in the school. Even when the test used is highly reliable, school mean scores will not be reliable if there are a small number of students in the school; conversely, if there are a large number of students in a school, the reliability of the school mean score will be high even when the reliability of the test is low.

Now, from Equations 7 and 8, we compute the two statistics that we need for our Monte Carlo studies—the variance of students within school (Equation 7) and the variance of observed school means (Equation 8):

$$\sigma^2_{X|S} = 8571.4285$$

$$\sigma^2_{\overline{X}} = 10000 - 8571.4285 = 1428.5715$$

Note that $\sigma^2_{\overline{X}} = \sigma^2_{\overline{T}}$. That is true because testing all the students in a school once would give us the same mean as the mean true score for the school. That is, testing an infinite number of students just once would produce the same average as testing that infinite number of students an infinite number of times. Both results would give us an error-free parameter for the school. Thus, the variance of those scores will be the same.

Now, let us apply this information to our first program. Again, note that the program is provided in the appendix. First, we pick a random normal variate ("NORMAL(0)" in the SAS program) and use that to generate a random school true mean score, given a normal distribution where the mean is 300 and the standard deviation is 37.796 (the square root of 1428.5715, as calculated above). Then, once we have the mean for the school, we draw 50 random students around that mean by drawing 50 random normal variates, multiplying each by the square root of 8571.4285, and adding that result to the school mean. We then output the 50 student scores and the school mean. Then, we do that all over again (simulating drawing a second random sample for the school around the same true mean score we originally chose for this school). This process—drawing a possible true mean for a school and then drawing two samples of 50 students each for that school—then is done 100,000 times.

The following is the output from running that computer program one time:

```
                      MANUFACTURED DATA FOR RILS--N=50, RATIO = .84                 1
                                                      11:31 Friday, October 25, 2002

                                   The MEANS Procedure

                               Analysis Variable : STUDSS

               N            Mean         Std Dev         Minimum         Maximum
         -----------------------------------------------------------------------
         5000000     300.2898281      99.9674695    -213.6537509     802.0227234
         -----------------------------------------------------------------------




                  MANUFACTURED DATA FOR RILS--N=50, RATIO = .84                     2
                                                      11:31 Friday, October 25, 2002

                                    The CORR Procedure

                             2  Variables:    MEANSS1  MEANSS2


                                    Simple Statistics

Variable         N        Mean       Std Dev           Sum       Minimum       Maximum

MEANSS1       100000   300.28983      39.93733      30028983     114.75254   470.16316
MEANSS2       100000   300.26852      39.91347      30026852     127.37330   478.89028




                      Pearson Correlation Coefficients, N = 100000
                               Prob > |r| under H0: Rho=0

                                         MEANSS1       MEANSS2

                           MEANSS1       1.00000       0.89381
                                                        <.0001

                           MEANSS2       0.89381       1.00000
                                          <.0001
```

The program generated data on 5,000,000 students, yielding a mean of 300.29 and a standard deviation of 99.97 (in contrast to the population parameters for these two statistics of 300 and 100, respectively, that we were trying to produce in our data set). The standard deviation of school mean scores is 39.9 (in contrast to population parameter of 40), and the correlation between two random draws is .894 (in contrast to the value calculated from Equation 6 of .893). All these results are within one standard error of the mean. If another sample of 100,000 schools had been drawn, it just as likely might have had a mean slightly less than 300 (rather than larger, as was the case for this sample) and a standard deviation slightly greater than 40 (rather than less, as in this sample). So, we have demonstrated that we can produce a random set of data that has statistics very similar to the population parameters we desire.

Study 2—Comparing the relative efficiency of various reporting statistics

From the first study, we showed that we know how to generate random scores for students that match some predetermined distribution. The results of that study, however, didn't produce anything beyond the parameters we started with, so we didn't learn anything new from it. In this study, we will put to use the knowledge of how to produce a data set in order to answer some useful questions:

> *There are many ways to report student scores. Some of them include:*
> 1. *Mean scaled score*
> 2. *Index (reporting scores in terms of performance levels and then converting each of the levels into a value, such as Below Basic = 1, Basic = 2, Proficient = 3, Advanced = 4)*
> 3. *Percentage of proficient students*
>
> *What is the reliability of each of these statistics for various school sizes? What is the relative efficiency of each statistic?*

To answer these questions, we will generate two sets of student scores for each school as we did in Study 1. After generating each set, we will convert them into them the various reporting statistics and compute the correlation across schools.

To start, we need to create student and school statistics. We will use the same values we used in Study 1. Then, we will convert the student scaled scores into various statistics. For purposes of this study, we will have 25 percent of our students Below Basic (a scaled score of less than 232), 25 percent Basic (a scaled score between 232 and 300), 40 percent Proficient (a scaled score between 300 and 428) and 10 percent Advanced (a scaled score at or above 428). Of course, if we were a particular state, we already would have established cut scores for our performance levels and would use those in place of these. But the percentages of students we have at each of our levels are not atypical of the results for many states.

To create an index, we will assign 1 point for Below Basic students, 2 points for Basic, 3 for Proficient and 4 for Advanced. Given the percentages of students at each of the levels, the average index for our example will be 2.35. As is the case for all the decisions in this paper, other choices could have been made. Each state, if it is using an index, will know the process for converting student performance level to the index. It might, or might not be, similar to the choice we made for this paper.

Performance	Points for	% of	Points x %

Level	Level	Students at Level	
Below Basic	1	25	25
Basic	2	25	50
Proficient	3	40	120
Advanced	4	10	40
Total			235

Finally, since the reliability of a pass/fail statistic is dependent on the percentage of students passing, we will consider three different cut scores for passing: Basic or above (referred to in the tables that follow as "Pass/Fail, pi = .25," since 25 percent of the students fail this test), Proficient or above ("Pass/Fail, pi = .50"), and Advanced ("Pass/Fail, pi = .90").

To see how the assignment of students to the various reporting statistics was done, refer to the computer program for Study 2 in the appendix. Note at the beginning of that program that the school size (NSTUD) has been set to 20. This program can be run repeatedly for different sized schools simply by changing that parameter at the beginning of the program.

The next few statements are calculations made from Equations 1-8 to determine $\sigma^2_{\overline{T}_0}$ (VARTBAR0), $\sigma^2_{T|S}$ (VARTWS), $\sigma^2_{\overline{T}}$ (VARTBAR), $\sigma_{\overline{T}}$ (SDTBAR), and $\sigma^2_{X|S}$ (VAROWS).

Once those basic computations are made, this program basically does the same thing as the program for Study 1. That is, it draws a random normal variate and uses that to create a school mean. Then it draws NSTUD random normal variates to create a simulated set of student scaled scores for that one school, converting the scaled score for each student into various reporting statistics as it goes along. It then computes the totals for the school. After that, the process is repeated to draw a second sample of size NSTUD for that school. We now have two randomly drawn sets of statistics for that first school. That process is repeated 100,000 times.

The results are provided in Table 1. For each of several sizes of schools, we computed the correlation between the pairs of statistics across the 100,000 schools generated. Note that the results for Proficient or above ("Pass/Fail, pi = .50") are printed in the row above those for Basic or above. We reversed the order of the reporting because the reliability of the percentages Proficient or above is higher for that of Basic or above (for the same reason that a test is more reliable if the average difficulty of the items is around .50 rather than some more extreme value).

Table 1

Correlation between Two Random Draws for Each School

Statistic	Number of Students										
	20	30	40	50	60	70	80	90	100	120	150
Mean scaled score	.73	.82	.87	.90	.91	.92	.93	.94	.95	.96	.97
Index	.70	.79	.85	.88	.90	.91	.92	.93	.94	.95	.96
Pass/Fail, pi=.50	.63	.74	.80	.84	.86	.88	.89	.91	.92	.93	.94
Pass/Fail, pi=.25	.59	.71	.77	.82	.84	.86	.88	.89	.90	.92	.94

Pass/Fail, pi=.90	.49	.61	.70	.74	.78	.81	.83	.85	.86	.88	.91

There are two obvious trends in the data, neither of which is surprising. The results for larger schools are more reliable than those for smaller schools, and the mean scaled score is the reporting statistic with the highest reliability, followed by the index, followed by each of the pass/fail statistics, with the pass/fail statistics with closer to 50 percent of the students passing are more reliable than those further away from that point. Mean scaled score is expected to be the most reliable statistic because it contains the most information: when we convert mean scaled scores to performance levels, some information is lost. For example, with scaled scores we would know which was the higher-scoring of two students at the Below Basic level. Once we convert those scores to Below Basic, we lose that information. Less information translates into lower reliability. Similarly, an index is more reliable than a pass/fail judgment, because we lose information when we make the pass/fail conversions. For example, if one student is Basic and another is Below Basic, we know which is the higher scoring of the two. But if we convert scores into pass/fail, with Proficient as the minimum level to be passing, we lose the information about which was the higher scoring of those two students.

To refine that finding somewhat, let's take a look at how many students are required to be in a school to have the reliability for the school be a constant value for various reporting statistics. For example, it takes 50 students to have a reliability of .90 when the reporting statistic is the mean scaled score, but 60 when the statistic is the index, about 85 when the statistic is Pass/Fail, pi = .50, 100 for pi = .25 and 140 for pi = .10. Table 2 summarizes the numbers of students required to obtain reliabilities of different sizes for different reporting statistics. Note that the exact reliability coefficient required does not appear in Table 1 for each of the reporting statistics. In those cases, interpolation was used. So, for example, there is no place in the table where "mean scaled score" has a reliability of .84. When there are 30 students per school, the reliability is .82; when there are 40 students per school, the reliability if .87. So interpolation tells us that the approximate school size needed for a reliability of .84 with a mean scaled score is 34, which is the value used in Table 2.

Table 2

Approximate Ratios of Numbers of Students Required to Maintain Constant Reliability

Statistic	r = .84		r = .88		r = .90		r = .92		Ave. Ratio
	N	Ratio	N	Ratio	N	Ratio	N	Ratio	
Mean scaled score	34		43		50		70		
Index	38	1.12	50	1.16	60	1.20	80	1.14	1.16
Pass/Fail, pi=.50	50	1.47	70	1.63	85	1.70	100	1.43	1.56
Pass/Fail, pi=.25	60	1.76	80	1.86	100	2.00	120	1.71	1.83
Pass/Fail, pi=.90	85	2.50	120	2.79	140	2.80	--	--	2.70

What Table 2 shows is that it requires about 16 percent more students to achieve the same level of reliability when an index is the reporting statistic instead of a mean scaled score, but over 50 percent more when pass/fail is the reporting statistic—and that is true only for the center of the distribution. When more extreme percentages of students pass (or fail), it can require twice as many students or more to maintain the same level of reliability as would be obtained with a mean scaled score. Thus, the decision to choose a pass/fail system of reporting should not be made lightly—a significant amount of reliability is sacrificed. A fair trade-off, for example, would be to report results for schools of 50 students if a mean scaled score were the reporting statistic, but require schools to have at least 80 students if a pass/fail statistic were chosen. Both results would be of approximately equal reliability.

Study 3—Computing the probability of classification errors for a status design

To this point, we have used reliability as the statistic of interest. We have done that only because we have been looking at the relative efficiency of various reporting statistics. However, a reliability coefficient is not of interest here; what we truly would like to know is the probability of classification errors. If a school is making Adequate Yearly Progress, it should be labeled as such; and the converse is true as well. The reliability of the process is not the most relevant statistic here; we are concerned about classification errors.

Under NCLB, a school with sufficiently high scores in any particular year (the school's "status" score) will be labeled as having made AYP. In this sense, AYP has nothing to do with "progress." Until the "safe harbor" provision is invoked, a school is not evaluated on its improvement from year to year, but whether its scores in any particular year are sufficiently high.

How high is sufficiently high? NCLB requires the starting point to be, at a minimum, the 20th percentile of schools. For the statistics being using in this report, the 20th percentile of schools is as follows:

- SS = 266
- Index = 2.05
- Percentage passing = 85 when pi = .10
- Percentage passing = 65 when pi = .25
- Percentage passing = 36 when pi = .50
- Percentage passing = 4 when pi = .90

So, for example, a school with a *true* mean scaled score greater than 266 *should* be classified as having made AYP. A school with an *observed* mean scaled score greater than 266 *will* be classified as having made AYP. The purpose of this study is to answer these two questions:

- How often *will* a school be classified as having made AYP when it *should* be classified as having made AYP?
- How often *will* a school be classified as not having made AYP when it *should* be classified as not having made AYP?

To perform this study, we first need to create a true score for a random school and then determine whether it should be classified as having made AYP. To do that, we need to know the percentage of students in its population that fall into each of the categories. Given the variance of students within school and the assumption of a normal distribution of students within the school, we can compute all those percentages once we know the true mean for the school. The computer program developed for Study 3 does that. It starts by selecting a mean for a randomly chosen school, then computes the percentages of students in each of the four performance levels *in that school* (called P1 – P4 in the program). So, once the school has been chosen, we know whether it *does* meet the NCLB criteria for satisfactory status. The program then proceeds to draw NSTUD random students for that school and computes whether that sample will meet the status requirement for NCLB. The code near the end of the program simply determines for each reporting statistic whether the sample meets AYP (all the status indicators ending with a “1”), whether the school population meets AYP (all the status indicators ending with a “2”), and then produces a 2x2 cross-tab of the results.

Table 3 provides the percentages obtained from our sample of 100,000 schools. Remember, by definition, 20 percent of the schools do not meet the AYP status criteria and 80 percent do. The first value is the percentage of samples that should make AYP and do; the second value (after the “+” sign) is the percentage that should not make AYP and do not.

Table 3

Probability a School Will Be *Correctly* Judged on Status

Statistic	N = 20	N = 50	N = 100
Mean scaled score	76+12	77+16	78+17
Index	76+12	76+16	77+17
Pass/Fail, pi=.50	72+12	76+15	77+17
Pass/Fail, pi=.25	75+11	74+16	76+17
Pass/Fail, pi=.90	72+8	74+12	75+14

Here is the first surprise. Although Study 2 confirmed that the various reporting statistics have reliability in the order listed, that additional reliability proved to have a very minor impact on the correct classification of schools. Even for small schools (just 20 students), the percentage of schools that were accurately classified ranged from 80 to 88 percent—and the mean scaled score was hardly more accurate than an index. For schools with 100 students, even the least reliable reporting statistic accurately classified 89 percent of the schools. While it is not a surprise that schools can be accurately classified on *status* (as we shall see, *improvement* is another story) even when Ns are fairly small, it is somewhat surprising that the least reliable statistic (pass/fail when only 10 percent of the students are passing) still produces accurate classification a substantial portion of the time.

Another finding is not surprising. More school are identified as not having made AYP than should be, and the less reliable the reporting statistic, the more of a problem this is. The reason why this is not a surprise is that the variance of observed scores is greater than the variance of true scores, and the degree of increase is a function of the reliability of the scores. Less reliable observed scores can have a variance far greater than that for true scores. To see how this plays out, let's take a look at the data in the first cell of Table 3. It tells us that 76 percent of the schools that have a true mean scaled score greater than 266 had an observed score greater than 266, and 12 percent of the schools that have a true mean scaled score less than 266 had an observed mean scaled score of less than 266. We know that 20 percent of the schools have an *observed* score less than 266. Those observations permit us to fill in the following information in Table 3A, along with some unknowns:

Table 3A

Probabilities of Observations, Given *N* = 20 and Reporting Statistic of *Mean Scaled Score*

Observed Status	True Status		Total
	< 266	> 266	
> 266	76	A	80
< 266	B	12	20
Total	C	D	100

Now, we know that 76 percent of the schools belong in the first cell, and that 80 percent of the schools have an observed score of 266 or above. That means that A = 4. In turn, that means that D = 16, which in turn means C = 84, which in turn means B = 8. Table 3B is the same as Table 3A, but with all the missing values filled in.

Table 3B

Probabilities of Observations, Given *N* = 20 and Reporting Statistic of *Mean Scaled Score*

Observed Status	True Status		Total
	< 266	> 266	
> 266	76	4	80
< 266	8	12	20
Total	84	16	100

The first observation is that even though only 16 percent of the schools have a true score below 266, 20 percent of the observed schools do. A statistic with lower reliability (either because of small N or the choice of a less reliable reporting statistic) will have a greater disparity than that. Second, although we commented earlier about the relatively accurate placement of schools into cells, note that of the 20 schools that would be identified because they had an observed score less than 266, only 12 deserve to be so labeled. Eight of out the 20, or a full 40 percent of the schools identified, are identified in error! Note that it is unlikely for a school with a true score less than 266 to have an observed score greater than 266—only 4 of the 80 "passing" schools have such low true scores. But of the schools that are identified, a substantial portion are identified in error. This type of analysis—looking at the percentage of identified schools that truly do not deserve to be identified—is something that every state should do as part of its process of looking at the classification error of its proposed accountability design.

Study 4—Computing the probability of classification errors for an improvement design: Part 1—When schools make no improvement

Under NCLB, if a school's status is not sufficient high, it still can avoid being labeled as not having made AYP if it reduces the percentage of students failing by 10 percent. That is, if a school had 40 percent of its students failing the year before, it makes AYP if the next year it has no more than 36 percent of its students failing. Even if a school makes no true improvement from year to year (i.e., its true score remains the same over that time), it might make AYP simply because the sample of students drawn the second year is sufficiently better than the sample drawn the first year. Study 4 is designed to determine how often this might occur.

Judging a sufficient amount of improvement for the pass/fail systems is straightforward, since NCLB prescribes it as reducing the percentage of students failing by 10 percent. For the mean scaled score and index, however, some judgment is required. If the standard deviation of students remained the same, the mean scaled score would need to increase to 465 to have 95 percent of the students passing, so sufficient improvement for a scaled score was established as the distance between a school's current observed score and 465, divided by 12 (the number of years NCLB provides schools to get 95 percent of their students proficient). Establishing goals for improvement for an index was even more judgmental; we established a goal of 3.5 for the index, figuring that such an average would likely have at least 95 percent of the students proficient. Thus, the annual goal for schools was established as the distance between their current index and 3.5, divided by 12.

The program that makes these calculations is provided in the appendix. This program is not too different from the previous one. The major difference is at the end of Year 1, it calculates a growth target for the school (GT) for each of the reporting statistics. It then draws a second sample of students under the assumption of no change in the school's true score for the second year, and determines whether the amount of improvement meets the NCLB requirements. The results are shown in Table 4. Note that the ideal would be to have 100 percent in each of the cells—all the schools in this model have made no improvement, and therefore should be judged as having made no improvement. Note also that we have added one additional reporting statistic—pass/fail when the standard is set so low that only 10 percent of the students fail. We decided this was important to do; with the status test, the results were symmetrical, so 10 percent passing gave the same results as 90 percent passing. But with the improvement test, the growth target is dependent on the stringency of the standard. If lots of students fail, schools' growth targets are large, and therefore it is unlikely that a school that has made no improvement will be judged as having made AYP. However, if few students fail, most schools will have small growth targets and therefore many more will have a sufficiently reduced percentage of students failing even if the school truly has done nothing to improve.

Table 4

Probability a School Will Be *Correctly* Judged on Improvement
If It Makes No Improvement

Statistic	N = 20	N = 50	N = 100
Mean scaled score	69	78	86
Index	63	70	78
Pass/Fail, pi=.50	63	70	77
Pass/Fail, pi=.25	58	61	66
Pass/Fail, pi=.10	55	57	59
Pass/Fail, pi=.90	82	94	98

These results show three things, none of which is particularly surprising:

1. The probability that a school's observed results will show that it has made AYP even when it has made no improvement decreases when the school gets larger.
2. The probability of such an event is considerably higher than the probability of misclassification on the basis of status. Remember, in Table 3 we saw that the vast majority of schools were correctly classified on status. While a majority are correctly classified in this table, the percentages are not nearly so high. Table 3 showed that 88 percent of the schools were correctly classified on status when there were just 20 students per school; Table 4 shows that only 86 percent of the schools are correctly classified as having made no improvement even when there are 100 students per school.
3. The choice of reporting statistic makes a substantial difference. In contrast to the status test, where the reporting statistic had only a minor effect, the choice of reporting statistic has a major effect here. When the standard is low, over 40 percent of the schools are reporting as improving a sufficient amount, even though they haven't improved at all. When the standard is very high, few schools make the required amount of improvement (but as we shall see in the next study, few make the requirement improvement even when they truly do improve quite a bit).

Study 5—Computing the probability of classification errors for an improvement design: Part 2—When schools make substantial improvement

In the previous study, we showed the probability that a school would show a sufficient amount of improvement in its observed results even when its true score remained the same over years. In this study, we will do basically the same thing, except this time we will posit a substantial amount of improvement in the school's true score.

The program for this study is identical to the one used in the previous study except for two lines. The first change is the calculation of "INCREMENT" right after the school's true mean is determined. The amount of the increment is *twice* the difference between the school's true mean and 465, divided by 12. Then, when the score for each observed draw is selected for Year 2, INCREMENT is added to each student's score. Thus, the assumption behind this program is that every student will improve by the same amount, and that amount is twice what the school must improve to meet its scaled score improvement target. Obviously, many different amounts and patterns of change could have been chosen for study; greater or lesser amounts of change, and differential amounts (for example, lower students gaining more). Readers are encouraged to create whatever models are of interest to them; we chose this one for its simplicity. It forms a sufficiently complete model for readers to follow.

The results are shown in Table 5.

Table 5

Probability a School Will Be *Correctly* Judged on Improvement If It Makes Improvement Equal to *Twice* Required Amount

Statistic	N = 20	N = 50	N = 100
Mean scaled score	69	78	86
Index	71	81	89
Pass/Fail, pi=.50	67	75	83
Pass/Fail, pi=.25	67	78	85
Pass/Fail, pi=.10	66	73	80
Pass/Fail, pi=.90	39	27	20

Some observations:

1. Note the perfectly symmetrical results for the mean scaled score. The percentage of schools that are correctly identified as not improving when they haven't improved (from Table 4) is exactly equal to the percentage of schools that are correctly identified as improving when they have improved twice the required amount (If schools had improved exactly the amount required, 50 percent of them would have been correctly classified as having improved).
2. Note that once again, the standard chosen for a pass/fail reporting statistic greatly affects the percentage of schools correctly classified—but now, in the opposite direction. When no one improved, a high standard correctly failed most schools, while a low standard incorrectly passed a large percentage. But in this example, where schools have improved and therefore should be labeled as improved, most do when the standard is low, but most do not when the standard is high.

In order to determine the total percentage of correctly classified schools, one would need to model the patterns of improvements among schools. To continue on with our examples and to keep things fairly simple, suppose half the schools in the state didn't improve at all and half improved by twice their growth target. The percentage of schools correctly identified would be the average of the values in the cells in Tables 4 and 5. Table 6 provides those averages.

Table 6

Probability Schools Will Be *Correctly* Judged on Improvement
If Half Make No Improvement
and Half Make Twice the Required Amount of Improvement

Statistic	N = 20	N = 50	N = 100
Mean scaled score	69	78	86
Index	67	76	84
Pass/Fail, pi=.50	65	73	80
Pass/Fail, pi=.25	63	70	76
Pass/Fail, pi=.10	57	65	71
Pass/Fail, pi=.90	58	61	59

With Table 6, the results are quite what would be expected. The percentage of times schools are accurately classified is highest for the mean scaled score, followed by the index, with the pass/fail systems behind, and those with the most extreme percentages of students passing showing the least reliability. However, it is important to remember that this result is dependent on the model for the distribution of improvement we chose. If few schools in a state are improving, the highest accuracy comes with a system that fails the most students. If all are improving a substantial amount, several reporting statistics have almost equally high classification rates.

Study 6—Computing the probability of classification errors over two years: Part 1—When schools make no improvement

While all the information in the previous studies is interesting and provides statistics about the likelihood of accurately classifying a school in one particular year, the most severe consequences for schools under NCLB are not dependent on the classification of the school in one particular year, but whether the school is classified as not having made AYP two consecutive years. The purpose of this study is to determine the probability that schools will be identified as having not made AYP two years in row if they have, in fact, not made any improvement over that time.

The program for this study is a logical extension over the previous ones. However, rather than generating just two years' worth of data, as we did in Study 4, we now must generate three years' worth, then determine whether the school made sufficient improvement between Year 1 and Year 2, *or* between Year 2 and Year 3. Satisfactory improvement in either of those periods will permit a school to avoid the prescribed corrective actions.

The program for this study builds on the program developed for Study 4. Now, however, we must compute a growth target for the school after Year 2 as well as after Year 1, and then determine whether the school made the required amount of improvement. To simplify the generation of the reports, the program sets unsatisfactory performance as a 100 and satisfactory performance as a 0 each year. Multiplying the results for the two years yields a 0 if the performance in either year was satisfactory. That programming trick means that simply computing the mean of the results across the years provides the percentage of schools failing to make AYP both years.

The results are provided in Table 7.

Table 7

Probability a School Will Fail to Make Sufficient Improvement Two Years in a Row If It Makes No Improvement

Statistic	N = 20	N = 50	N = 100
Mean scaled score	42	59	74
Index	34	46	59
Pass/Fail, pi=.50	33	44	57
Pass/Fail, pi=.25	26	30	38
Pass/Fail, pi=.10	24	25	27
Pass/Fail, pi=.90	64	87	95

Again, remember that none of these schools made improvement, and therefore all of them should have been identified as not having made sufficient improvement two years in a row. For some of the cells, far fewer than half the schools are so identified. For some of the reporting statistics (note particularly pass/fail when there is a low standard), the majority of schools are not identified, even for the largest schools, despite the fact that none of them made improvement and therefore should be identified. Unless the standard for proficiency was extremely high, the mean scaled score was the statistic that most often correctly identified schools as not improving two consecutive years. But even with that statistic, a majority of small schools made AYP improvement at least one of the two years. Thus, even when schools are not improving, it is likely that a substantial portion of them will make AYP in at least one of the two years, and the less reliable statistic, the more likely that is to occur. Viewed in its simplest terms, this is a case where unreliability is working to the unfair advantage of non-improving schools. If, for example, a school has an unreliably high result for Year 2, it likely will make AYP in Year 2. If it has an unreliably low result for Year 2, it likely will not make AYP in Year 2, but it likely will in Year 3. So either an unusually high or low result for Year 2 means that a school will not be identified two consecutive years.

It is worthwhile comparing the results in this table to the ones in Table 4. As would be expected, far fewer schools are identified when the requirement is failure to improve two years in a row rather than just one. In fact, almost (but not quite for most of the reporting statistics) twice as many schools made AYP in one of the two years as made it in any one year.

Study 7—Computing the probability of classification errors over two years: Part 2—When schools make substantial improvement

Study 7 is a natural complement to Study 6, just as Study 5 was a natural complement to Study 4. In the first study, we examined what the likelihood that a school would make AYP even if it did not improve. In the second study, we determine the likelihood that a school will make AYP if it *does* improve.

For this study, we will presume that each student in the school improves each year the amount equal to one-twelfth the distance between the school's true score and 465. That is the same as the assumption made for Study 5, and just as was true for that study, it needs to be noted that this is just one of a wide range of possible improvement assumptions that could be made—researchers for various states will need to determine what improvement assumptions made sense for them to pursue.

The program for this study is much like the one for Study 6. The only difference is adding the established amount of improvement to each observed result for Year 2 and then twice that amount for Year 3.

Table 8 provides the results.

Table 8

Probability a School Will Fail to Make Sufficient Improvement Two Years in a Row If It Makes Improvement Equal to Required Amount Each Year

Statistic	N = 20	N = 50	N = 100
Mean scaled score	17	17	17
Index	12	11	9
Pass/Fail, pi=.50	13	13	12
Pass/Fail, pi=.25	13	8	6
Pass/Fail, pi=.10	13	10	7
Pass/Fail, pi=.90	45	70	83

As can be seen from Table 8, only a small percentage of the schools that truly improve fail to make AYP at least one of the two years, with the exception of the pass/fail system with a very high standard. One particularly noteworthy result is that the percentage of schools failing to make AYP both years is highest with the mean scaled score (with the exception of the pass/fail system with a very high standard), and the percentage does not vary with school size. For all the other reporting statistics, the probability that a school will make AYP at least one of the two years decreases as the school gets larger.

It is important to consider what the probability is of correct classification of a school. Suppose, for example, that a state decides to use a pass/fail reporting statistic with 50 percent of the students originally passing. Suppose further that there are 200 schools in the state, all with 20 students per school, and that 100 of the schools do not improve while the other 100 improve an amount each year equal to the one-twelfth of the distance between their starting true scaled score mean and 465. If that were the case, the following would be the expected results for the 200 schools, using the data from Tables 7 and 8.

Table 9

Results for Hypothetical State

True Status	Identification Status		Total
	Failed Both Years	Passed at least One Year	
Improved	14	86	100
Did Not Improve	33	67	100
Total	47	153	200

Even though half the schools improved and half did not, only 47 schools would be identified for improvement. But of those 47 schools, 14, or 30 percent of those identified, would be schools that truly had improved and therefore would be incorrectly identified as not having improved. Those are the results for just one reporting statistic for just one school size. Using those same assumptions (half the schools improve, half do not), Table 10 provides the percentage of schools that would be identified that had truly improved (and therefore should not have been identified).

Table 10

Proportion of Schools Failing to Make AYP Two Consecutive Years that Truly Improved

Statistic	N = 20	N = 50	N = 100
Mean scaled score	29	22	19
Index	28	19	13
Pass/Fail, pi=.50	30	23	17
Pass/Fail, pi=.25	33	21	14
Pass/Fail, pi=.10	35	29	21
Pass/Fail, pi=.90	41	44	47

Under these assumptions, the reporting statistic that works best is pass/fail with a fairly low standard (one that passes 75 percent of the students to start). Other statistics work well for smaller schools, but once schools get large, this is clearly the best choice. In addition to identifying the smallest percentage of truly improving schools, it also identifies a relatively small percentage of schools (improving or not). Out of 200 schools with 100 students per school, only 44 would be identified, and of that number, only 6 of the improving schools would be identified. Note that all this is relative; there certainly may be some policy-makers that would be surprised to find that a relatively excellent finding is that 14 percent of the schools identified are identified in error—especially given the rather extreme differences between "improving" and "non-improving" schools.

Study 8—Generating two sets of scores for each school, using two content areas

To this point, all the studies have involved generating data for just one content area. However, NCLB requires that schools have satisfactory scores in reading and mathematics, so it is necessary to examine the probabilities of being identified when two tests are used instead of one. Study 8 will start this new line of investigation. The next set of studies will parallel the ones done to this point, but include scores for two correlated content areas.

The key word in the previous sentence is "correlated." One cannot draw two uncorrelated samples for a school and pretend that those represent reading and mathematics scores. How a school does in one content area is highly correlated with how it does in another, so for the data to model actual school scores correctly, the scores for the two areas must be correlated.

This correlation must be employed in two ways. The true scores for schools in reading and math are highly correlated, as are the observed scores for individual students. So, once we have drawn the reading score for a school, we must make sure that the math score we choose for that school is correlated with the reading score. Then, once we have the true reading and math scores for a school, and we are choosing a sample of students within the school, whenever we choose a random student and determine that student's reading score, we must choose a correlated (but not perfectly correlated) math score for the student at the same time.

Operationally, this is done by determining the desired correlation between the two variables, drawing two random normal values, and then using that information to compute the two scores. The following SAS code applies these principles:

```
*Choose two normal random variables to create school true mean for reading
and math;
      Z1 = NORMAL(0);
      Z2 = NORMAL(0);
*Set the correlation of true means between reading and math at the school
level;
      SCHRMCORR = .95;
*Select a random school and compute its true mean reading and math scores;
      READ = SDTBAR * Z1 + 300;
      RSQ = SCHRMCORR**2;
      MATHZ = SCHRMCORR*Z1 + SQRT(1-RSQ) * Z2;
      MATH = SDTBAR * MATHZ + 300;
```

The first two lines of code select two random normal values, while the third line establishes the correlation between true scores of schools as .95.[5] From this point, the process for selecting values for reading and math is straightforward. Draw a score for reading as done in the previous studies. Then, to draw an appropriately correlated value for math, compute a regressed score for reading (multiply the z-score for reading by the correlation between the two content areas), then add in a random value that has the appropriate variance to bring the variance of math scores back to their desired amount (multiply the square root of 1 minus the correlation squared by the second random z-score). That provides you with a new variable that has a mean of 0, a variance of 1, and is correlated with the first variable. Then, simply transform that new z-score to have the desired mean and variance.

In the program supplied in the appendix, note that this process must be done twice—once to get the school mean true score and then again to get the individual student results. The logic of both applications is identical.

[5] We took the data from one state and computed the correlation between school means, by size of school, and the correlation between student observed scores in reading and math. The correlation between school means was a little below .9 for small schools and approaching .95 for the larger schools. The correlation between student observed scores was a little below .8. Values of .95 for the correlation of school mean true scores and .75 for student observed scores within school reproduced these values in the observed data. A state would want to determine its own correlations between content areas for schools and students and adjust the values used in our programs accordingly.

Table 11 provides the results of one application when we generated data for 100,000 schools.

Table 11

Correlations between School Mean Observed Scores and Student Observed Scores, Using Program for Study 8

Statistic	N = 20	N = 50	N = 100
Correlation between school means	.89	.92	.94
Correlation between student observed scores	.77	.78	.78

As can be seen from the data in Table 11, the program produced data that closely match the desired results.

Study 9—Computing the probability of classification errors for a status design, using two content areas

This study is the same as Study 3, except that a second content area has been added. Note that schools must pass a status test for both reading and math. When only one content area was being used, it was clear that 20 percent of the observed scores should fail, since the 20th percentile of school observed scores was chosen as the required status score. Now, since 20 percent will fail reading and 20 percent will fail math, the percentage that fail will be larger than 20 percent (but not 40 percent, since reading and math scores are correlated). In fact, it generally turned out that about 25 percent of the schools failed either one test or the other—although there were some very interesting exceptions.

The program for Study 9 is considerably longer than the program for Study 3, since we now have to keep track of two sets of scores rather than one. But the basic outline of the program is the same: compute true scores for the school and then determine all the population values, draw a sample of students around the true scores, and then compare the sample results to the population values. Note again that the status of a school is "OK" only if it is sufficiently high in *both* reading and math. Having an observed score in either content area that is below the 20th percentile means that the school as a whole does not make AYP.

Table 12 provides the results for this study. This table can be directly compared to Table 3, where schools were judged on just one content area. Again, the first value provided in each cell is the percentage of schools that should have made AYP (because their true scores in both reading and math were greater than the 20th percentile of observed scores) actually made AYP (because their observed scores in both reading and math were greater than those values), while the second value is the percentage of true fails that actually failed. Obviously, the percentage of schools correctly judged as passing is lower than in Table 3 (since fewer schools now are passing), while the percentage of schools correctly judged as failing is higher (again, since now more schools are failing).

Table 12

Probability a School Will Be *Correctly* Judged on Status on Two Content Areas

Statistic	N = 20	N = 50	N = 100
Mean scaled score	71+15	73+19	72+21
Index	70+15	71+20	72+21
Pass/Fail, pi=.50	64+16	71+19	72+21
Pass/Fail, pi=.25	69+15	69+20	72+21
Pass/Fail, pi=.90	64+12	68+16	69+19

Table 12A compares the sum of the two values (for the total percentage of correct judgments) when schools must pass only one content area versus when they must pass two. Most of the results are not a surprise. The probability of correct classification increases when schools are larger, the probability of correct classification is higher when using a more reliable reporting statistic (although, again, not much higher), and the probability of correct classification with one test is higher than with two (although not much higher in this case as well).

Table 12A

Percentages of Times Schools Are *Correctly* Judged on Status, Comparing One Content Area with Two

Statistic	N = 20		N = 50		N = 100	
	One	Two	One	Two	One	Two
Mean scaled score	88	86	93	92	95	93
Index	88	85	92	91	94	93
Pass/Fail, pi=.50	84	80	91	90	94	93
Pass/Fail, pi=.25	86	84	90	89	93	93
Pass/Fail, pi=.90	80	76	86	84	89	88

What is even more interesting is simply the number of schools identified on the basis of observed scores than should be identified based on true scores. Again, this happens because the variance of observed scores is greater than the variance of true scores; since we are trying to identify schools in one tail of the distribution, we will identify more schools on the basis of observed scores than true scores. Table 12B provides the percentage of schools identified on the basis of true scores and the percentage identified on the basis of observed scores.

The gap between the two percentages decreases as the school size increases. When there are 100 students per school, the differences are quite small. There is one result that seems particularly aberrant (so much so that we rechecked the finding several times), and that is the one for Pass/Fail, pi = .50, when N = 20. For that cell, the percentage of schools identified on the basis of true scores is considerably smaller than the percentage based on observed scores. Since Pass/Fail with pi = .50 is likely to be reporting statistic of choice for many states, that result warrants closer examination.

Table 12B

Percentages of Schools Passing AYP on the Basis of True Scores versus The Percentages Passing on the Basis of Observed Scores

Statistic	N = 20		N = 50		N = 100	
	Observed	True	Observed	True	Observed	True
Mean scaled score	81	75	78	76	76	75
Index	80	75	77	74	77	75
Pass/Fail, pi=.50	81	67	78	75	77	75
Pass/Fail, pi=.25	79	74	77	73	76	75
Pass/Fail, pi=.90	83	68	78	74	77	74

Table 12C provides the probabilities in more detail, showing the percentage of schools in each of the four possible cells. A total of 80 percent of the judgments are correct; 64 percent of those who should pass, do; and 16 percent of those who should not pass, do not. But notice that of every 33 who fail, 17 have true scores higher than the required amount and therefore should pass. While this finding is unique to this particular cell, it should be of concern, particularly since this reporting statistic is likely to be used by many states: *In this cell, a majority of schools identified as failing were schools that truly* should *have been identified as passing.* That seems to be an unacceptably high error rate.

Table 12C

Probabilities of Observations, Given *N* = 20 and Reporting Statistic of *Pass/Fail, pi = .50,* Using Two Content Areas

Observed Status	True Status		Total
	% Passing > 36 on Both Tests *"Should Pass"*	% Passing < 36 on at Least One Test *"Should Fail"*	
% Passing > 36 on Both Tests *"Does Pass"*	64	3	67
% Passing < 36 on at Least One Test *"Does Fail"*	17	16	33
Total	81	19	100

Study 10—Computing the probability of classification errors for an improvement design, using two content areas: Part 1—When schools make no progress

This study is the companion to Study 4, except now we compute two scores for each school rather than one. To make sufficient improvement, a school has to make improvement in both content areas. Needless to say, the probability that this will happen by chance is less than it is when just one content area is tested. The unknown going into this study, however, is how much that probability will decrease.

The program for doing this study is a simple extension of the one used in Study 4. Again, the only real change is that we are computing two scores for each student, using the same logic and programming as we did in the two studies before this one.

The results are provided in Table 13. Table 13 includes the results from Table 4, where just one content area was included, alongside the new results, using two content areas.

Table 13

**Percentages of Times Schools Are *Correctly* Judged on Improvement,
Comparing One Content Area with Two,
When There Is No Improvement**

Statistic	N = 20		N = 50		N = 100	
	One Content Area	Two Content Areas	One Content Area	Two Content Areas	One Content Area	Two Content Areas
Mean scaled score	69	79	78	86	86	92
Index	63	76	70	82	78	88
Pass/Fail, pi=.50	63	78	70	84	77	89
Pass/Fail, pi=.25	58	74	61	77	66	81
Pass/Fail, pi=.10	55	72	57	74	59	76
Pass/Fail, pi=.90	82	92	94	98	98	99

Given that the model is for no improvement on the part of any school, no school should pass AYP on the basis of improvement. The percentage that do on the basis of two tests is considerably less than on the basis of one test; in fact, for some of the cells, the percentage that are incorrectly identified as passing is reduced by half or more.

Study 11—Computing the probability of classification errors for an improvement design, using two content areas: Part 2—When schools make substantial progress

Study 11 is the companion to both Study 5 (where just one content area was included) and Study 10 (where schools made no improvement). In Study 11, we model what happens when schools make the degree of improvement modeled for Study 5, but do it in both content areas at the same time.

Again, the program for doing this study is provided in the appendix, but it is a straightforward extension of previous programs. This program is simply a duplicate of the one used for Study 10, but now each student's score is incremented by a certain amount, depending upon the true score for the school in reading and math (the increments are determined separately for each content area). As was true for Study 5, a large number of possible patterns for improvement could be modeled; we just chose one to provide a preliminary look. States could posit different amounts from these and study the affect of those improvements on the probabilities their schools will be identified or not.

The results are provided in Table 14. As we did with Table 13, we have reproduced the results for one content area in this table so that the results for two content areas can be directly compared to it.

As would be expected, the percentage of schools that are correctly judged on the basis of two tests is less than the percentage on one. Most of these changes are almost identical to the increase in the percentage of correctly identified schools when there was no improvement (Table 13). Thus, adding a second content area appears to have a fairly uniform effect—there are similar decreases in the percentage of schools identified whether they have truly make improvement or not. Again, looking at the row for "Pass/Fail, pi=.50," since that is a reporting statistic likely to be used by many states, note that even though these schools have all made tremendous gains, over a quarter of them will be incorrectly judged as not having made AYP even when there are as many as 100 students in the school.

Table 14

Percentage of Schools *Correctly* Judged on Improvement When They Make Improvement Equal to *Twice* Required Amount in Two Content Areas, Comparing the Results for One Content Area with the Results for Two

Statistic	N = 20		N = 50		N = 100	
	One Content Area	Two Content Areas	One Content Area	Two Content Areas	One Content Area	Two Content Areas
Mean scaled score	69	58	78	69	86	79
Index	71	60	81	72	89	82
Pass/Fail, pi=.50	67	53	75	62	83	73
Pass/Fail, pi=.25	67	52	78	65	85	76
Pass/Fail, pi=.10	66	51	73	58	80	68
Pass/Fail, pi=.90	39	24	27	14	20	10

Study 12—Computing the probability of classification errors over two years, using two content areas: Part 1—When schools make no progress

Study 12 is a direct extension of Study 6, adding the second content area. In this study, we are determining the probability that a school will fail the improvement test for AYP two years in a row if it makes no real improvement.

The program for doing this is, once again, a direct extension of the program used for Study 6, adding the process for generating two scores for each student that was demonstrated in Study 8. The program is long, simply because there is a lot to keep track of, but it is not complex. We simply compute scores for schools in three years and determine whether the results in Year 2 are sufficiently higher than those in Year 1 to make AYP, and the same thing from Year 2 to Year 3. A school has to meet both improvement targets in one year to make AYP for that year.

Table 15 provides the results. Again, we have incorporated the results from Table 7 (when just one content area was included) into this table so that the effects of adding a second content area can be easily seen.

Once again, the first obvious result is that fewer schools make AYP when a second test is added. In this case, those percentages increase rather dramatically. Note once again that none of these schools is supposed to make AYP on the basis of improvement—none improved. Again, an obvious result is that the larger the school, the less likely it is to make AYP in either of the two years on the basis of chance alone. Note again the dramatic difference in the percentage of schools identified in a pass/fail system depending on the difficulty of the standard—the higher the standard, the more likely it is that schools will fail to make AYP. Finally, note the dramatic decrease in the number of schools identified in comparison to the number identified on the basis of one year's results only (Table 13).

Table 15

Percentage of Schools Failing to Make Sufficient Improvement in Observed Scores in Two Content Areas Two Years in a Row When They Make No Real Improvement, Comparing the Results for One Content Area with the Results for Two

Statistic	N = 20		N = 50		N = 100	
	One Content Area	Two Content Areas	One Content Area	Two Content Areas	One Content Area	Two Content Areas
Mean scaled score	42	60	59	74	74	85
Index	34	55	46	66	59	77
Pass/Fail, pi=.50	33	57	44	68	57	78
Pass/Fail, pi=.25	26	51	30	56	38	63
Pass/Fail, pi=.10	24	48	25	52	27	55
Pass/Fail, pi=.90	64	84	87	96	95	99

Study 13—Computing the probability of classification errors over two years, using two content areas: Part 2—When schools make substantial progress

Study 13 is the natural complement to Study 12—computing the probability that schools will be judged as having made sufficient improvement on both content areas if they do, in fact, make substantial progress. Again, we modeled substantial improvement in both content areas the same way we did in Study 11.

The computer program to do this study is provided in the Appendix. Again, while the program is long and complex in order to keep track of all the results for schools in each of the three years, the basic logic of it is straightforward. A sample is drawn each year; in the second and third years, the amount of improvement made by the school (equal to its target) is added to every student's score in both content areas. We then simply compute the probability that a school's improvement is equal to or greater than its required improvement target under NCLB.

The results are provided in Table 16. Given what we have observed to this point, there are no surprises. The probability that schools will fail the improvement portion of AYP is greater when they have to pass two tests rather than one. The probability that they will fail is less if they make improvement than if they make none. The probability that they will pass increases if they are a large school, but not by much—at least, not by nearly as much as the probability increases that they will be identified if they have not made improvement. So NCLB is a system that clearly favors small schools; if they have not made improvement, the odds are far smaller that they will be identified than they will if they are a large school—but if they have made improvement, their odds are not hurt that much by being a small school.

Table 16

Percentage of Schools Failing to Make Sufficient Improvement in Observed Scores in Two Content Areas Two Years in a Row When They Make Improvement Equal to Required Amount Each Year, Comparing the Results for One Content Area with the Results for Two

Statistic	N = 20		N = 50		N = 100	
	One Content Area	Two Content Areas	One Content Area	Two Content Areas	One Content Area	Two Content Areas
Mean scaled score	17	31	17	31	17	31
Index	12	28	11	25	9	22
Pass/Fail, pi=.50	13	33	13	32	12	30
Pass/Fail, pi=.25	13	32	8	24	6	20
Pass/Fail, pi=.10	13	32	10	29	7	23
Pass/Fail, pi=.90	45	70	70	87	83	93

Table 17 provides an overall perspective on this issue. Suppose 100 of the 200 schools in a state truly improve by the required amount while the other 100 have the same true scores year after year. Suppose further that the state chose "pass/fail" as its reporting statistic, and set the passing score so that 50 percent of the students passed the first year. Suppose further that all the schools in the state consisted of 50 students. The results it would observe on improvement only would be those in Table 17.

Table 17

Probabilities of Observations, Given *N* = 50 and Reporting Statistic of *Pass/Fail, pi = .50,* Using Two Content Areas

True Status	Identification Status		Total
	Failed Both Years	Passed at least One Year	
Improved	32	68	100
Did Not Improve	68	32	100
Total	100	100	100

Half the schools would be judged as improving while half would be judged as not improving. That's good news, since, in reality, half did improve and half did not. But of the 100 identified as "not improving," only 68 did not truly improve; almost a third of the schools that would be put in that category would be mislabeled. They would be schools that truly improved, but due to the luck of the draw of their student sample each year, did not show sufficient improvement in the scores of those students.

This study is the last one in which we use the Monte Carlo method for studying reliability. The remaining area of concern is the impact of requiring all subgroups to meet the improvement standards for a school to show AYP. Modeling the performance of subgroups with the Monte Carlo method is too complex, so the remaining studies in this report will use the "random draws with replacement" method. The first several studies using that method will duplicate several of the studies already concluded, and show that the results are very much the same (lending credence to both approaches). Then, we will extend that method to study the impact of subgroups on the probabilities of correct classification.

Study 14: Drawing Random Samples with Replacement

The remaining studies presented in this paper were conducted using the random draws with replacement method. As discussed in Chapter 2, one appeal of this method is its simplicity. The data used in the studies described here were generated by drawing random samples of data from a state assessment student data file. The data were drawn from a single elementary grade in which students were tested in Reading and Mathematics. Before presenting the studies, we will describe the characteristics of the original data. As shown in Table 18A, there were four major subgroups of students represented in the state that will be include in the studies presented in this paper: economically disadvantaged, students with disabilities, African-American/Black, and White. Statewide, the percentage of English language learners and students from other racial/ethnic groups were too small to include in these studies. Table 18B presents performance data in terms of scaled scores and performance levels. Note that in many respects, the data presented here are quite similar to the data generated in the Monte Carlo studies described previously. The major difference is that the standard deviation of the student scaled scores is close to 50 rather than 100. There are also five performance levels included in this example as compared to four levels in the Monte Carlo studies. However, the percentage of students at the highest level (Distinguished) is very small, and the distribution of students across the performance levels is consistent with the distribution of performance in the Monte Carlo studies.

Table 18A
Demographics

Number of Students: 63,000
Number of Schools: 911
Subgroups Economically Disadvantaged: 62% Students with Disabilities: 10% African-American/Black: 51% White: 46%

Table 18B
Performance Data

	Reading	Mathematics
Scaled Score Range	100 - 500	100 - 500
Student Mean (sd)	306.97 (51.23)	315.16 (51.85)
School Mean (sd)	304.66 (30.44)	311.49 (32.61)
Performance Levels		
5 - Distinguished	1.08%	1.67%
4 - Advanced	14.34%	10.80%
3 - Proficient	44.11%	40.83%
2 - Basic	24.13%	23.42%
1 - Below Basic	16.34%	23.27%
Student-level correlation between reading and mathematics		.78
School-level correlation between reading and mathematics		.91

Consistent with the Monte Carlo studies, the statistics examined in these studies will be scaled scores, an index, and percentage of proficient students based on cut scores for passing set at pi=10, pi=25, pi=50, and pi=90. Table 18C provides the scaled score that defines each of the pass/fail cut scores and the 20th percentile status cut for each of the reporting statistics.

Table 18C
Cut Scores for Reporting Statistics

	Reading		Mathematics	
	Cut Score	20th percentile	Cut Score	20th percentile
Scaled Score		286		293
Index		2.20		1.97
Pass/Fail, pi = .10	242	83%	250	83%
Pass/Fail, pi = .25	278	62%	284	62%
Pass/Fail, pi = .50	312	32%	318	29%
Pass/Fail, pi = .90	364	2%	378	02%

The purpose of this study is to demonstrate the procedures used to draw the random samples with replacement from the original pool of data and show that each randomly drawn sample has the same characteristics as the original sample. For the studies described in this paper, student samples were drawn at the state level. That is, random samples of 63,000 students were drawn with replacement from the statewide pool of students. As a results, the number of students per school and district varies from sample to sample. In fact, the number of schools is also allowed to vary from sample to sample because students from some very small schools may not be drawn at all in a particular sample.

The samples were drawn in a two-step process. The first step was to determine which records in the original sample to draw for each sample. This was accomplished by generating a random set of 63,000 numbers between 1 and 63,000. As shown below, this was accomplished using a DO loop to generate the record numbers to be drawn from the original file. Note that the same record number may be selected several times, while some record numbers will not be selected at all.

```
do i= 1 to 63000;
  record = round((uniform(0)*63000) + .5);
output;
end;
```

The second step was to create the sample file, by pulling the selected records from the original data file. This was accomplished by merging the selection file containing the record numbers drawn above with the original data file. Records repeated multiple times in the selection file were drawn multiple times. Record numbers not included in the selection file were not drawn. The result of this process is three student data files, each containing 63,000 records drawn from the original student data file. The program used to generate three random sets of data is included in the appendix. The process can be expanded easily to generate any number of random samples.

How do the three sample files compare to the original data? As shown in Table 19A below, each sample includes approximately 63 percent of the original 63,000 student records. In each sample approximately 23,000 of the original student records are included a single time; 12,000 records are included two times; 4,000 records are included three times; 1,000 records are included four times; and diminishing numbers of records are included up to eight times.

Table 19A

Characteristics of Three Random Samples Drawn with Replacement:
Number of Records Drawn from the Original File

Number of Times Drawn	Sample A	Sample B	Sample C
1	23,051	23,197	23,207
2	11,701	11,579	11,624
3	3,829	3,807	3,831
4	970	977	972
5	201	204	192
6	28	40	28
7	2	8	4
8	0	0	1
Total Unique Records Drawn	39,871	39,812	39,859

In terms of student performance statewide, there is little difference among the three samples. As shown in Table 19B, student performance across the samples is stable in terms of scaled scores and performance levels. At the school level, however, there is slightly more variation among the samples. Particularly, the number of schools in the three samples ranges from 894 to 900, compared to the 911 schools in the original sample. As mentioned previously, the schools lost in the sampling process are schools with very few students. In this example, all of the schools missing from one or more of the samples are schools with fewer than 10 students at the tested grade.

Table 19B

Characteristics of Three Random Samples Drawn with Replacement
Mean Scaled Scores and Performance Level Distributions

Reading	Original Data	Sample A	Sample B	Sample C
Scaled Score				
Range	100 - 500	100 - 500	100 - 500	100 – 500
Student Mean (sd)	306.97 (51.23)	306.87 (51.15)	307.23 (50.98)	306.96 (51.24)
School Mean (sd)	304.66 (30.44)	305.22 (29.49)	305.36 (30.38)	305.84 (29.75)
Performance Levels				
5 - Distinguished	1.08%	1.05%	1.04%	1.04%
4 - Advanced	14.34%	14.28%	14.44%	14.44%
3 - Proficient	44.11%	44.13%	44.16%	44.03%
2 - Basic	24.13%	24.18%	23.99%	24.30%
1 - Below Basic	16.34%	16.36%	16.37%	16.18%
Mathematics	Original Data	Sample A	Sample B	Sample C
Scaled Score				
Range	100 - 500	100 - 500	100 - 500	100 - 500
Student Mean (sd)	315.16 (51.85)	315.01 (51.96)	315.44 (51.34)	315.39 (51.69)
School Mean (sd)	311.49 (32.61)	311.99 (32.73)	312.66 (32.50)	312.61 (32.67)
Performance Levels				
5 - Distinguished	1.67%	1.60%	1.62%	1.76%
4 - Advanced	10.80%	10.95%	10.69%	10.90%
3 - Proficient	40.83%	40.60%	41.15%	40.80%
2 - Basic	23.42%	23.63%	23.64%	23.45%
1 - Below Basic	23.27%	23.22%	22.89%	23.09%
Number of Schools	911	900	894	896
Student-level Correlation, Reading and Mathematics	.776	.775	.774	.772
School-level Correlation, Reading and Mathematics	.906	.864	.902	.876

Study 15 – Replicating Monte Carlo Studies

The first set of studies conducted with the random draws with replacement method replicated several of the conducted using the Monte Carlo method. The purpose of the replication was twofold. First, as explained in Chapter 2, because it is possible to make errors in logic along the way that will lead to errors in computing, we recommend that results be obtained by at least two methods and compared. Second, before proceeding to studies analyzing the impact of adding subgroups, we wished to establish a logical thread between the results obtained from the Monte Carlo studies and subsequent results.

The Monte Carlo studies replicated using the random draws method are as follows:

- Study 6 – Computing the probability of classification errors over two years: Part 1 – When schools make no improvement (Table 7)
- Study 7 – Computing the probability of classification errors over two years: Part 2 – When schools make substantial improvement (Table 8)

- Study 12 – Computing the probability of classification errors over two years, using two content areas: Part 1 – When schools make no progress (Table 15)
- Study 13 – Computing the probability of classification errors over two years, using two content areas: Part 2 – When schools make substantial progress (Table 16)

The design, procedures, and results for those studies are discussed fully earlier in this chapter. Rather than repeat that discussion here, to allow comparisons, we will simply provide tables with results obtained from the random draws method for the replicated studies. Additionally, the decision rules and programming used with the random draws method to determine whether schools have met the AYP improvement goal are the same as those used for the Monte Carlo studies.

Note that the results presented for these studies and the remaining studies conducted using the random draws with replacement method are the combined results from 25 replications of each study. That is, each study was repeated twenty-five times with different sets of three randomly drawn student samples. Results across those replications were aggregated to produce the data presented in this paper. The purpose of repeating the studies was a) to determine the consistency of results using the method and b) to present a stable set of results. Results of status and improvement analyses for overall schools and large subgroups (N>30) were very consistent across studies and varied no more than 1-3 percentage points after 5-10 replications. Results of analyses involving small subgroups (n<30) stabilized after 15 replications.

Table 20

Probability a School Will Fail to Make Sufficient Improvement Two Years in a Row
If It Makes No Improvement
(Complements Table 7)

Reading	N=10-30	N=31-69	N=70-130
Mean scaled score	48	56	67
Index	35	41	50
Pass/Fail, pi=.50	38	45	56
Pass/Fail, pi=.25	23	30	38
Pass/Fail, pi=.10	13	20	26
Pass/Fail, pi=.90	76	88	94
Mathematics	N=10-30	N=31-69	N=70-130
Mean scaled score	49	62	72
Index	35	43	53
Pass/Fail, pi=.50	37	45	54
Pass/Fail, pi=.25	23	28	37
Pass/Fail, pi=.10	14	19	24
Pass/Fail, pi=.90	79	88	94

Table 21

Probability a School Will Fail to Make Sufficient Improvement Two Years in a Row If it Makes Improvement Equal to Required Amount Each Year (Complements Table 8)

Reading	N=10-30	N=31-69	N=70-130
Mean scaled score	19	18	18
Index	12	9	8
Pass/Fail, pi=.50	17	15	15
Pass/Fail, pi=.25	12	12	10
Pass/Fail, pi=.10	8	10	11
Pass/Fail, pi=.90	57	64	73
Mathematics	N=10-30	N=31-69	N=70-130
Mean scaled score	18	16	18
Index	11	8	7
Pass/Fail, pi=.50	16	12	13
Pass/Fail, pi=.25	8	8	7
Pass/Fail, pi=.10	5	8	8
Pass/Fail, pi=.90	62	70	76

Table 22

Percentage of Schools Failing to Make Sufficient Improvement in Observed Scores in Two Content Areas Two Years in a Row When They Make No Real Improvement (Complements Table 15)

Reading or Mathematics	N=10-30	N=31-69	N=70-130
Mean scaled score	68	75	82
Index	59	64	70
Pass/Fail, pi=.50	62	68	77
Pass/Fail, pi=.25	45	54	61
Pass/Fail, pi=.10	28	40	49
Pass/Fail, pi=.90	91	97	98

Table 23

Percentage of Schools Failing to Make Sufficient Improvement in Observed Scores in Two Content Areas Two Years in a Row When They Make Improvement Equal to Required Amount Each Year (Complements Table 16)

Reading or Mathematics	N=10-30	N=31-69	N=70-130
Mean scaled score	34	32	33
Index	27	22	18
Pass/Fail, pi=.50	38	33	33
Pass/Fail, pi=.25	25	26	25
Pass/Fail, pi=.10	15	23	25
Pass/Fail, pi=.90	80	86	90

Overall, the results presented in Tables 20-23 are very comparable to the results of the Monte Carlo studies. With few exceptions, the results obtained from the two methods are equivalent in terms of estimates of reliability/misclassifications and trends across various size schools. The largest differences between results from the two methods are found for small schools and/or extreme cut scores. In Tables 8 & 22, for example, there is a 17 percentage point difference between the percentage of small schools identified with a cut score of pi=.90. In both studies, however, the percentage of small schools identified at pi=.90 was by far the largest among all of the reporting statistics. Similarly, as shown in Tables 15 & 22 and Tables 16 & 23, there are differences between methods in results for the extreme cut score pi=.10.

Study 16 – Computing the probability of classification errors over two years for two content areas, including subgroups: Part 1 – When schools make no improvement

A core component of the NCLB definition of AYP is the requirement that a state's accountability program includes "separate measurable annual objectives for continuous and substantial improvement" for each of the following subgroups of students within a school:

- economically disadvantaged students;
- students from major racial and ethnic groups;
- students with disabilities; and
- students with limited English proficiency.

As currently interpreted, this will require states to perform the same set of analyses described throughout this paper for each major subgroup within a school. A school in which any subgroup fails to make AYP in a given year will be classified as failing to make sufficient improvement for that year. As shown in Table 24, the number of tests that a school must pass in order to successfully make AYP in a given year grows steadily as the number of subgroups within the school increases. It would not be uncommon for a given school to have six or more subgroups (e.g., Economically Disadvantaged, Students with Disabilities, African-American/Black, Asian, Hispanic, White, Limited English Proficient).

Table 24

Number of AYP Tests A School Must Pass Each Year
Based on the Number of Subgroups within the School

Number of Subgroups	Number of AYP tests in Reading	Number of AYP tests in Mathematics	Total Number of AYP tests
0	1	1	2
1	2	2	4
2	3	3	6
3	4	4	8
4	5	5	10
5	6	6	12
6	7	7	14
7	8	8	16

This study will extend the previous studies to include measures of improvement for the four subgroups of students identified in this state: economically disadvantaged students, students with disabilities, African-American/Black students, and White students. As in previous studies, results will be provided first for a single content area, reading and mathematics, and then for the two content areas combined. This study will include subgroups of any size in the computations. Subsequent studies will examine the impact on classification errors of raising the minimum subgroup size.

The programming for this study was conducted in two phases. In the first phase, the previously described programs were rerun separately for each subgroup of students. That produced a series of school-level data files containing a pass/fail decision in each content area for each subgroup. There are more efficient approaches that would analyze all subgroups simultaneously, but the complexity of the already complex and lengthy program would be increased significantly. In the second phase of the analysis, the individual data files for each subgroup were combined and an AYP classification was made for each school based on all of the AYP tests. A school that failed any of the AYP tests in a given year was classified as not making AYP for that year. We then computed the probability that a school failed to make AYP two years in a row. The program used to combine the data files and assign the AYP classification is contained in the appendix.

Table 25A

Probability a School or One of Its Subgroups Will Fail to Make Sufficient Improvement Two Years in a Row If It Makes No Improvement

Reading	N=10-30	N=31-69	N=70-130
Mean scaled score	86	93	95
Index	80	90	92
Pass/Fail, pi=.50	76	88	96
Pass/Fail, pi=.25	56	77	82
Pass/Fail, pi=.10	34	50	63
Pass/Fail, pi=.90	97	99	99
Mathematics	**N=10-30**	**N=31-69**	**N=70-130**
Mean scaled score	87	94	98
Index	83	91	92
Pass/Fail, pi=.50	74	88	93
Pass/Fail, pi=.25	58	74	80
Pass/Fail, pi=.10	35	47	56
Pass/Fail, pi=.90	96	99	99

Table 25B

Percentage of Schools Failing to Make Sufficient Improvement in Observed Scores in Two Content Areas Two Years in a Row When the School and All Subgroups Make No Real Improvement

Reading or Mathematics	N=10-30	N=31-69	N=70-130
Mean scaled score	95	99	99
Index	93	97	98
Pass/Fail, pi=.50	90	97	98
Pass/Fail, pi=.25	75	91	93
Pass/Fail, pi=.10	51	71	80
Pass/Fail, pi=.90	99	99	99

The results of the study are provided in Tables 25A and 25B. The trend found in previous studies continues here as we move from one content area to two content areas to overall school plus subgroups. The probability that school will fail the improvement portion of AYP increases as the number of hurdles that must be cleared in terms of AYP tests increases. As shown in Table 25B, if a school has not made improvement, the probability that at least one of its major subgroups will fail AYP each year is nearly 1.00. Only with a very low cut score (i.e., pi=10) and a very small school (which probably does not have more than one subgroup) is there close to an even chance that a school that has not made improvement will pass AYP.

Study 17 – Computing the probability of classification errors over two years for two content areas, including subgroups: Part 2 – When schools make substantial improvement

Study 17 repeats the analyses conducted in Study 16 for the case when schools and the subgroups within the schools do make substantial improvement. In previous studies, we set out to model the situation where students made enough improvement (in terms of mean scaled score) for the school to meet its AYP goal. Therefore, we modeled improvement by adding a constant to each student's score in Year 2 and Year 3. With subgroups, however, the process became substantially less straightforward than it was for the overall school. Consistent with the process used for the overall school, we wished to model the situation where the students in each subgroup made enough improvement (in terms of mean scaled score) for the subgroup to meet its AYP goal.

Recall that the amount of improvement was determined by computing the increase in scaled score necessary to close the gap between the current school mean and the 90th percentile scaled score by 2014. In studies conducted with the random draws method, the 90th percentile scaled scores were 380 in reading and 400 in mathematics. Under this model, the level of improvement for each subgroup would be based on the difference between its mean scaled score and the goal. An example of the improvement computations for an individual school scoring near the state average is provided in Table 26.

Table 26

Annual Improvement Required for Each Subgroup to Meet Its Scaled Score Improvement Goal

Subgroup	Current Mean Scaled Score	2014 Goal	Gap	Annual Improvement Goal (Gap divided by 12)
Economically Disadvantaged	304	400	96	8.00
Students with Disabilities	270	400	130	10.83
African-American/Black	295	400	105	8.75
White	336	400	64	5.33
Overall School	315	400	85	7.08

As shown in Table 26, each subgroup has an annual improvement increment that is different from the overall school. Of course, the subgroups are not independent from the overall school, and more importantly, with the possible exception of the racial/ethnic subgroups, the subgroups are not independent of each other. That is, a single student might be a member of multiple subgroups. Therefore, the annual improvement increment for an individual student's score would vary as a function of the student's multiple group memberships. For the purpose of estimating the reliability of the accountability system, we decided not to address that complexity in the system. As previously described, in these studies, each subgroup was analyzed separately and treated independently. In each subgroup analysis, student scores were increased by increment equal to the annual improvement goal for that subgroup.

Table 27A

Probability a School and One of its Subgroups Will Fail to Make Sufficient Improvement Two Years in a Row When They Make Improvement Equal to the Required Amount

Reading	N=10-30	N=31-69	N=70-130
Mean scaled score	64	69	71
Index	65	67	65
Pass/Fail, pi=.50	64	75	78
Pass/Fail, pi=.25	44	58	62
Pass/Fail, pi=.10	24	36	42
Pass/Fail, pi=.90	92	98	99
Mathematics	N=10-30	N=31-69	N=70-130
Mean scaled score	60	69	72
Index	62	66	64
Pass/Fail, pi=.50	65	72	73
Pass/Fail, pi=.25	42	46	51
Pass/Fail, pi=.10	21	27	34
Pass/Fail, pi=.90	92	98	99

Table 27B

Probability a School and All Subgroups Will Fail to Make Sufficient Improvement Two Years in a Row When They Make Improvement Equal to the Required Amount

Reading or Mathematics	N=10-30	N=31-69	N=70-130
Mean scaled score	80	85	87
Index	80	85	85
Pass/Fail, pi=50%	81	91	92
Pass/Fail, pi=25%	61	75	80
Pass/Fail, pi=10%	35	52	62
Pass/Fail, pi=90%	97	99	99

The results of the study provided in Tables 27A and 27B look quite similar to the results of the previous study examining performance under the no improvement condition. When subgroups of any size are included with two content areas and a reasonable pass/fail cut is identified, approximately 90 percent of schools were identified as failing to make AYP two years in a row even under our model in which each subgroup made significant progress in average scaled score two years in a row.

One trend that begins to emerge in this study is the relationship between school size and the probability of being identified as failing to make AYP. A review of each of the reporting statistics in Table 27B reveals that the probability of being identified as failing to make AYP when the school has improved increases as school size increases, suggesting that the system is fairer for small schools than large schools, although hardly fair for anyone. These results are consistent with previous research conducted by Gong (2001). To a large extent, the effect appears to be due to the increased likelihood that in large schools there will be multiple small subgroups, and therefore multiple conjunctive pass/fail tests, included in the analysis.

Study 18 – Computing the probability of classification errors over two years for two content areas, including subgroups: Part 3 – When the minimum subgroup size is increased

Study 16 and Study 17 introduced the idea of including subgroups in AYP analysis and identified several of the conceptual and programming concerns that must be considered in designing and evaluating an accountability system. Those studies, however, placed no restrictions on the size of the subgroups included in the analysis. In attempting to design valid and reliable accountability systems that meet the requirements of NCLB, states will identify a minimum size for subgroups.

Study 18 replicates studies 16 and 17 for subgroups of 10 students or more and subgroups more than 30 students. Results are provided only for classifications across two content areas. The programming for this study is identical to previous studies with the exception that subgroups that did not meet the minimum size criterion were excluded.

Tables 28A – 28D provide the results.

Table 28A

Percentage of Schools Failing to Make Sufficient Improvement in Observed Scores in Two Content Areas Two Years in a Row
When the School and All Subgroups Make No Real Improvement
Minimum Subgroup Size of 10

Reading or Mathematics	N=10-30	N=31-69	N=70-130
Mean scaled score	78	91	97
Index	70	85	93
Pass/Fail, pi=.50	73	87	95
Pass/Fail, pi=.25	53	75	88
Pass/Fail, pi=.10	31	53	72
Pass/Fail, pi=.90	94	99	99

Table 28B

Percentage of Schools Failing to Make Sufficient Improvement in Observed Scores in Two Content Areas Two Years in a Row
When the School and All Subgroups Make No Real Improvement
Minimum Subgroup Size of 31

Reading or Mathematics	**N=10-30**	**N=31-69**	**N=70-130**
Mean scaled score	68	82	93
Index	59	73	86
Pass/Fail, pi=.50	62	76	89
Pass/Fail, pi=.25	45	63	80
Pass/Fail, pi=.10	28	47	66
Pass/Fail, pi=.90	91	97	99

Table 28C

Probability a School and All Subgroups Will Fail to Make Sufficient Improvement Two Years in a Row When They Make Improvement Equal to the Required Amount Minimum Subgroup Size of 10

Reading or Mathematics	N=10-30	N=31-69	N=70-130
Mean scaled score	48	63	74
Index	44	54	64
Pass/Fail, pi=.50	52	65	77
Pass/Fail, pi=.25	34	52	64
Pass/Fail, pi=.10	18	38	52
Pass/Fail, pi=.90	87	96	99

Table 28D

Probability a School and All Subgroups Will Fail to Make Sufficient Improvement Two Years in a Row When They Make Improvement Equal to the Required Amount Minimum Subgroup Size of 31

Reading or Mathematics	N=10-30	N=31-69	N=70-130
Scaled Score	34	46	59
Index	27	34	44
Pass/Fail, pi=.50	38	45	61
Pass/Fail, pi=.25	25	37	49
Pass/Fail, pi=.10	15	31	43
Pass/Fail, pi=.90	80	90	96

As can be seen from Tables 28A – 28D, placing a minimum size on the subgroups to some extent mitigates the impact of adding subgroups to the accountability design. However, under the improvement condition, a large percentage of schools are still identified as failing to make AYP two years in a row. Table 29 provides a comparison for schools of 70-130 students of the results under the improvement condition for the overall school and various minimum size subgroups.

Table 29

Probability a School of 70-130 Students Will Fail to Make Sufficient Improvement Two Years in a Row When It Makes Improvement Equal to the Required Amount Under Varying Inclusion Rules for Subgroups

Reading or Mathematics	Overall School	Subgroup No min N	Subgroup Min N=10	Subgroup Min N=31
Scaled Score	33	87	74	59
Index	18	85	64	44
Pass/Fail, pi=.50	33	92	77	61
Pass/Fail, pi=.25	25	80	64	49
Pass/Fail, pi=.10	25	62	52	43
Pass/Fail, pi=.90	90	99	99	96

The results in Table 29 show that compared to the results for the overall school, when subgroups with a minimum size of 31 students are included, approximately twice as many schools are identified as failing to make AYP two years in a row. With some minor variation, the gap between the overall school result and subgroup result is consistent across the reporting statistics. When evaluating these results, it is important to recall that the analysis was based on a model in which students in all subgroups made improvement equal to their improvement goal two years in a row. The increase in the percentage of schools identified as failing to make AYP two years in a row is a function of increased sampling error across multiple pass/fail tests of progress.

As expected, the results in Table 29 also show that the percentage of schools identified as failing AYP progressively decreases as the minimum sample size increases. When evaluating their results, a state must determine how much of the decrease is due to increased reliability due to larger sample sizes and how much of the decrease is due to the elimination of subgroups from the school accountability system. For the data used in this series of studies, Table 30 provides the number of schools with included for each of the four subgroups as the minimum size is increased.

Table 30

Impact of Minimum Subgroup Size on the Number of Schools with Subgroups Total Number of Schools = 911

Subgroup	No minimum	Minimum N=10	Minimum N=31
Economically Disadvantaged	896	641	274
Students with Disabilities	779	246	8
African-American/Black Students	801	623	412
White Students	781	561	350

The figures in Table 30 show a titanic decrease in the number of schools with subgroups as the minimum sample size increases from 1 to 10 to 31. Most striking, is the decrease in number of schools in which students with disabilities, the lowest performing subgroup statewide (see Table 26) would be included in the accountability system. With a minimum sample size of 31, students with disabilities are removed from the accountability system in all but eight schools. An alternative way to view the question, is to examine the number of subgroups that are included in a school's AYP classification as the minimum subgroup sample size increases. Table 31 presents the number of schools with 0, 1, 2, 3, or 4 subgroups under different minimum sample sizes. Of course, the reported number of schools with multiple subgroups may be overestimated in Table 31, if there is a significant overlap in subgroup membership. On the other hand, the data presented in this paper represent a single grade level. When school results are aggregated across several grade levels, the number of students within each subgroup will increase. States must take both of these factors into account when evaluating the reliability of their systems.

Table 31

Impact of Minimum Subgroup Size
on the Number of Subgroups within A School
Total Number of Schools = 911

Number of Subgroups	No minimum	Minimum N=10	Minimum N=31
None	3	100	246
One	48	42	157
Two	51	314	385
Three	193	287	116
Four	616	168	7

Study 19 – Computing the probability of classification errors over two years for two content areas, including subgroups: Part 4 – When the status and improvement are considered

To this point, the studies conducted with the random draws with replacement method have focused on the probability of determining whether schools have made AYP under varying improvement models. The results of those studies indicate that when subgroups of moderate size are included in the model, a substantial percentage of schools will be identified as failing to make AYP two years in a row even when students in each of the subgroups has made significant progress each year.

Under NCLB, however, annual improvement is not the sole, or even primary, determiner of whether a school has made AYP. Schools that pass the status test will be classified as having made AYP and will not be evaluated for improvement. As shown in Study 9 examining overall school results, more than 70 percent of schools will be classified as making AYP on the basis of status alone. [6] The final study presented in this paper examines the impact on classification errors of considering status and improvement when evaluating school performance across two content areas and multiple subgroups.

The design of a study involving status and improvement in two content areas for multiple subgroups across multiple years can appear to become quite complex very quickly. At the heart, or center, of the study, however, there are three basic yes/no questions that must be addressed for the overall school and each subgroup in each content area, and in each year:

[6] Of course, over time increasingly larger percentages of schools will have to show regular improvement in order to meet the steadily rising status line.

1. Is the group exempt from AYP on the basis of size?
2. Did the group make AYP on the basis of status?
3. Did the group make AYP on the basis of improvement?

If the answer to *any one* of those three questions is 'yes', then that subgroup is considered to have made AYP for the year. This is the one key point in the system where there are multiple methods to pass the test rather than conjunctive tests that must be passed.

Working out from center the following questions are addressed for each year:

- Did the overall school and each subgroup make AYP in reading?
- Did the overall school and each subgroup make AYP in mathematics?

If the answer to both of those questions is 'yes', then the school is considered to have made AYP for the year. Conversely, if the answer to either question is 'no', then the school has not made AYP for the year. In our study, therefore, with five groups (i.e., the overall school and up to four subgroups) and two content areas, there would be as many as 10 pass/fail tests or yes/no questions that would have to be passed for a school to make AYP in a given year.

The final question, of course, addresses performance across years:

- Did the school make AYP in Year 1?
- Did the school make AYP in Year 2?

If the answer to either of those questions is 'yes', then the school will not be identified as failing AYP two years in a row.

The results of the status and improvement analyses under the no improvement and improvement conditions are provided in Table 32A and Table 32B. The tables provide the results for reading and mathematics combined with a minimum subgroup size of more than 30 students. In considering the following results, recall the impact of setting the minimum subgroup size to 31. Small schools (n=10-30), of course will have no subgroups. Overall, as shown in Tables 30 and 31, the number of subgroups within a school in this state is dramatically reduced with a minimum subgroup size of 31.

Table 32A

Probability a School and All Subgroups Will Fail to Make AYP on Status or Improvement Two Years in a Row When It Makes No Improvement Minimum Subgroup Size of 31

Reading or Mathematics	N=10-30	N=31-69	N=70-130
Scaled Score	16	16	26
Index	17	15	26
Pass/Fail, pi=.50	20	19	30
Pass/Fail, pi=.25	15	15	26
Pass/Fail, pi=.10	9	11	20
Pass/Fail, pi=.90	45	35	41

Table 32B

Probability a School and All Subgroups Will Fail to Make AYP on Status or Improvement Two Years in a Row When It Makes Improvement Equal to the Required Amount Minimum Subgroup Size of 31

Reading or Mathematics	N=10-30	N=31-69	N=70-130
Scaled Score	5	3	6
Index	5	3	5
Pass/Fail, pi=.50	9	5	8
Pass/Fail, pi=.25	5	2	5
Pass/Fail, pi=.10	4	4	5
Pass/Fail, pi=.90	30	16	21

The results in Table 32B paint a very different picture than those presented in Table 28D. When schools make improvement and both status and improvement are considered, the percentage of schools identified as failing to make AYP two years in a row drops below 10 percent for all but the most stringent, pi=.90, reporting statistic. In contrast, when improvement alone was considered, approximately half of the larger schools that were modeled to make improvement were identified as failing to make AYP two years in a row. Considering status, therefore, has apparently significantly improved the reliability of the accountability system. Under the improvement model, the percentage of schools that will be classified as not making AYP two years in a row will be very low.

An alternate interpretation of these results, however, is that the consideration of status has simply temporarily removed a large number of schools from the accountability system. As the status bar is raised over time, the number of schools that will be *required* to improve performance on an annual basis will increase and the ability of the accountability system to detect that improved performance will be critical.

To conclude this study, we will consider the relatively ordinary case of schools with 70-130 students and the pass/fail system with a proficient cut score set at pi=.50. As shown in Table 28D, 49 percent of these schools fail to make AYP on improvement alone. When status and improvement are considered together (see Table 32B), this percentage drops to 8 percent. When status is considered alone, however, the percentage of these schools identified as failing to make AYP two years in a row is 12 percent. Therefore, there are 12 percent of the schools that are relying on the improvement portion of the accountability system to identify them as having made AYP in at least one year. When improvement is evaluated, however, two-thirds of those schools are still identified as failing to make AYP two years in a row even though they have been modeled to make significant improvement each year. For the group of schools below the status bar, the probability of making AYP on the basis of improvement is .33, much lower than the figure of .50 reported in Table 28D for all schools.

Appendix A

Computer Programs Used in Report

Study 1—Generating two sets of scores for each school

```
OPTIONS LINESIZE=112 PAGESIZE=63 FORMCHAR = '|----|-|---';
TITLE 'MANUFACTURED DATA--N=50, RATIO = .84';

DATA STUDENT SCHOOL;
  NSTUD = 50;

  DO I = 1 TO 100000;
*Choose a school mean;
    SCHMEAN = 37.796 * NORMAL(0) + 300;
          SUMSS = 0;

          DO REP = 1 TO NSTUD;
*Randomly choose 50 students for each school and compute the mean of those students;
             STUDSS = SCHMEAN + SQRT(8571.4285)*NORMAL(0);
                SUMSS = SUMSS + STUDSS;
                OUTPUT STUDENT;
                END;
          MEANSS1 = SUMSS / NSTUD;

          SUMSS = 0;
*Choose another random sample of students and compute the mean for them;
          DO REP = 1 TO NSTUD;
             STUDSS = SCHMEAN + SQRT(8400)*NORMAL(0);
                SUMSS = SUMSS + STUDSS;
                END;
          MEANSS2 = SUMSS / NSTUD;
          OUTPUT SCHOOL;
        END;

PROC MEANS DATA=STUDENT;
    VAR STUDSS;
PROC CORR DATA=SCHOOL;
    VAR MEANSS1 MEANSS2;
RUN;
```

Study 2— Comparing the relative efficiency of various reporting statistics

```
OPTIONS LINESIZE=112 PAGESIZE=63 FORMCHAR = '|----|-|---';
TITLE 'MANUFACTURED DATA--N=20, RATIO = .84';

DATA MANUSTUD;

  NSTUD = 20;
  VARTBAR0 = 1600 - 1000/NSTUD;
  VARTWS = (9000 - VARTBAR0)*NSTUD/(NSTUD - 1);
  VARTBAR = VARTBAR0 - VARTWS/NSTUD;
  SDTBAR = SQRT (VARTBAR);
  VAROWS = 1000 + VARTWS;

  DO I = 1 TO 100000;
    SCHMEAN = SDTBAR * NORMAL(0) + 300;
          SUMSS = 0;
          SUMIND = 0;
          SUMPF1 = 0;
          SUMPF2 = 0;
          SUMPF3 = 0;
      DO REP = 1 TO NSTUD;
             STUDSS = SCHMEAN + SQRT(VAROWS)*NORMAL(0);
                IF STUDSS < 232 THEN STUDIND = 1;
                IF 232 <= STUDSS < 300 THEN STUDIND = 2;
                IF 300 <= STUDSS < 428 THEN STUDIND = 3;
                IF 428 <= STUDSS THEN STUDIND = 4;
                IF STUDIND = 1 THEN PF25 = 0; ELSE PF25 = 100;
                IF STUDIND <= 2 THEN PF50 = 0; ELSE PF50 = 100;
                IF STUDIND <= 3 THEN PF90 = 0; ELSE PF90 = 100;
                SUMSS = SUMSS + STUDSS;
                SUMIND = SUMIND + STUDIND;
                SUMPF1 = SUMPF1 + PF25;
                SUMPF2 = SUMPF2 + PF50;
                SUMPF3 = SUMPF3 + PF90;
                END;
          MEANSS1 = SUMSS / NSTUD;
          MEANIND1 = SUMIND / NSTUD;
          PCTPF125 = SUMPF1 / NSTUD;
          PCTPF150 = SUMPF2 / NSTUD;
          PCTPF190 = SUMPF3 / NSTUD;

          SUMSS = 0;
          SUMIND = 0;
          SUMPF1 = 0;
          SUMPF2 = 0;
          SUMPF3 = 0;
      DO REP = 1 TO NSTUD;
             STUDSS = SCHMEAN + SQRT(VAROWS)*NORMAL(0);
                IF STUDSS < 232 THEN STUDIND = 1;
                IF 232 <= STUDSS < 300 THEN STUDIND = 2;
                IF 300 <= STUDSS < 428 THEN STUDIND = 3;
                IF 428 <= STUDSS THEN STUDIND = 4;
                IF STUDIND = 1 THEN PF25 = 0; ELSE PF25 = 100;
                IF STUDIND <= 2 THEN PF50 = 0; ELSE PF50 = 100;
                IF STUDIND <= 3 THEN PF90 = 0; ELSE PF90 = 100;
                SUMSS = SUMSS + STUDSS;
                SUMIND = SUMIND + STUDIND;
                SUMPF1 = SUMPF1 + PF25;
                SUMPF2 = SUMPF2 + PF50;
                SUMPF3 = SUMPF3 + PF90;
                END;
          MEANSS2 = SUMSS / NSTUD;
          MEANIND2 = SUMIND / NSTUD;
          PCTPF225 = SUMPF1 / NSTUD;
          PCTPF250 = SUMPF2 / NSTUD;
          PCTPF290 = SUMPF3 / NSTUD;
          OUTPUT;
        END;
PROC CORR;
```

```
    VAR MEANSS1--PCTPF190 MEANSS2--PCTPF290;
RUN;
```

Study 3—Computing the probability of classification errors for a status design

```
TITLE 'MANUFACTURED DATA--N=20, RATIO = .84';

DATA MANUSTUD;
  NSTUD=20;
  VARTBAR0 = 1600 - 1000/NSTUD;
  VARTWS = (9000 - VARTBAR0)*NSTUD/(NSTUD - 1);
  VARTBAR = VARTBAR0 - VARTWS/NSTUD;
  SDTBAR = SQRT (VARTBAR);
  VAROWS = 1000 + VARTWS;

  DO I = 1 TO 100000;
    SCHMEAN = SDTBAR * NORMAL(0) + 300;
        Z1 = (232 - SCHMEAN)/SQRT(VAROWS);
        Z2 = (300 - SCHMEAN)/SQRT(VAROWS);
        Z3 = (428 - SCHMEAN)/SQRT(VAROWS);
        P1 = PROBNORM(Z1);
        P2 = PROBNORM(Z2) - PROBNORM(Z1);
        P3 = PROBNORM(Z3) - PROBNORM(Z2);
        P4 = 1 - PROBNORM(Z3);
        INDEX = P1 + 2*P2 + 3*P3 + 4*P4;
          SUMSS = 0;
          SUMIND = 0;
          SUMPF1 = 0;
          SUMPF2 = 0;
          SUMPF3 = 0;
          SUMSSI = 0;
          SUMINDI = 0;
          SUMPF1I = 0;
          SUMPF2I = 0;
          SUMPF3I = 0;
      DO REP = 1 TO NSTUD;
             STUDSS = SCHMEAN + SQRT(VAROWS)*NORMAL(0);

                IF STUDSS < 232 THEN STUDIND = 1;
                IF 232 <= STUDSS < 300 THEN STUDIND = 2;
                IF 300 <= STUDSS < 428 THEN STUDIND = 3;
                IF 428 <= STUDSS THEN STUDIND = 4;
                IF STUDIND = 1 THEN PF25 = 0; ELSE PF25 = 100;
                IF STUDIND <= 2 THEN PF50 = 0; ELSE PF50 = 100;
                IF STUDIND <= 3 THEN PF90 = 0; ELSE PF90 = 100;
                SUMSS = SUMSS + STUDSS;
                SUMIND = SUMIND + STUDIND;
                SUMPF1 = SUMPF1 + PF25;
                SUMPF2 = SUMPF2 + PF50;
                SUMPF3 = SUMPF3 + PF90;
               END;
          MEANSS1 = SUMSS / NSTUD;
          IF MEANSS1 < 266 THEN STATSS1 = 'NOT OK'; ELSE STATSS1 = '    OK';
          MEANIND1 = SUMIND / NSTUD;
          IF MEANIND1 < 2.05 THEN STATIND1 = 'NOT OK'; ELSE STATIND1 = '    OK';
          PCTPF125 = SUMPF1 / NSTUD;
          IF PCTPF125 < 65 THEN STATPF251 = 'NOT OK'; ELSE STATPF251 = '    OK';
          PCTPF150 = SUMPF2 / NSTUD;
          IF PCTPF150 < 36 THEN STATPF501 = 'NOT OK'; ELSE STATPF501 = '    OK';
          PCTPF190 = SUMPF3 / NSTUD;
          IF PCTPF190 < 4 THEN STATPF901 = 'NOT OK'; ELSE STATPF901 = '    OK';

          IF SCHMEAN < 266 THEN STATSS2 = 'NOT OK'; ELSE STATSS2 = '    OK';
      IF INDEX < 2.05 THEN STATIND2 = 'NOT OK'; ELSE STATIND2 = '    OK';
          IF P1 < .35 THEN STATPF252 = '    OK'; ELSE STATPF252 = 'NOT OK';
          IF (P1 + P2) < .64 THEN STATPF502 = '    OK'; ELSE STATPF502 = 'NOT OK';
          IF P4 < .04 THEN STATPF902 = 'NOT OK'; ELSE STATPF902 = '    OK';
          OUTPUT;
        END;
PROC FREQ;
   TABLES STATSS1*STATSS2;
   TABLES STATIND1*STATIND2;
   TABLES STATPF251*STATPF252;
   TABLES STATPF501*STATPF502;
   TABLES STATPF901*STATPF902;

RUN;
```

Study 4—Computing the probability of classification errors for an improvement design: Part 1—When schools make no improvement

```
TITLE 'MANUFACTURED DATA--N=20, RATIO = .84';
TITLE2 'PROBABILITY OF CORRECT CLASSIFICATION WHEN SCHOOL MAKES NO PROGRESS';

DATA MANUSTUD;
  NSTUD=20;

 VARTBAR0 = 1600 - 1000/NSTUD;
  VARTWS = (9000 - VARTBAR0)*NSTUD/(NSTUD - 1);
  VARTBAR = VARTBAR0 - VARTWS/NSTUD;
  SDTBAR = SQRT (VARTBAR);
  VAROWS = 1000 + VARTWS;

  DO I = 1 TO 100000;
    SCHMEAN = SDTBAR * NORMAL(0) + 300;

  *YEAR 1;

          SUMSS = 0;
          SUMIND = 0;
          SUMPF1 = 0;
          SUMPF2 = 0;
          SUMPF3 = 0;
          SUMPF4 = 0;
          DO REP = 1 TO NSTUD;
             STUDSS = SCHMEAN + SQRT(VAROWS)*NORMAL(0);

                IF STUDSS < 232 THEN STUDIND = 1;
                IF 232 <= STUDSS < 300 THEN STUDIND = 2;
                IF 300 <= STUDSS < 428 THEN STUDIND = 3;
                IF 428 <= STUDSS THEN STUDIND = 4;
                IF STUDIND = 1 THEN PF25 = 0; ELSE PF25 = 100;
                IF STUDIND <= 2 THEN PF50 = 0; ELSE PF50 = 100;
                IF STUDIND <= 3 THEN PF90 = 0; ELSE PF90 = 100;
                IF STUDSS < 172 THEN PF10 = 0; ELSE PF10 = 100;
                SUMSS = SUMSS + STUDSS;
                SUMIND = SUMIND + STUDIND;
                SUMPF1 = SUMPF1 + PF25;
                SUMPF2 = SUMPF2 + PF50;
                SUMPF3 = SUMPF3 + PF90;
                SUMPF4 = SUMPF4 + PF10;
                END;
          MEANSS1 = SUMSS / NSTUD;
          IF MEANSS1 < 220 THEN GROUP = 1;
          IF 220 <= MEANSS1 < 260 THEN GROUP = 2;
          IF 260 <= MEANSS1 < 300 THEN GROUP = 3;
          IF 300 <= MEANSS1 THEN GROUP = 4;
          MEANIND1 = SUMIND / NSTUD;
          PCTPF125 = SUMPF1 / NSTUD;
          PCTPF150 = SUMPF2 / NSTUD;
          PCTPF190 = SUMPF3 / NSTUD;
          PCTPF110 = SUMPF4 / NSTUD;
          GTSS1 = (465 - MEANSS1) / 12;
          GTIND1 = (3.5 - MEANIND1) / 12;
          GTPF125 = (100 - PCTPF125) / 10;
          GTPF150 = (100 - PCTPF150) / 10;
          GTPF190 = (100 - PCTPF190) / 10;
          GTPF110 = (100 - PCTPF110) / 10;

  *YEAR 2;

          SUMSS = 0;
          SUMIND = 0;
          SUMPF1 = 0;
          SUMPF2 = 0;
          SUMPF3 = 0;
          SUMPF4 = 0;
      DO REP = 1 TO NSTUD;
```

```
            STUDSS = SCHMEAN + SQRT(VAROWS)*NORMAL(0);
                IF STUDSS < 232 THEN STUDIND = 1;
                IF 232 <= STUDSS < 300 THEN STUDIND = 2;
                IF 300 <= STUDSS < 428 THEN STUDIND = 3;
                IF 428 <= STUDSS THEN STUDIND = 4;
                IF STUDIND = 1 THEN PF25 = 0; ELSE PF25 = 100;
                IF STUDIND <= 2 THEN PF50 = 0; ELSE PF50 = 100;
                IF STUDIND <= 3 THEN PF90 = 0; ELSE PF90 = 100;
                IF STUDSS < 172 THEN PF10 = 0; ELSE PF10 = 100;
                SUMSS = SUMSS + STUDSS;
                SUMIND = SUMIND + STUDIND;
                SUMPF1 = SUMPF1 + PF25;
                SUMPF2 = SUMPF2 + PF50;
                SUMPF3 = SUMPF3 + PF90;
                SUMPF4 = SUMPF4 + PF10;
                END;
         MEANSS2 = SUMSS / NSTUD;
         MEANIND2 = SUMIND / NSTUD;
         PCTPF225 = SUMPF1 / NSTUD;
         PCTPF250 = SUMPF2 / NSTUD;
         PCTPF290 = SUMPF3 / NSTUD;
         PCTPF210 = SUMPF4 / NSTUD;

      IF (MEANSS2 - MEANSS1) < GTSS1 THEN AYPSS1 = 100; ELSE AYPSS1 = 0;
         IF (MEANIND2 - MEANIND1) < GTIND1 THEN AYPIND1 = 100; ELSE AYPIND1 = 0;
         IF (PCTPF225 - PCTPF125) < GTPF125 THEN AYPPF251 = 100; ELSE AYPPF251 = 0;
         IF (PCTPF250 - PCTPF150) < GTPF150 THEN AYPPF501 = 100; ELSE AYPPF501 = 0;
         IF (PCTPF290 - PCTPF190) < GTPF190 THEN AYPPF901 = 100; ELSE AYPPF901 = 0;
         IF (PCTPF210 - PCTPF110) < GTPF110 THEN AYPPF101 = 100; ELSE AYPPF101 = 0;
         OUTPUT;
    END;
PROC FREQ;
    TABLES AYPSS1--AYPPF101;
RUN;
```

Study 5—Computing the probability of classification errors for an improvement design: Part 2—When schools make substantial improvement

```
TITLE 'MANUFACTURED DATA--N=20, RATIO = .84';
TITLE2 'PROBABILITY OF CORRECT CLASSIFICATION WHEN SCHOOL MAKES TWICE REQUIRED PROGRESS';

DATA MANUSTUD;
  NSTUD=20;

 VARTBAR0 = 1600 - 1000/NSTUD;
  VARTWS = (9000 - VARTBAR0)*NSTUD/(NSTUD - 1);
  VARTBAR = VARTBAR0 - VARTWS/NSTUD;
  SDTBAR = SQRT (VARTBAR);
  VAROWS = 1000 + VARTWS;

  DO I = 1 TO 100000;
    SCHMEAN = SDTBAR * NORMAL(0) + 300;

  *YEAR 1;

          SUMSS = 0;
          SUMIND = 0;
          SUMPF1 = 0;
          SUMPF2 = 0;
          SUMPF3 = 0;
          SUMPF4 = 0;
          DO REP = 1 TO NSTUD;
             STUDSS = SCHMEAN + SQRT(VAROWS)*NORMAL(0);
             INCREMENT = 2 * (465 - SCHMEAN)/ 12;
                IF STUDSS < 232 THEN STUDIND = 1;
                IF 232 <= STUDSS < 300 THEN STUDIND = 2;
                IF 300 <= STUDSS < 428 THEN STUDIND = 3;
                IF 428 <= STUDSS THEN STUDIND = 4;
                IF STUDIND = 1 THEN PF25 = 0; ELSE PF25 = 100;
                IF STUDIND <= 2 THEN PF50 = 0; ELSE PF50 = 100;
                IF STUDIND <= 3 THEN PF90 = 0; ELSE PF90 = 100;
                IF STUDSS < 172 THEN PF10 = 0; ELSE PF10 = 100;
                SUMSS = SUMSS + STUDSS;
                SUMIND = SUMIND + STUDIND;
                SUMPF1 = SUMPF1 + PF25;
                SUMPF2 = SUMPF2 + PF50;
                SUMPF3 = SUMPF3 + PF90;
                SUMPF4 = SUMPF4 + PF10;
                END;
          MEANSS1 = SUMSS / NSTUD;
          IF MEANSS1 < 220 THEN GROUP = 1;
          IF 220 <= MEANSS1 < 260 THEN GROUP = 2;
          IF 260 <= MEANSS1 < 300 THEN GROUP = 3;
          IF 300 <= MEANSS1 THEN GROUP = 4;
          MEANIND1 = SUMIND / NSTUD;
          PCTPF125 = SUMPF1 / NSTUD;
          PCTPF150 = SUMPF2 / NSTUD;
          PCTPF190 = SUMPF3 / NSTUD;
          PCTPF110 = SUMPF4 / NSTUD;
          GTSS1 = (465 - MEANSS1) / 12;
          GTIND1 = (3.5 - MEANIND1) / 12;
          GTPF125 = (100 - PCTPF125) / 10;
          GTPF150 = (100 - PCTPF150) / 10;
          GTPF190 = (100 - PCTPF190) / 10;
          GTPF110 = (100 - PCTPF110) / 10;

  *YEAR 2;

          SUMSS = 0;
          SUMIND = 0;
          SUMPF1 = 0;
          SUMPF2 = 0;
          SUMPF3 = 0;
          SUMPF4 = 0;
```

```
      DO REP = 1 TO NSTUD;
            STUDSS = SCHMEAN + SQRT(VAROWS)*NORMAL(0) + INCREMENT;
                IF STUDSS < 232 THEN STUDIND = 1;
                IF 232 <= STUDSS < 300 THEN STUDIND = 2;
                IF 300 <= STUDSS < 428 THEN STUDIND = 3;
                IF 428 <= STUDSS THEN STUDIND = 4;
                IF STUDIND = 1 THEN PF25 = 0; ELSE PF25 = 100;
                IF STUDIND <= 2 THEN PF50 = 0; ELSE PF50 = 100;
                IF STUDIND <= 3 THEN PF90 = 0; ELSE PF90 = 100;
                IF STUDSS < 172 THEN PF10 = 0; ELSE PF10 = 100;
                SUMSS = SUMSS + STUDSS;
                SUMIND = SUMIND + STUDIND;
                SUMPF1 = SUMPF1 + PF25;
                SUMPF2 = SUMPF2 + PF50;
                SUMPF3 = SUMPF3 + PF90;
                SUMPF4 = SUMPF4 + PF10;
                END;
         MEANSS2 = SUMSS / NSTUD;
         MEANIND2 = SUMIND / NSTUD;
         PCTPF225 = SUMPF1 / NSTUD;
         PCTPF250 = SUMPF2 / NSTUD;
         PCTPF290 = SUMPF3 / NSTUD;
         PCTPF210 = SUMPF4 / NSTUD;

      IF (MEANSS2 - MEANSS1) < GTSS1 THEN AYPSS1 = 100; ELSE AYPSS1 = 0;
         IF (MEANIND2 - MEANIND1) < GTIND1 THEN AYPIND1 = 100; ELSE AYPIND1 = 0;
         IF (PCTPF225 - PCTPF125) < GTPF125 THEN AYPPF251 = 100; ELSE AYPPF251 = 0;
         IF (PCTPF250 - PCTPF150) < GTPF150 THEN AYPPF501 = 100; ELSE AYPPF501 = 0;
         IF (PCTPF290 - PCTPF190) < GTPF190 THEN AYPPF901 = 100; ELSE AYPPF901 = 0;
         IF (PCTPF210 - PCTPF110) < GTPF110 THEN AYPPF101 = 100; ELSE AYPPF101 = 0;
         OUTPUT;
    END;
PROC FREQ;
    TABLES AYPSS1--AYPPF101;
RUN;
```

Study 6—Computing the probability of classification errors over two years: Part 1—When schools make no improvement

```
TITLE 'MANUFACTURED DATA--N=20, RATIO = .84';
TITLE2 'PROBABILITY OF FAILING TWO YEARS IN A ROW WHEN SCHOOL NO MAKES PROGRESS';

DATA MANUSTUD;
  NSTUD=20;
  INCREMENT = 0;
  VARTBAR0 = 1600 - 1000/NSTUD;
  VARTWS = (9000 - VARTBAR0)*NSTUD/(NSTUD - 1);
  VARTBAR = VARTBAR0 - VARTWS/NSTUD;
  SDTBAR = SQRT (VARTBAR);
  VAROWS = 1000 + VARTWS;

  DO I = 1 TO 100000;

  *YEAR 1;
    SCHMEAN = SDTBAR * NORMAL(0) + 300;

          SUMSS = 0;
          SUMIND = 0;
          SUMPF1 = 0;
          SUMPF2 = 0;
          SUMPF3 = 0;
          SUMPF4 = 0;
          DO REP = 1 TO NSTUD;
             STUDSS = SCHMEAN + SQRT(VAROWS)*NORMAL(0);
                SS2 = STUDSS + INCREMENT;
                IF STUDSS < 232 THEN STUDIND = 1;
                IF 232 <= STUDSS < 300 THEN STUDIND = 2;
                IF 300 <= STUDSS < 428 THEN STUDIND = 3;
                IF 428 <= STUDSS THEN STUDIND = 4;
                IF STUDIND = 1 THEN PF25 = 0; ELSE PF25 = 100;
                IF STUDIND <= 2 THEN PF50 = 0; ELSE PF50 = 100;
                IF STUDIND <= 3 THEN PF90 = 0; ELSE PF90 = 100;
                IF STUDSS < 172 THEN PF10 = 0; ELSE PF10 = 100;
                SUMSS = SUMSS + STUDSS;
                SUMIND = SUMIND + STUDIND;
                SUMPF1 = SUMPF1 + PF25;
                SUMPF2 = SUMPF2 + PF50;
                SUMPF3 = SUMPF3 + PF90;
                SUMPF4 = SUMPF4 + PF10;
                END;
          MEANSS1 = SUMSS / NSTUD;
          IF MEANSS1 < 266 THEN GROUP = 1;
          IF 266 <= MEANSS1 < 260 THEN GROUP = 2;
          IF 260 <= MEANSS1 < 300 THEN GROUP = 3;
          IF 300 <= MEANSS1 THEN GROUP = 4;
          MEANIND1 = SUMIND / NSTUD;
          PCTPF125 = SUMPF1 / NSTUD;
          PCTPF150 = SUMPF2 / NSTUD;
          PCTPF190 = SUMPF3 / NSTUD;
          PCTPF110 = SUMPF4 / NSTUD;
          GTSS1 = (465 - MEANSS1) / 12;
          GTIND1 = (3.5 - MEANIND1) / 12;
          GTPF125 = (100 - PCTPF125) / 10;
          GTPF150 = (100 - PCTPF150) / 10;
          GTPF190 = (100 - PCTPF190) / 10;
          GTPF110 = (100 - PCTPF110) / 10;
  *YEAR 2;

          SUMSS = 0;
          SUMIND = 0;
          SUMPF1 = 0;
          SUMPF2 = 0;
          SUMPF3 = 0;
          SUMPF4 = 0;
      DO REP = 1 TO NSTUD;
             STUDSS = SCHMEAN + SQRT(VAROWS)*NORMAL(0) + INCREMENT;
```

```
                IF STUDSS < 232 THEN STUDIND = 1;
                IF 232 <= STUDSS < 300 THEN STUDIND = 2;
                IF 300 <= STUDSS < 428 THEN STUDIND = 3;
                IF 428 <= STUDSS THEN STUDIND = 4;
                IF STUDIND = 1 THEN PF25 = 0; ELSE PF25 = 100;
                IF STUDIND <= 2 THEN PF50 = 0; ELSE PF50 = 100;
                IF STUDIND <= 3 THEN PF90 = 0; ELSE PF90 = 100;
                IF STUDSS < 172 THEN PF10 = 0; ELSE PF10 = 100;
                SUMSS = SUMSS + STUDSS;
                SUMIND = SUMIND + STUDIND;
                SUMPF1 = SUMPF1 + PF25;
                SUMPF2 = SUMPF2 + PF50;
                SUMPF3 = SUMPF3 + PF90;
                SUMPF4 = SUMPF4 + PF10;
                END;
         MEANSS2 = SUMSS / NSTUD;
         MEANIND2 = SUMIND / NSTUD;
         PCTPF225 = SUMPF1 / NSTUD;
         PCTPF250 = SUMPF2 / NSTUD;
         PCTPF290 = SUMPF3 / NSTUD;
         PCTPF210 = SUMPF4 / NSTUD;
         GTSS2 = (465 - MEANSS2) / 11;
         GTIND2 = (3.5 - MEANIND2) / 11;
         GTPF225 = (100 - PCTPF225) / 10;
         GTPF250 = (100 - PCTPF250) / 10;
         GTPF290 = (100 - PCTPF290) / 10;
         GTPF210 = (100 - PCTPF210) / 10;
      IF (MEANSS2 - MEANSS1) < GTSS1 THEN AYPSS1 = 100; ELSE AYPSS1 = 0;
         IF (MEANIND2 - MEANIND1) < GTIND1 THEN AYPIND1 = 100; ELSE AYPIND1 = 0;
         IF (PCTPF225 - PCTPF125) < GTPF125 THEN AYPPF251 = 100; ELSE AYPPF251 = 0;
         IF (PCTPF250 - PCTPF150) < GTPF150 THEN AYPPF501 = 100; ELSE AYPPF501 = 0;
         IF (PCTPF290 - PCTPF190) < GTPF190 THEN AYPPF901 = 100; ELSE AYPPF901 = 0;
         IF (PCTPF210 - PCTPF110) < GTPF110 THEN AYPPF101 = 100; ELSE AYPPF101 = 0;

*YEAR 3;

         SUMSS = 0;
         SUMIND = 0;
         SUMPF1 = 0;
         SUMPF2 = 0;
         SUMPF3 = 0;
         SUMPF4 = 0;
      DO REP = 1 TO NSTUD;
            STUDSS = SCHMEAN + SQRT(VAROWS)*NORMAL(0) + 2*INCREMENT;
                IF STUDSS < 232 THEN STUDIND = 1;
                IF 232 <= STUDSS < 300 THEN STUDIND = 2;
                IF 300 <= STUDSS < 428 THEN STUDIND = 3;
                IF 428 <= STUDSS THEN STUDIND = 4;
                IF STUDIND = 1 THEN PF25 = 0; ELSE PF25 = 100;
                IF STUDIND <= 2 THEN PF50 = 0; ELSE PF50 = 100;
                IF STUDIND <= 3 THEN PF90 = 0; ELSE PF90 = 100;
                IF STUDSS < 172 THEN PF10 = 0; ELSE PF10 = 100;
                SUMSS = SUMSS + STUDSS;
                SUMIND = SUMIND + STUDIND;
                SUMPF1 = SUMPF1 + PF25;
                SUMPF2 = SUMPF2 + PF50;
                SUMPF3 = SUMPF3 + PF90;
                SUMPF4 = SUMPF4 + PF10;
                END;
         MEANSS3 = SUMSS / NSTUD;
         MEANIND3 = SUMIND / NSTUD;
         PCTPF325 = SUMPF1 / NSTUD;
         PCTPF350 = SUMPF2 / NSTUD;
         PCTPF390 = SUMPF3 / NSTUD;
         PCTPF310 = SUMPF4 / NSTUD;
         IF (MEANSS3 - MEANSS2) < GTSS2 THEN AYPSS2 = 100; ELSE AYPSS2 = 0;
         IF (MEANIND3 - MEANIND2) < GTIND2 THEN AYPIND2 = 100; ELSE AYPIND2 = 0;
         IF (PCTPF325 - PCTPF225) < GTPF225 THEN AYPPF252 = 100; ELSE AYPPF252 = 0;
         IF (PCTPF350 - PCTPF250) < GTPF250 THEN AYPPF502 = 100; ELSE AYPPF502 = 0;
         IF (PCTPF390 - PCTPF290) < GTPF290 THEN AYPPF902 = 100; ELSE AYPPF902 = 0;
         IF (PCTPF310 - PCTPF210) < GTPF210 THEN AYPPF102 = 100; ELSE AYPPF102 = 0;
```

```
          TWOYRSSS = AYPSS1 * AYPSS2/100;
          TWOYRSIND = AYPIND1 * AYPIND2/100;
          TWOYRSPF25 = AYPPF251*AYPPF252/100;
      TWOYRSPF50 = AYPPF501*AYPPF502/100;
      TWOYRSPF90 = AYPPF901*AYPPF902/100;
      TWOYRSPF10 = AYPPF101*AYPPF102/100;

          OUTPUT;
    END;
PROC TABULATE;
    VAR TWOYRSSS TWOYRSIND TWOYRSPF10 TWOYRSPF50 TWOYRSPF25 TWOYRSPF90;
       CLASS GROUP;
       TABLES (TWOYRSSS TWOYRSIND TWOYRSPF10 TWOYRSPF25 TWOYRSPF50 TWOYRSPF90) ,
               (GROUP ALL)*MEAN*F=5.1;
RUN;
```

Study 7—Computing the probability of classification errors over two years: Part 2—When schools make substantial improvement

```
TITLE 'MANUFACTURED DATA--N=20, RATIO = .84';
TITLE2 'PROBABILITY OF FAILING TWO YEARS IN A ROW WHEN SCHOOL MAKES PROGRESS';

DATA MANUSTUD;
  NSTUD=20;

  VARTBAR0 = 1600 - 1000/NSTUD;
  VARTWS = (9000 - VARTBAR0)*NSTUD/(NSTUD - 1);
  VARTBAR = VARTBAR0 - VARTWS/NSTUD;
  SDTBAR = SQRT (VARTBAR);
  VAROWS = 1000 + VARTWS;

  DO I = 1 TO 100000;

  *YEAR 1;
    SCHMEAN = SDTBAR * NORMAL(0) + 300;
    INCREMENT = (465 - SCHMEAN)/12;
          SUMSS = 0;
          SUMIND = 0;
          SUMPF1 = 0;
          SUMPF2 = 0;
          SUMPF3 = 0;
          SUMPF4 = 0;
          DO REP = 1 TO NSTUD;
             STUDSS = SCHMEAN + SQRT(VAROWS)*NORMAL(0);

                IF STUDSS < 232 THEN STUDIND = 1;
                IF 232 <= STUDSS < 300 THEN STUDIND = 2;
                IF 300 <= STUDSS < 428 THEN STUDIND = 3;
                IF 428 <= STUDSS THEN STUDIND = 4;
                IF STUDIND = 1 THEN PF25 = 0; ELSE PF25 = 100;
                IF STUDIND <= 2 THEN PF50 = 0; ELSE PF50 = 100;
                IF STUDIND <= 3 THEN PF90 = 0; ELSE PF90 = 100;
                IF STUDSS < 172 THEN PF10 = 0; ELSE PF10 = 100;
                SUMSS = SUMSS + STUDSS;
                SUMIND = SUMIND + STUDIND;
                SUMPF1 = SUMPF1 + PF25;
                SUMPF2 = SUMPF2 + PF50;
                SUMPF3 = SUMPF3 + PF90;
                SUMPF4 = SUMPF4 + PF10;
                END;
          MEANSS1 = SUMSS / NSTUD;
          IF MEANSS1 < 266 THEN GROUP = 1;
          IF 266 <= MEANSS1 < 260 THEN GROUP = 2;
          IF 260 <= MEANSS1 < 300 THEN GROUP = 3;
          IF 300 <= MEANSS1 THEN GROUP = 4;
          MEANIND1 = SUMIND / NSTUD;
          PCTPF125 = SUMPF1 / NSTUD;
          PCTPF150 = SUMPF2 / NSTUD;
          PCTPF190 = SUMPF3 / NSTUD;
          PCTPF110 = SUMPF4 / NSTUD;
          GTSS1 = (465 - MEANSS1) / 12;
          GTIND1 = (3.5 - MEANIND1) / 12;
          GTPF125 = (100 - PCTPF125) / 10;
          GTPF150 = (100 - PCTPF150) / 10;
          GTPF190 = (100 - PCTPF190) / 10;
          GTPF110 = (100 - PCTPF110) / 10;
  *YEAR 2;

          SUMSS = 0;
          SUMIND = 0;
          SUMPF1 = 0;
          SUMPF2 = 0;
          SUMPF3 = 0;
          SUMPF4 = 0;
      DO REP = 1 TO NSTUD;
             STUDSS = SCHMEAN + SQRT(VAROWS)*NORMAL(0) + INCREMENT;
```

```
                IF STUDSS < 232 THEN STUDIND = 1;
                IF 232 <= STUDSS < 300 THEN STUDIND = 2;
                IF 300 <= STUDSS < 428 THEN STUDIND = 3;
                IF 428 <= STUDSS THEN STUDIND = 4;
                IF STUDIND = 1 THEN PF25 = 0; ELSE PF25 = 100;
                IF STUDIND <= 2 THEN PF50 = 0; ELSE PF50 = 100;
                IF STUDIND <= 3 THEN PF90 = 0; ELSE PF90 = 100;
                IF STUDSS < 172 THEN PF10 = 0; ELSE PF10 = 100;
                SUMSS = SUMSS + STUDSS;
                SUMIND = SUMIND + STUDIND;
                SUMPF1 = SUMPF1 + PF25;
                SUMPF2 = SUMPF2 + PF50;
                SUMPF3 = SUMPF3 + PF90;
                SUMPF4 = SUMPF4 + PF10;
                END;
         MEANSS2 = SUMSS / NSTUD;
         MEANIND2 = SUMIND / NSTUD;
         PCTPF225 = SUMPF1 / NSTUD;
         PCTPF250 = SUMPF2 / NSTUD;
         PCTPF290 = SUMPF3 / NSTUD;
         PCTPF210 = SUMPF4 / NSTUD;
         GTSS2 = (465 - MEANSS2) / 11;
         GTIND2 = (3.5 - MEANIND2) / 11;
         GTPF225 = (100 - PCTPF225) / 10;
         GTPF250 = (100 - PCTPF250) / 10;
         GTPF290 = (100 - PCTPF290) / 10;
         GTPF210 = (100 - PCTPF210) / 10;
      IF (MEANSS2 - MEANSS1) < GTSS1 THEN AYPSS1 = 100; ELSE AYPSS1 = 0;
         IF (MEANIND2 - MEANIND1) < GTIND1 THEN AYPIND1 = 100; ELSE AYPIND1 = 0;
         IF (PCTPF225 - PCTPF125) < GTPF125 THEN AYPPF251 = 100; ELSE AYPPF251 = 0;
         IF (PCTPF250 - PCTPF150) < GTPF150 THEN AYPPF501 = 100; ELSE AYPPF501 = 0;
         IF (PCTPF290 - PCTPF190) < GTPF190 THEN AYPPF901 = 100; ELSE AYPPF901 = 0;
         IF (PCTPF210 - PCTPF110) < GTPF110 THEN AYPPF101 = 100; ELSE AYPPF101 = 0;

*YEAR 3;

         SUMSS = 0;
         SUMIND = 0;
         SUMPF1 = 0;
         SUMPF2 = 0;
         SUMPF3 = 0;
         SUMPF4 = 0;
      DO REP = 1 TO NSTUD;
            STUDSS = SCHMEAN + SQRT(VAROWS)*NORMAL(0) + 2*INCREMENT;
                IF STUDSS < 232 THEN STUDIND = 1;
                IF 232 <= STUDSS < 300 THEN STUDIND = 2;
                IF 300 <= STUDSS < 428 THEN STUDIND = 3;
                IF 428 <= STUDSS THEN STUDIND = 4;
                IF STUDIND = 1 THEN PF25 = 0; ELSE PF25 = 100;
                IF STUDIND <= 2 THEN PF50 = 0; ELSE PF50 = 100;
                IF STUDIND <= 3 THEN PF90 = 0; ELSE PF90 = 100;
                IF STUDSS < 172 THEN PF10 = 0; ELSE PF10 = 100;
                SUMSS = SUMSS + STUDSS;
                SUMIND = SUMIND + STUDIND;
                SUMPF1 = SUMPF1 + PF25;
                SUMPF2 = SUMPF2 + PF50;
                SUMPF3 = SUMPF3 + PF90;
                SUMPF4 = SUMPF4 + PF10;
                END;
         MEANSS3 = SUMSS / NSTUD;
         MEANIND3 = SUMIND / NSTUD;
         PCTPF325 = SUMPF1 / NSTUD;
         PCTPF350 = SUMPF2 / NSTUD;
         PCTPF390 = SUMPF3 / NSTUD;
         PCTPF310 = SUMPF4 / NSTUD;
         IF (MEANSS3 - MEANSS2) < GTSS2 THEN AYPSS2 = 100; ELSE AYPSS2 = 0;
         IF (MEANIND3 - MEANIND2) < GTIND2 THEN AYPIND2 = 100; ELSE AYPIND2 = 0;
         IF (PCTPF325 - PCTPF225) < GTPF225 THEN AYPPF252 = 100; ELSE AYPPF252 = 0;
         IF (PCTPF350 - PCTPF250) < GTPF250 THEN AYPPF502 = 100; ELSE AYPPF502 = 0;
         IF (PCTPF390 - PCTPF290) < GTPF290 THEN AYPPF902 = 100; ELSE AYPPF902 = 0;
         IF (PCTPF310 - PCTPF210) < GTPF210 THEN AYPPF102 = 100; ELSE AYPPF102 = 0;
```

```
          TWOYRSSS = AYPSS1 * AYPSS2/100;
          TWOYRSIND = AYPIND1 * AYPIND2/100;
          TWOYRSPF25 = AYPPF251*AYPPF252/100;
      TWOYRSPF50 = AYPPF501*AYPPF502/100;
      TWOYRSPF90 = AYPPF901*AYPPF902/100;
      TWOYRSPF10 = AYPPF101*AYPPF102/100;

          OUTPUT;
    END;
PROC TABULATE;
    VAR TWOYRSSS TWOYRSIND TWOYRSPF10 TWOYRSPF50 TWOYRSPF25 TWOYRSPF90;
       CLASS GROUP;
       TABLES (TWOYRSSS TWOYRSIND TWOYRSPF10 TWOYRSPF25 TWOYRSPF50 TWOYRSPF90) ,
               (GROUP ALL)*MEAN*F=5.1;
RUN;
```

Study 8—Generating two sets of scores for each school, using two content areas

```
TITLE 'MANUFACTURED DATA--N=20, RATIO = .84';

DATA STUDENT SCHOOL;
  NSTUD = 20;
  INCREMENT = 0;
  VARTBAR0 = 1600 - 1000/NSTUD;
  VARTWS = (9000 - VARTBAR0)*NSTUD/(NSTUD - 1);
  VARTBAR = VARTBAR0 - VARTWS/NSTUD;
  SDTBAR = SQRT (VARTBAR);
  VAROWS = 1000 + VARTWS;

  DO I = 1 TO 1000;
*Choose two normal random variables to create school true mean for reading and math;
    Z1 = NORMAL(0);
       Z2 = NORMAL(0);
*Set the correlation of true means between reading and math at the school level;
       SCHRMCORR = .95;
*Set the correlation between reading and math observed scores at the student level within school;
    STURMCORR = .75;

*Select a random school and compute its true mean reading and math scores;
    READ = SDTBAR * Z1 + 300;
       RSQ = SCHRMCORR**2;
       MATHZ = SCHRMCORR*Z1 + SQRT(1-RSQ) * Z2;
       MATH = SDTBAR * MATHZ + 300;


         SUMREAD = 0;
         SUMMATH = 0;

         DO REP = 1 TO NSTUD;
*Randomly choose NSTUD students for each school and compute their scores in reading and math;
         Z3 = NORMAL(0);
                Z4 = NORMAL(0);
                RSQ = STURMCORR**2;
            READSS = READ + SQRT(VAROWS)*Z3;
                MATHZ = STURMCORR * Z3 + SQRT (1-RSQ) * Z4;
                MATHSS = MATH + SQRT(VAROWS)*MATHZ;
                SUMREAD = SUMREAD + READSS;
                SUMMATH = SUMMATH + MATHSS;
                OUTPUT STUDENT;
                END;
         MEANREAD1 = SUMREAD / NSTUD;
         MEANMATH1 = SUMMATH / NSTUD;

         SUMREAD = 0;
         SUMMATH = 0;
*Choose another random sample of students and compute the mean for them;
         DO REP = 1 TO NSTUD;
            Z3 = NORMAL(0);
                Z4 = NORMAL(0);
                RSQ = STURMCORR**2;
            READSS = READ + SQRT(VAROWS)*Z3;
                MATHZ = STURMCORR * Z3 + SQRT (1-RSQ) * Z4;
                MATHSS = MATH + SQRT(VAROWS)*MATHZ;
                SUMREAD = SUMREAD + READSS;
                SUMMATH = SUMMATH + MATHSS;
                OUTPUT STUDENT;
                END;
         MEANREAD2 = SUMREAD / NSTUD;
         MEANMATH2 = SUMMATH / NSTUD;
         OUTPUT SCHOOL;
       END;

PROC CORR DATA=STUDENT;
    VAR READSS MATHSS;
PROC CORR DATA=SCHOOL;
    VAR MEANREAD1 MEANMATH1 MEANREAD2 MEANMATH2;
```

```
RUN;
```

Study 9—Computing the probability of classification errors for a status design, using two content areas

```
TITLE 'MANUFACTURED DATA--N=100, RATIO = .84';
TITLE2 ' PROBABILITY OF CORRECT CLASSIFICATION WITH TWO CONTENT AREAS ';

DATA SCHOOL;
  NSTUD = 100;
  INCREMENT = 0;
  VARTBAR0 = 1600 - 1000/NSTUD;
  VARTWS = (9000 - VARTBAR0)*NSTUD/(NSTUD - 1);
  VARTBAR = VARTBAR0 - VARTWS/NSTUD;
  SDTBAR = SQRT (VARTBAR);
  VAROWS = 1000 + VARTWS;

  DO I = 1 TO 100000;
*Choose two normal random variables to create school true mean for reading and math;
    Z1 = NORMAL(0);
        Z2 = NORMAL(0);
*Set the correlation of true means between reading and math at the school level;
        SCHRMCORR = .95;
*Set the correlation between reading and math observed scores at the student level within school;
    STURMCORR = .75;

*Select a random school and compute its true mean reading and math scores;
    READ = SDTBAR * Z1 + 300;
        RSQ = SCHRMCORR**2;
        MATHZ = SCHRMCORR*Z1 + SQRT(1-RSQ) * Z2;
        MATH = SDTBAR * MATHZ + 300;

*Compute true statistics for school;

        ZR1 = (232 - READ)/SQRT(VAROWS);
        ZR2 = (300 - READ)/SQRT(VAROWS);
        ZR3 = (428 - READ)/SQRT(VAROWS);
        PR1 = PROBNORM(ZR1);
        PR2 = PROBNORM(ZR2) - PROBNORM(ZR1);
        PR3 = PROBNORM(ZR3) - PROBNORM(ZR2);
        PR4 = 1 - PROBNORM(ZR3);
        INDEXR = PR1 + 2*PR2 + 3*PR3 + 4*PR4;

        ZM1 = (232 - MATH)/SQRT(VAROWS);
        ZM2 = (300 - MATH)/SQRT(VAROWS);
        ZM3 = (428 - MATH)/SQRT(VAROWS);
        PM1 = PROBNORM(ZM1);
        PM2 = PROBNORM(ZM2) - PROBNORM(ZM1);
        PM3 = PROBNORM(ZM3) - PROBNORM(ZM2);
        PM4 = 1 - PROBNORM(ZM3);
        INDEXM = PM1 + 2*PM2 + 3*PM3 + 4*PM4;

*Initialize sums for school;
      SUMSSR = 0;
          SUMINDR = 0;
          SUMPF1R = 0;
          SUMPF2R = 0;
          SUMPF3R = 0;
          SUMSSM = 0;
          SUMINDM = 0;
          SUMPF1M = 0;
          SUMPF2M = 0;
          SUMPF3M = 0;

      DO REP = 1 TO NSTUD;
*Randomly choose NSTUD students for each school and compute their scores in reading and math;
         Z3 = NORMAL(0);
                Z4 = NORMAL(0);
                RSQ = STURMCORR**2;
            READSS = READ + SQRT(VAROWS)*Z3;
                MATHZ = STURMCORR * Z3 + SQRT (1-RSQ) * Z4;
                MATHSS = MATH + SQRT(VAROWS)*MATHZ;
                SUMREAD = SUMREAD + READSS;
                SUMMATH = SUMMATH + MATHSS;
            STUDSS = SCHMEAN + SQRT(VAROWS)*NORMAL(0);

                IF READSS < 232 THEN READIND = 1;
                IF 232 <= READSS < 300 THEN READIND = 2;
                IF 300 <= READSS < 428 THEN READIND = 3;
                IF 428 <= READSS THEN READIND = 4;
                IF READIND = 1 THEN PFR25 = 0; ELSE PFR25 = 100;
                IF READIND <= 2 THEN PFR50 = 0; ELSE PFR50 = 100;
                IF READIND <= 3 THEN PFR90 = 0; ELSE PFR90 = 100;

         IF MATHSS < 232 THEN MATHIND = 1;
                IF 232 <= MATHSS < 300 THEN MATHIND = 2;
                IF 300 <= MATHSS < 428 THEN MATHIND = 3;
                IF 428 <= MATHSS THEN MATHIND = 4;
                IF MATHIND = 1 THEN PFM25 = 0; ELSE PFM25 = 100;
                IF MATHIND <= 2 THEN PFM50 = 0; ELSE PFM50 = 100;
                IF MATHIND <= 3 THEN PFM90 = 0; ELSE PFM90 = 100;

                SUMSSR = SUMSSR + READSS;
                SUMINDR = SUMINDR + READIND;
                SUMPF1R = SUMPF1R + PFR25;
                SUMPF2R = SUMPF2R + PFR50;
                SUMPF3R = SUMPF3R + PFR90;

                SUMSSM = SUMSSM + MATHSS;
                SUMINDM = SUMINDM + MATHIND;
                SUMPF1M = SUMPF1M + PFM25;
                SUMPF2M = SUMPF2M + PFM50;
                SUMPF3M = SUMPF3M + PFM90;
               END;
          MEANSSR = SUMSSR / NSTUD;
          MEANSSM = SUMSSM / NSTUD;
          IF (MEANSSR < 266) OR (MEANSSM < 266) THEN STATSS1 = 'NOT OK'; ELSE STATSS1 = '    OK';

          MEANINDR = SUMINDR / NSTUD;
          MEANINDM = SUMINDM / NSTUD;
          IF (MEANINDR < 2.05) OR (MEANINDM < 2.05) THEN STATIND1 = 'NOT OK'; ELSE STATIND1 = '
OK';

          PCTPF125R = SUMPF1R / NSTUD;
          PCTPF125M = SUMPF1M / NSTUD;
          IF (PCTPF125R < 65) OR (PCTPF125M < 65) THEN STATPF251 = 'NOT OK'; ELSE STATPF251 = '
OK';

          PCTPF150R = SUMPF2R / NSTUD;
          PCTPF150M = SUMPF2M / NSTUD;
          IF (PCTPF150R < 36) OR (PCTPF150M < 36) THEN STATPF501 = 'NOT OK'; ELSE STATPF501 = '
OK';

          PCTPF190R = SUMPF3R / NSTUD;
          PCTPF190M = SUMPF3M / NSTUD;
          IF (PCTPF190R < 4) OR (PCTPF190M < 4) THEN STATPF901 = 'NOT OK'; ELSE STATPF901 = '
OK';

          IF (READ < 266) OR (MATH < 266) THEN STATSS2 = 'NOT OK'; ELSE STATSS2 = '    OK';
           IF (INDEXR < 2.05) OR (INDEXM < 2.05) THEN STATIND2 = 'NOT OK'; ELSE STATIND2 = '
OK';
          IF (PR1 < .35) AND (PM1 < .35) THEN STATPF252 = '    OK'; ELSE STATPF252 = 'NOT OK';
          IF ((PR1 + PR2) < .64) AND ((PM1 + PM2) < .64) THEN STATPF502 = '    OK'; ELSE
STATPF502 = 'NOT OK';
          IF (PR4 < .04) OR (PM4 <.04) THEN STATPF902 = 'NOT OK'; ELSE STATPF902 = '    OK';
          OUTPUT;
        END;
PROC FREQ;
   TABLES STATSS1*STATSS2;
   TABLES STATIND1*STATIND2;
   TABLES STATPF251*STATPF252;
   TABLES STATPF501*STATPF502;
   TABLES STATPF901*STATPF902;
```

```
RUN;
```

Study 10—Computing the probability of classification errors for an improvement design, using two content areas: Part 1—When schools make no progress

```
TITLE 'MANUFACTURED DATA--N=100, RATIO = .84';
TITLE2 'PROBABILITY OF CORRECT CLASSIFICATION WHEN SCHOOL MAKES NO PROGRESS--TWO CONTENT AREAS';

DATA MANUSTUD;
  NSTUD=100;
  INCREMENT = 0;
  VARTBAR0 = 1600 - 1000/NSTUD;
  VARTWS = (9000 - VARTBAR0)*NSTUD/(NSTUD - 1);
  VARTBAR = VARTBAR0 - VARTWS/NSTUD;
  SDTBAR = SQRT (VARTBAR);
  VAROWS = 1000 + VARTWS;

  DO I = 1 TO 100000;
*Choose two normal random variables to create school true mean for reading and math;
    Z1 = NORMAL(0);
        Z2 = NORMAL(0);
*Set the correlation of true means between reading and math at the school level;
        SCHRMCORR = .95;
*Set the correlation between reading and math observed scores at the student level within school;
    STURMCORR = .75;

*Select a random school and compute its true mean reading and math scores;
    READ = SDTBAR * Z1 + 300;
        RSQ = SCHRMCORR**2;
        MATHZ = SCHRMCORR*Z1 + SQRT(1-RSQ) * Z2;
        MATH = SDTBAR * MATHZ + 300;
    INCREMENTR = 0;
    INCREMENTM = 0;
  *YEAR 1;

*Initialize sums for school;
      SUMSSR = 0;
          SUMINDR = 0;
          SUMPF1R = 0;
          SUMPF2R = 0;
          SUMPF3R = 0;
          SUMPF4R = 0;
          SUMSSM = 0;
          SUMINDM = 0;
          SUMPF1M = 0;
          SUMPF2M = 0;
          SUMPF3M = 0;
          SUMPF4M = 0;

          DO REP = 1 TO NSTUD;

*Randomly choose NSTUD students for each school and compute their scores in reading and math;
         Z3 = NORMAL(0);
                Z4 = NORMAL(0);
                RSQ = STURMCORR**2;
            READSS = READ + SQRT(VAROWS)*Z3;
                MATHZ = STURMCORR * Z3 + SQRT (1-RSQ) * Z4;
                MATHSS = MATH + SQRT(VAROWS)*MATHZ;

                IF READSS < 232 THEN READIND = 1;
                IF 232 <= READSS < 300 THEN READIND = 2;
                IF 300 <= READSS < 428 THEN READIND = 3;
                IF 428 <= READSS THEN READIND = 4;
                IF READIND = 1 THEN PF25R = 0; ELSE PF25R = 100;
                IF READIND <= 2 THEN PF50R = 0; ELSE PF50R = 100;
                IF READIND <= 3 THEN PF90R = 0; ELSE PF90R = 100;
                IF READSS < 172 THEN PF10R = 0; ELSE PF10R = 100;
                SUMSSR = SUMSSR + READSS;
                SUMINDR = SUMINDR + READIND;
                SUMPF1R = SUMPF1R + PF25R;
                SUMPF2R = SUMPF2R + PF50R;
                SUMPF3R = SUMPF3R + PF90R;
                SUMPF4R = SUMPF4R + PF10R;

                IF MATHSS < 232 THEN MATHIND = 1;
                IF 232 <= MATHSS < 300 THEN MATHIND = 2;
                IF 300 <= MATHSS < 428 THEN MATHIND = 3;
                IF 428 <= MATHSS THEN MATHIND = 4;
                IF MATHIND = 1 THEN PF25M = 0; ELSE PF25M = 100;
                IF MATHIND <= 2 THEN PF50M = 0; ELSE PF50M = 100;
                IF MATHIND <= 3 THEN PF90M = 0; ELSE PF90M = 100;
                IF MATHSS < 172 THEN PF10M = 0; ELSE PF10M = 100;
                SUMSSM = SUMSSM + MATHSS;
                SUMINDM = SUMINDM + MATHIND;
                SUMPF1M = SUMPF1M + PF25M;
                SUMPF2M = SUMPF2M + PF50M;
                SUMPF3M = SUMPF3M + PF90M;
                SUMPF4M = SUMPF4M + PF10M;
                END;
          MEANSSR1 = SUMSSR / NSTUD;
          MEANINDR1 = SUMINDR / NSTUD;
          PCTPF125R = SUMPF1R / NSTUD;
          PCTPF150R = SUMPF2R / NSTUD;
          PCTPF190R = SUMPF3R / NSTUD;
          PCTPF110R = SUMPF4R / NSTUD;
          GTSSR1 = (465 - MEANSSR1) / 12;
          GTINDR1 = (3.5 - MEANINDR1) / 12;
          GTPF125R = (100 - PCTPF125R) / 10;
          GTPF150R = (100 - PCTPF150R) / 10;
          GTPF190R = (100 - PCTPF190R) / 10;
          GTPF110R = (100 - PCTPF110R) / 10;

          MEANSSM1 = SUMSSM / NSTUD;
          MEANINDM1 = SUMINDM / NSTUD;
          PCTPF125M = SUMPF1M / NSTUD;
          PCTPF150M = SUMPF2M / NSTUD;
          PCTPF190M = SUMPF3M / NSTUD;
          PCTPF110M = SUMPF4M / NSTUD;
          GTSSM1 = (465 - MEANSSM1) / 12;
          GTINDM1 = (3.5 - MEANINDM1) / 12;
          GTPF125M = (100 - PCTPF125M) / 10;
          GTPF150M = (100 - PCTPF150M) / 10;
          GTPF190M = (100 - PCTPF190M) / 10;
          GTPF110M = (100 - PCTPF110M) / 10;

  *YEAR 2;

*Initialize sums for school;
      SUMSSR = 0;
          SUMINDR = 0;
          SUMPF1R = 0;
          SUMPF2R = 0;
          SUMPF3R = 0;
          SUMPF4R = 0;
          SUMSSM = 0;
          SUMINDM = 0;
          SUMPF1M = 0;
          SUMPF2M = 0;
          SUMPF3M = 0;
          SUMPF4M = 0;

          DO REP = 1 TO NSTUD;

*Randomly choose NSTUD students for each school and compute their scores in reading and math;
         Z3 = NORMAL(0);
                Z4 = NORMAL(0);
                RSQ = STURMCORR**2;
            READSS = READ + SQRT(VAROWS)*Z3 + INCREMENTR;
                MATHZ = STURMCORR * Z3 + SQRT (1-RSQ) * Z4;
                MATHSS = MATH + SQRT(VAROWS)*MATHZ + INCREMENTM;

                IF READSS < 232 THEN READIND = 1;
                IF 232 <= READSS < 300 THEN READIND = 2;
                IF 300 <= READSS < 428 THEN READIND = 3;
                IF 428 <= READSS THEN READIND = 4;
                IF READIND = 1 THEN PF25R = 0; ELSE PF25R = 100;
                IF READIND <= 2 THEN PF50R = 0; ELSE PF50R = 100;
                IF READIND <= 3 THEN PF90R = 0; ELSE PF90R = 100;
                IF READSS < 172 THEN PF10R = 0; ELSE PF10R = 100;
                SUMSSR = SUMSSR + READSS;
                SUMINDR = SUMINDR + READIND;
                SUMPF1R = SUMPF1R + PF25R;
                SUMPF2R = SUMPF2R + PF50R;
                SUMPF3R = SUMPF3R + PF90R;
                SUMPF4R = SUMPF4R + PF10R;

                IF MATHSS < 232 THEN MATHIND = 1;
                IF 232 <= MATHSS < 300 THEN MATHIND = 2;
                IF 300 <= MATHSS < 428 THEN MATHIND = 3;
                IF 428 <= MATHSS THEN MATHIND = 4;
                IF MATHIND = 1 THEN PF25M = 0; ELSE PF25M = 100;
                IF MATHIND <= 2 THEN PF50M = 0; ELSE PF50M = 100;
                IF MATHIND <= 3 THEN PF90M = 0; ELSE PF90M = 100;
                IF MATHSS < 172 THEN PF10M = 0; ELSE PF10M = 100;
                SUMSSM = SUMSSM + MATHSS;
                SUMINDM = SUMINDM + MATHIND;
                SUMPF1M = SUMPF1M + PF25M;
                SUMPF2M = SUMPF2M + PF50M;
                SUMPF3M = SUMPF3M + PF90M;
                SUMPF4M = SUMPF4M + PF10M;
                END;
          MEANSSR2 = SUMSSR / NSTUD;
          MEANINDR2 = SUMINDR / NSTUD;
          PCTPF225R = SUMPF1R / NSTUD;
          PCTPF250R = SUMPF2R / NSTUD;
          PCTPF290R = SUMPF3R / NSTUD;
          PCTPF210R = SUMPF4R / NSTUD;
          GTSSR2 = (465 - MEANSSR2) / 12;
          GTINDR2 = (3.5 - MEANINDR2) / 12;
          GTPF225R = (100 - PCTPF225R) / 10;
          GTPF250R = (100 - PCTPF250R) / 10;
          GTPF290R = (100 - PCTPF290R) / 10;
          GTPF210R = (100 - PCTPF210R) / 10;

          MEANSSM2 = SUMSSM / NSTUD;
          MEANINDM2 = SUMINDM / NSTUD;
          PCTPF225M = SUMPF1M / NSTUD;
          PCTPF250M = SUMPF2M / NSTUD;
          PCTPF290M = SUMPF3M / NSTUD;
          PCTPF210M = SUMPF4M / NSTUD;
          GTSSM2 = (465 - MEANSSM2) / 12;
          GTINDM2 = (3.5 - MEANINDM2) / 12;
          GTPF225M = (100 - PCTPF225M) / 10;
          GTPF250M = (100 - PCTPF250M) / 10;
          GTPF290M = (100 - PCTPF290M) / 10;
          GTPF210M = (100 - PCTPF210M) / 10;


      IF ((MEANSSR2 - MEANSSR1) < GTSSR1) OR ((MEANSSM2 - MEANSSM1) < GTSSM1)
                    THEN AYPSS1 = 100; ELSE AYPSS1 = 0;
          IF ((MEANINDR2 - MEANINDR1) < GTINDR1) OR ((MEANINDM2 - MEANINDM1) < GTINDM1)
                    THEN AYPIND1 = 100; ELSE AYPIND1 = 0;
          IF ((PCTPF225R - PCTPF125R) < GTPF125R) OR ((PCTPF225M - PCTPF125M) < GTPF125M)
                    THEN AYPPF251 = 100; ELSE AYPPF251 = 0;
          IF ((PCTPF250R - PCTPF150R) < GTPF150R) OR ((PCTPF250M - PCTPF150M) < GTPF150M)
                    THEN AYPPF501 = 100; ELSE AYPPF501 = 0;
          IF ((PCTPF290R - PCTPF190R) < GTPF190R) OR ((PCTPF290M - PCTPF190M) < GTPF190M)
                    THEN AYPPF901 = 100; ELSE AYPPF901 = 0;
          IF ((PCTPF210R - PCTPF110R) < GTPF110R) OR ((PCTPF210M - PCTPF110M) < GTPF110M)
                    THEN AYPPF101 = 100; ELSE AYPPF101 = 0;
          OUTPUT;
    END;
PROC FREQ;
```

```
    TABLES AYPSS1 AYPIND1 AYPPF251 AYPPF501 AYPPF901 AYPPF101;
RUN;
```

Study 12—Computing the probability of classification errors over two years, using two content areas: Part 1—When schools make no progress

```
TITLE 'MANUFACTURED DATA--N=20, RATIO = .84';
TITLE2 'PROBABILITY OF FAILING TWO YEARS IN A ROW WHEN SCHOOL MAKES NO PROGRESS--TWO CONTENT
AREAS';

DATA MANUSTUD;
  NSTUD=20;
  INCREMENT = 0;
  VARTBAR0 = 1600 - 1000/NSTUD;
  VARTWS = (9000 - VARTBAR0)*NSTUD/(NSTUD - 1);
  VARTBAR = VARTBAR0 - VARTWS/NSTUD;
  SDTBAR = SQRT (VARTBAR);
  VAROWS = 1000 + VARTWS;

  DO I = 1 TO 100000;
*Choose two normal random variables to create school true mean for reading and math;
    Z1 = NORMAL(0);
       Z2 = NORMAL(0);
*Set the correlation of true means between reading and math at the school level;
       SCHRMCORR = .95;
*Set the correlation between reading and math observed scores at the student level within school;
    STURMCORR = .75;

*Select a random school and compute its true mean reading and math scores;
    READ = SDTBAR * Z1 + 300;
       RSQ = SCHRMCORR**2;
       MATHZ = SCHRMCORR*Z1 + SQRT(1-RSQ) * Z2;
       MATH = SDTBAR * MATHZ + 300;
    INCREMENTR = 0;
    INCREMENTM = 0;

  *YEAR 1;

*Initialize sums for school;
      SUMSSR = 0;
          SUMINDR = 0;
          SUMPF1R = 0;
          SUMPF2R = 0;
          SUMPF3R = 0;
          SUMPF4R = 0;
          SUMSSM = 0;
          SUMINDM = 0;
          SUMPF1M = 0;
          SUMPF2M = 0;
          SUMPF3M = 0;
          SUMPF4M = 0;

          DO REP = 1 TO NSTUD;

*Randomly choose NSTUD students for each school and compute their scores in reading and math;
         Z3 = NORMAL(0);
                Z4 = NORMAL(0);
                RSQ = STURMCORR**2;
             READSS = READ + SQRT(VAROWS)*Z3;
                MATHZ = STURMCORR * Z3 + SQRT (1-RSQ) * Z4;
                MATHSS = MATH + SQRT(VAROWS)*MATHZ;

                IF READSS < 232 THEN READIND = 1;
                IF 232 <= READSS < 300 THEN READIND = 2;
                IF 300 <= READSS < 428 THEN READIND = 3;
                IF 428 <= READSS THEN READIND = 4;
                IF READIND = 1 THEN PF25R = 0; ELSE PF25R = 100;
                IF READIND <= 2 THEN PF50R = 0; ELSE PF50R = 100;
                IF READIND <= 3 THEN PF90R = 0; ELSE PF90R = 100;
                IF READSS < 172 THEN PF10R = 0; ELSE PF10R = 100;
                SUMSSR = SUMSSR + READSS;
                SUMINDR = SUMINDR + READIND;
                SUMPF1R = SUMPF1R + PF25R;
                SUMPF2R = SUMPF2R + PF50R;
                SUMPF3R = SUMPF3R + PF90R;
                SUMPF4R = SUMPF4R + PF10R;

                IF MATHSS < 232 THEN MATHIND = 1;
                IF 232 <= MATHSS < 300 THEN MATHIND = 2;
                IF 300 <= MATHSS < 428 THEN MATHIND = 3;
                IF 428 <= MATHSS THEN MATHIND = 4;
                IF MATHIND = 1 THEN PF25M = 0; ELSE PF25M = 100;
                IF MATHIND <= 2 THEN PF50M = 0; ELSE PF50M = 100;
                IF MATHIND <= 3 THEN PF90M = 0; ELSE PF90M = 100;
                IF MATHSS < 172 THEN PF10M = 0; ELSE PF10M = 100;
                SUMSSM = SUMSSM + MATHSS;
                SUMINDM = SUMINDM + MATHIND;
                SUMPF1M = SUMPF1M + PF25M;
                SUMPF2M = SUMPF2M + PF50M;
                SUMPF3M = SUMPF3M + PF90M;
                SUMPF4M = SUMPF4M + PF10M;
                END;
          MEANSSR1 = SUMSSR / NSTUD;
          MEANINDR1 = SUMINDR / NSTUD;
          PCTPF125R = SUMPF1R / NSTUD;
          PCTPF150R = SUMPF2R / NSTUD;
          PCTPF190R = SUMPF3R / NSTUD;
          PCTPF110R = SUMPF4R / NSTUD;
          GTSSR1 = (465 - MEANSSR1) / 12;
          GTINDR1 = (3.5 - MEANINDR1) / 12;
          GTPF125R = (100 - PCTPF125R) / 10;
          GTPF150R = (100 - PCTPF150R) / 10;
          GTPF190R = (100 - PCTPF190R) / 10;
          GTPF110R = (100 - PCTPF110R) / 10;

          MEANSSM1 = SUMSSM / NSTUD;
          MEANINDM1 = SUMINDM / NSTUD;
          PCTPF125M = SUMPF1M / NSTUD;
          PCTPF150M = SUMPF2M / NSTUD;
          PCTPF190M = SUMPF3M / NSTUD;
          PCTPF110M = SUMPF4M / NSTUD;
          GTSSM1 = (465 - MEANSSM1) / 12;
          GTINDM1 = (3.5 - MEANINDM1) / 12;
          GTPF125M = (100 - PCTPF125M) / 10;
          GTPF150M = (100 - PCTPF150M) / 10;
          GTPF190M = (100 - PCTPF190M) / 10;
          GTPF110M = (100 - PCTPF110M) / 10;

  *YEAR 2;

*Initialize sums for school;
      SUMSSR = 0;
          SUMINDR = 0;
          SUMPF1R = 0;
          SUMPF2R = 0;
          SUMPF3R = 0;
          SUMPF4R = 0;
          SUMSSM = 0;
          SUMINDM = 0;
          SUMPF1M = 0;
          SUMPF2M = 0;
          SUMPF3M = 0;
          SUMPF4M = 0;

          DO REP = 1 TO NSTUD;

*Randomly choose NSTUD students for each school and compute their scores in reading and math;
         Z3 = NORMAL(0);
                Z4 = NORMAL(0);
                RSQ = STURMCORR**2;
             READSS = READ + SQRT(VAROWS)*Z3 + INCREMENTR;
                MATHZ = STURMCORR * Z3 + SQRT (1-RSQ) * Z4;
                MATHSS = MATH + SQRT(VAROWS)*MATHZ + INCREMENTM;
```

```
              IF READSS < 232 THEN READIND = 1;
              IF 232 <= READSS < 300 THEN READIND = 2;
              IF 300 <= READSS < 428 THEN READIND = 3;
              IF 428 <= READSS THEN READIND = 4;
              IF READIND = 1 THEN PF25R = 0; ELSE PF25R = 100;
              IF READIND <= 2 THEN PF50R = 0; ELSE PF50R = 100;
              IF READIND <= 3 THEN PF90R = 0; ELSE PF90R = 100;
              IF READSS < 172 THEN PF10R = 0; ELSE PF10R = 100;
              SUMSSR = SUMSSR + READSS;
              SUMINDR = SUMINDR + READIND;
              SUMPF1R = SUMPF1R + PF25R;
              SUMPF2R = SUMPF2R + PF50R;
              SUMPF3R = SUMPF3R + PF90R;
              SUMPF4R = SUMPF4R + PF10R;

              IF MATHSS < 232 THEN MATHIND = 1;
              IF 232 <= MATHSS < 300 THEN MATHIND = 2;
              IF 300 <= MATHSS < 428 THEN MATHIND = 3;
              IF 428 <= MATHSS THEN MATHIND = 4;
              IF MATHIND = 1 THEN PF25M = 0; ELSE PF25M = 100;
              IF MATHIND <= 2 THEN PF50M = 0; ELSE PF50M = 100;
              IF MATHIND <= 3 THEN PF90M = 0; ELSE PF90M = 100;
              IF MATHSS < 172 THEN PF10M = 0; ELSE PF10M = 100;
              SUMSSM = SUMSSM + MATHSS;
              SUMINDM = SUMINDM + MATHIND;
              SUMPF1M = SUMPF1M + PF25M;
              SUMPF2M = SUMPF2M + PF50M;
              SUMPF3M = SUMPF3M + PF90M;
              SUMPF4M = SUMPF4M + PF10M;
              END;
       MEANSSR2 = SUMSSR / NSTUD;
       MEANINDR2 = SUMINDR / NSTUD;
       PCTPF225R = SUMPF1R / NSTUD;
       PCTPF250R = SUMPF2R / NSTUD;
       PCTPF290R = SUMPF3R / NSTUD;
       PCTPF210R = SUMPF4R / NSTUD;
       GTSSR2 = (465 - MEANSSR2) / 11;
       GTINDR2 = (3.5 - MEANINDR2) / 11;
       GTPF225R = (100 - PCTPF225R) / 10;
       GTPF250R = (100 - PCTPF250R) / 10;
       GTPF290R = (100 - PCTPF290R) / 10;
       GTPF210R = (100 - PCTPF210R) / 10;

       MEANSSM2 = SUMSSM / NSTUD;
       MEANINDM2 = SUMINDM / NSTUD;
       PCTPF225M = SUMPF1M / NSTUD;
       PCTPF250M = SUMPF2M / NSTUD;
       PCTPF290M = SUMPF3M / NSTUD;
       PCTPF210M = SUMPF4M / NSTUD;
       GTSSM2 = (465 - MEANSSM2) / 11;
       GTINDM2 = (3.5 - MEANINDM2) / 11;
       GTPF225M = (100 - PCTPF225M) / 10;
       GTPF250M = (100 - PCTPF250M) / 10;
       GTPF290M = (100 - PCTPF290M) / 10;
       GTPF210M = (100 - PCTPF210M) / 10;

    IF ((MEANSSR2 - MEANSSR1) < GTSSR1) OR ((MEANSSM2 - MEANSSM1) < GTSSM1)
                  THEN AYPSS1 = 100; ELSE AYPSS1 = 0;
       IF ((MEANINDR2 - MEANINDR1) < GTINDR1) OR ((MEANINDM2 - MEANINDM1) < GTINDM1)
                  THEN AYPIND1 = 100; ELSE AYPIND1 = 0;
       IF ((PCTPF225R - PCTPF125R) < GTPF125R) OR ((PCTPF225M - PCTPF125M) < GTPF125M)
                  THEN AYPPF251 = 100; ELSE AYPPF251 = 0;
       IF ((PCTPF250R - PCTPF150R) < GTPF150R) OR ((PCTPF250M - PCTPF150M) < GTPF150M)
                  THEN AYPPF501 = 100; ELSE AYPPF501 = 0;
       IF ((PCTPF290R - PCTPF190R) < GTPF190R) OR ((PCTPF290M - PCTPF190M) < GTPF190M)
                  THEN AYPPF901 = 100; ELSE AYPPF901 = 0;
       IF ((PCTPF210R - PCTPF110R) < GTPF110R) OR ((PCTPF210M - PCTPF110M) < GTPF110M)
                  THEN AYPPF101 = 100; ELSE AYPPF101 = 0;

*YEAR 3;
```

```
*Initialize sums for school;
      SUMSSR = 0;
          SUMINDR = 0;
          SUMPF1R = 0;
          SUMPF2R = 0;
          SUMPF3R = 0;
          SUMPF4R = 0;
          SUMSSM = 0;
          SUMINDM = 0;
          SUMPF1M = 0;
          SUMPF2M = 0;
          SUMPF3M = 0;
          SUMPF4M = 0;

          DO REP = 1 TO NSTUD;

*Randomly choose NSTUD students for each school and compute their scores in reading and math;
         Z3 = NORMAL(0);
                Z4 = NORMAL(0);
                RSQ = STURMCORR**2;
             READSS = READ + SQRT(VAROWS)*Z3 + 2*INCREMENTR;
                MATHZ = STURMCORR * Z3 + SQRT (1-RSQ) * Z4;
                MATHSS = MATH + SQRT(VAROWS)*MATHZ + 2*INCREMENTM;

                IF READSS < 232 THEN READIND = 1;
                IF 232 <= READSS < 300 THEN READIND = 2;
                IF 300 <= READSS < 428 THEN READIND = 3;
                IF 428 <= READSS THEN READIND = 4;
                IF READIND = 1 THEN PF25R = 0; ELSE PF25R = 100;
                IF READIND <= 2 THEN PF50R = 0; ELSE PF50R = 100;
                IF READIND <= 3 THEN PF90R = 0; ELSE PF90R = 100;
                IF READSS < 172 THEN PF10R = 0; ELSE PF10R = 100;
                SUMSSR = SUMSSR + READSS;
                SUMINDR = SUMINDR + READIND;
                SUMPF1R = SUMPF1R + PF25R;
                SUMPF2R = SUMPF2R + PF50R;
                SUMPF3R = SUMPF3R + PF90R;
                SUMPF4R = SUMPF4R + PF10R;

                IF MATHSS < 232 THEN MATHIND = 1;
                IF 232 <= MATHSS < 300 THEN MATHIND = 2;
                IF 300 <= MATHSS < 428 THEN MATHIND = 3;
                IF 428 <= MATHSS THEN MATHIND = 4;
                IF MATHIND = 1 THEN PF25M = 0; ELSE PF25M = 100;
                IF MATHIND <= 2 THEN PF50M = 0; ELSE PF50M = 100;
                IF MATHIND <= 3 THEN PF90M = 0; ELSE PF90M = 100;
                IF MATHSS < 172 THEN PF10M = 0; ELSE PF10M = 100;
                SUMSSM = SUMSSM + MATHSS;
                SUMINDM = SUMINDM + MATHIND;
                SUMPF1M = SUMPF1M + PF25M;
                SUMPF2M = SUMPF2M + PF50M;
                SUMPF3M = SUMPF3M + PF90M;
                SUMPF4M = SUMPF4M + PF10M;
                END;
          MEANSSR3 = SUMSSR / NSTUD;
          MEANINDR3 = SUMINDR / NSTUD;
          PCTPF325R = SUMPF1R / NSTUD;
          PCTPF350R = SUMPF2R / NSTUD;
          PCTPF390R = SUMPF3R / NSTUD;
          PCTPF310R = SUMPF4R / NSTUD;

          MEANSSM3 = SUMSSM / NSTUD;
          MEANINDM3 = SUMINDM / NSTUD;
          PCTPF325M = SUMPF1M / NSTUD;
          PCTPF350M = SUMPF2M / NSTUD;
          PCTPF390M = SUMPF3M / NSTUD;
          PCTPF310M = SUMPF4M / NSTUD;

      IF ((MEANSSR3 - MEANSSR2) < GTSSR2) OR ((MEANSSM3 - MEANSSM2) < GTSSM2)
                    THEN AYPSS2 = 100; ELSE AYPSS2 = 0;
          IF ((MEANINDR3 - MEANINDR2) < GTINDR2) OR ((MEANINDM3 - MEANINDM2) < GTINDM2)
                    THEN AYPIND2 = 100; ELSE AYPIND2 = 0;
          IF ((PCTPF325R - PCTPF225R) < GTPF225R) OR ((PCTPF325M - PCTPF225M) < GTPF225M)
                    THEN AYPPF252 = 100; ELSE AYPPF252 = 0;
          IF ((PCTPF350R - PCTPF250R) < GTPF250R) OR ((PCTPF350M - PCTPF250M) < GTPF250M)
                    THEN AYPPF502 = 100; ELSE AYPPF502 = 0;
          IF ((PCTPF390R - PCTPF290R) < GTPF290R) OR ((PCTPF390M - PCTPF290M) < GTPF290M)
                    THEN AYPPF902 = 100; ELSE AYPPF902 = 0;
          IF ((PCTPF310R - PCTPF210R) < GTPF210R) OR ((PCTPF310M - PCTPF210M) < GTPF210M)
                    THEN AYPPF102 = 100; ELSE AYPPF102 = 0;
       TWOYRSSS = AYPSS1 * AYPSS2/100;
          TWOYRSIND = AYPIND1 * AYPIND2/100;
          TWOYRSPF25 = AYPPF251*AYPPF252/100;
       TWOYRSPF50 = AYPPF501*AYPPF502/100;
       TWOYRSPF90 = AYPPF901*AYPPF902/100;
       TWOYRSPF10 = AYPPF101*AYPPF102/100;
          OUTPUT;
     END;
 PROC FREQ;
     TABLES TWOYRSSS--TWOYRSPF10;

 RUN;
```

Study 13—Computing the probability of classification errors over two years, using two content areas: Part 2—When schools make substantial progress

```
TITLE 'MANUFACTURED DATA--N=20, RATIO = .84';
TITLE2 'PROBABILITY OF FAILING TWO YEARS IN A ROW WHEN SCHOOL MAKES PROGRESS--TWO CONTENT AREAS';

DATA MANUSTUD;
  NSTUD=20;
  INCREMENT = 0;
  VARTBAR0 = 1600 - 1000/NSTUD;
  VARTWS = (9000 - VARTBAR0)*NSTUD/(NSTUD - 1);
  VARTBAR = VARTBAR0 - VARTWS/NSTUD;
  SDTBAR = SQRT (VARTBAR);
  VAROWS = 1000 + VARTWS;

  DO I = 1 TO 100000;
*Choose two normal random variables to create school true mean for reading and math;
    Z1 = NORMAL(0);
        Z2 = NORMAL(0);
*Set the correlation of true means between reading and math at the school level;
        SCHRMCORR = .95;
*Set the correlation between reading and math observed scores at the student level within school;
    STURMCORR = .75;

*Select a random school and compute its true mean reading and math scores;
    READ = SDTBAR * Z1 + 300;
        RSQ = SCHRMCORR**2;
        MATHZ = SCHRMCORR*Z1 + SQRT(1-RSQ) * Z2;
        MATH = SDTBAR * MATHZ + 300;
    INCREMENTR = (465 - READ)/ 12;
    INCREMENTM = (465 - MATH)/ 12;

  *YEAR 1;

*Initialize sums for school;
      SUMSSR = 0;
          SUMINDR = 0;
          SUMPF1R = 0;
          SUMPF2R = 0;
          SUMPF3R = 0;
          SUMPF4R = 0;
          SUMSSM = 0;
          SUMINDM = 0;
          SUMPF1M = 0;
          SUMPF2M = 0;
          SUMPF3M = 0;
          SUMPF4M = 0;

          DO REP = 1 TO NSTUD;

*Randomly choose NSTUD students for each school and compute their scores in reading and math;
         Z3 = NORMAL(0);
                Z4 = NORMAL(0);
                RSQ = STURMCORR**2;
             READSS = READ + SQRT(VAROWS)*Z3;
                MATHZ = STURMCORR * Z3 + SQRT (1-RSQ) * Z4;
                MATHSS = MATH + SQRT(VAROWS)*MATHZ;

                IF READSS < 232 THEN READIND = 1;
                IF 232 <= READSS < 300 THEN READIND = 2;
                IF 300 <= READSS < 428 THEN READIND = 3;
                IF 428 <= READSS THEN READIND = 4;
                IF READIND = 1 THEN PF25R = 0; ELSE PF25R = 100;
                IF READIND <= 2 THEN PF50R = 0; ELSE PF50R = 100;
                IF READIND <= 3 THEN PF90R = 0; ELSE PF90R = 100;
                IF READSS < 172 THEN PF10R = 0; ELSE PF10R = 100;
                SUMSSR = SUMSSR + READSS;
                SUMINDR = SUMINDR + READIND;
                SUMPF1R = SUMPF1R + PF25R;
                SUMPF2R = SUMPF2R + PF50R;
                SUMPF3R = SUMPF3R + PF90R;
                SUMPF4R = SUMPF4R + PF10R;

                IF MATHSS < 232 THEN MATHIND = 1;
                IF 232 <= MATHSS < 300 THEN MATHIND = 2;
                IF 300 <= MATHSS < 428 THEN MATHIND = 3;
                IF 428 <= MATHSS THEN MATHIND = 4;
                IF MATHIND = 1 THEN PF25M = 0; ELSE PF25M = 100;
                IF MATHIND <= 2 THEN PF50M = 0; ELSE PF50M = 100;
                IF MATHIND <= 3 THEN PF90M = 0; ELSE PF90M = 100;
                IF MATHSS < 172 THEN PF10M = 0; ELSE PF10M = 100;
                SUMSSM = SUMSSM + MATHSS;
                SUMINDM = SUMINDM + MATHIND;
                SUMPF1M = SUMPF1M + PF25M;
                SUMPF2M = SUMPF2M + PF50M;
                SUMPF3M = SUMPF3M + PF90M;
                SUMPF4M = SUMPF4M + PF10M;
                END;
         MEANSSR1 = SUMSSR / NSTUD;
         MEANINDR1 = SUMINDR / NSTUD;
         PCTPF125R = SUMPF1R / NSTUD;
         PCTPF150R = SUMPF2R / NSTUD;
         PCTPF190R = SUMPF3R / NSTUD;
         PCTPF110R = SUMPF4R / NSTUD;
         GTSSR1 = (465 - MEANSSR1) / 12;
         GTINDR1 = (3.5 - MEANINDR1) / 12;
         GTPF125R = (100 - PCTPF125R) / 10;
         GTPF150R = (100 - PCTPF150R) / 10;
         GTPF190R = (100 - PCTPF190R) / 10;
         GTPF110R = (100 - PCTPF110R) / 10;

         MEANSSM1 = SUMSSM / NSTUD;
         MEANINDM1 = SUMINDM / NSTUD;
         PCTPF125M = SUMPF1M / NSTUD;
         PCTPF150M = SUMPF2M / NSTUD;
         PCTPF190M = SUMPF3M / NSTUD;
         PCTPF110M = SUMPF4M / NSTUD;
         GTSSM1 = (465 - MEANSSM1) / 12;
         GTINDM1 = (3.5 - MEANINDM1) / 12;
         GTPF125M = (100 - PCTPF125M) / 10;
         GTPF150M = (100 - PCTPF150M) / 10;
         GTPF190M = (100 - PCTPF190M) / 10;
         GTPF110M = (100 - PCTPF110M) / 10;

  *YEAR 2;

*Initialize sums for school;
      SUMSSR = 0;
         SUMINDR = 0;
         SUMPF1R = 0;
         SUMPF2R = 0;
         SUMPF3R = 0;
         SUMPF4R = 0;
         SUMSSM = 0;
         SUMINDM = 0;
         SUMPF1M = 0;
         SUMPF2M = 0;
         SUMPF3M = 0;
         SUMPF4M = 0;

         DO REP = 1 TO NSTUD;

*Randomly choose NSTUD students for each school and compute their scores in reading and math;
         Z3 = NORMAL(0);
                Z4 = NORMAL(0);
                RSQ = STURMCORR**2;
            READSS = READ + SQRT(VAROWS)*Z3 + INCREMENTR;
                MATHZ = STURMCORR * Z3 + SQRT (1-RSQ) * Z4;
                MATHSS = MATH + SQRT(VAROWS)*MATHZ + INCREMENTM;
```

```
              IF READSS < 232 THEN READIND = 1;
              IF 232 <= READSS < 300 THEN READIND = 2;
              IF 300 <= READSS < 428 THEN READIND = 3;
              IF 428 <= READSS THEN READIND = 4;
              IF READIND = 1 THEN PF25R = 0; ELSE PF25R = 100;
              IF READIND <= 2 THEN PF50R = 0; ELSE PF50R = 100;
              IF READIND <= 3 THEN PF90R = 0; ELSE PF90R = 100;
              IF READSS < 172 THEN PF10R = 0; ELSE PF10R = 100;
              SUMSSR = SUMSSR + READSS;
              SUMINDR = SUMINDR + READIND;
              SUMPF1R = SUMPF1R + PF25R;
              SUMPF2R = SUMPF2R + PF50R;
              SUMPF3R = SUMPF3R + PF90R;
              SUMPF4R = SUMPF4R + PF10R;

              IF MATHSS < 232 THEN MATHIND = 1;
              IF 232 <= MATHSS < 300 THEN MATHIND = 2;
              IF 300 <= MATHSS < 428 THEN MATHIND = 3;
              IF 428 <= MATHSS THEN MATHIND = 4;
              IF MATHIND = 1 THEN PF25M = 0; ELSE PF25M = 100;
              IF MATHIND <= 2 THEN PF50M = 0; ELSE PF50M = 100;
              IF MATHIND <= 3 THEN PF90M = 0; ELSE PF90M = 100;
              IF MATHSS < 172 THEN PF10M = 0; ELSE PF10M = 100;
              SUMSSM = SUMSSM + MATHSS;
              SUMINDM = SUMINDM + MATHIND;
              SUMPF1M = SUMPF1M + PF25M;
              SUMPF2M = SUMPF2M + PF50M;
              SUMPF3M = SUMPF3M + PF90M;
              SUMPF4M = SUMPF4M + PF10M;
              END;
       MEANSSR2 = SUMSSR / NSTUD;
       MEANINDR2 = SUMINDR / NSTUD;
       PCTPF225R = SUMPF1R / NSTUD;
       PCTPF250R = SUMPF2R / NSTUD;
       PCTPF290R = SUMPF3R / NSTUD;
       PCTPF210R = SUMPF4R / NSTUD;
       GTSSR2 = (465 - MEANSSR2) / 11;
       GTINDR2 = (3.5 - MEANINDR2) / 11;
       GTPF225R = (100 - PCTPF225R) / 10;
       GTPF250R = (100 - PCTPF250R) / 10;
       GTPF290R = (100 - PCTPF290R) / 10;
       GTPF210R = (100 - PCTPF210R) / 10;

       MEANSSM2 = SUMSSM / NSTUD;
       MEANINDM2 = SUMINDM / NSTUD;
       PCTPF225M = SUMPF1M / NSTUD;
       PCTPF250M = SUMPF2M / NSTUD;
       PCTPF290M = SUMPF3M / NSTUD;
       PCTPF210M = SUMPF4M / NSTUD;
       GTSSM2 = (465 - MEANSSM2) / 11;
       GTINDM2 = (3.5 - MEANINDM2) / 11;
       GTPF225M = (100 - PCTPF225M) / 10;
       GTPF250M = (100 - PCTPF250M) / 10;
       GTPF290M = (100 - PCTPF290M) / 10;
       GTPF210M = (100 - PCTPF210M) / 10;

    IF ((MEANSSR2 - MEANSSR1) < GTSSR1) OR ((MEANSSM2 - MEANSSM1) < GTSSM1)
                  THEN AYPSS1 = 100; ELSE AYPSS1 = 0;
       IF ((MEANINDR2 - MEANINDR1) < GTINDR1) OR ((MEANINDM2 - MEANINDM1) < GTINDM1)
                  THEN AYPIND1 = 100; ELSE AYPIND1 = 0;
       IF ((PCTPF225R - PCTPF125R) < GTPF125R) OR ((PCTPF225M - PCTPF125M) < GTPF125M)
                  THEN AYPPF251 = 100; ELSE AYPPF251 = 0;
       IF ((PCTPF250R - PCTPF150R) < GTPF150R) OR ((PCTPF250M - PCTPF150M) < GTPF150M)
                  THEN AYPPF501 = 100; ELSE AYPPF501 = 0;
       IF ((PCTPF290R - PCTPF190R) < GTPF190R) OR ((PCTPF290M - PCTPF190M) < GTPF190M)
                  THEN AYPPF901 = 100; ELSE AYPPF901 = 0;
       IF ((PCTPF210R - PCTPF110R) < GTPF110R) OR ((PCTPF210M - PCTPF110M) < GTPF110M)
                  THEN AYPPF101 = 100; ELSE AYPPF101 = 0;

*YEAR 3;
```

```
*Initialize sums for school;
      SUMSSR = 0;
         SUMINDR = 0;
         SUMPF1R = 0;
         SUMPF2R = 0;
         SUMPF3R = 0;
         SUMPF4R = 0;
         SUMSSM = 0;
         SUMINDM = 0;
         SUMPF1M = 0;
         SUMPF2M = 0;
         SUMPF3M = 0;
         SUMPF4M = 0;

         DO REP = 1 TO NSTUD;

*Randomly choose NSTUD students for each school and compute their scores in reading and math;
         Z3 = NORMAL(0);
                Z4 = NORMAL(0);
                RSQ = STURMCORR**2;
            READSS = READ + SQRT(VAROWS)*Z3 + 2*INCREMENTR;
                MATHZ = STURMCORR * Z3 + SQRT (1-RSQ) * Z4;
                MATHSS = MATH + SQRT(VAROWS)*MATHZ + 2*INCREMENTM;

                IF READSS < 232 THEN READIND = 1;
                IF 232 <= READSS < 300 THEN READIND = 2;
                IF 300 <= READSS < 428 THEN READIND = 3;
                IF 428 <= READSS THEN READIND = 4;
                IF READIND = 1 THEN PF25R = 0; ELSE PF25R = 100;
                IF READIND <= 2 THEN PF50R = 0; ELSE PF50R = 100;
                IF READIND <= 3 THEN PF90R = 0; ELSE PF90R = 100;
                IF READSS < 172 THEN PF10R = 0; ELSE PF10R = 100;
                SUMSSR = SUMSSR + READSS;
                SUMINDR = SUMINDR + READIND;
                SUMPF1R = SUMPF1R + PF25R;
                SUMPF2R = SUMPF2R + PF50R;
                SUMPF3R = SUMPF3R + PF90R;
                SUMPF4R = SUMPF4R + PF10R;

                IF MATHSS < 232 THEN MATHIND = 1;
                IF 232 <= MATHSS < 300 THEN MATHIND = 2;
                IF 300 <= MATHSS < 428 THEN MATHIND = 3;
                IF 428 <= MATHSS THEN MATHIND = 4;
                IF MATHIND = 1 THEN PF25M = 0; ELSE PF25M = 100;
                IF MATHIND <= 2 THEN PF50M = 0; ELSE PF50M = 100;
                IF MATHIND <= 3 THEN PF90M = 0; ELSE PF90M = 100;
                IF MATHSS < 172 THEN PF10M = 0; ELSE PF10M = 100;
                SUMSSM = SUMSSM + MATHSS;
                SUMINDM = SUMINDM + MATHIND;
                SUMPF1M = SUMPF1M + PF25M;
                SUMPF2M = SUMPF2M + PF50M;
                SUMPF3M = SUMPF3M + PF90M;
                SUMPF4M = SUMPF4M + PF10M;
                END;
         MEANSSR3 = SUMSSR / NSTUD;
         MEANINDR3 = SUMINDR / NSTUD;
         PCTPF325R = SUMPF1R / NSTUD;
         PCTPF350R = SUMPF2R / NSTUD;
         PCTPF390R = SUMPF3R / NSTUD;
         PCTPF310R = SUMPF4R / NSTUD;

         MEANSSM3 = SUMSSM / NSTUD;
         MEANINDM3 = SUMINDM / NSTUD;
         PCTPF325M = SUMPF1M / NSTUD;
         PCTPF350M = SUMPF2M / NSTUD;
         PCTPF390M = SUMPF3M / NSTUD;
         PCTPF310M = SUMPF4M / NSTUD;

      IF ((MEANSSR3 - MEANSSR2) < GTSSR2) OR ((MEANSSM3 - MEANSSM2) < GTSSM2)
                    THEN AYPSS2 = 100; ELSE AYPSS2 = 0;
         IF ((MEANINDR3 - MEANINDR2) < GTINDR2) OR ((MEANINDM3 - MEANINDM2) < GTINDM2)
```

```
                    THEN AYPIND2 = 100; ELSE AYPIND2 = 0;
          IF ((PCTPF325R - PCTPF225R) < GTPF225R) OR ((PCTPF325M - PCTPF225M) < GTPF225M)
                    THEN AYPPF252 = 100; ELSE AYPPF252 = 0;
          IF ((PCTPF350R - PCTPF250R) < GTPF250R) OR ((PCTPF350M - PCTPF250M) < GTPF250M)
                    THEN AYPPF502 = 100; ELSE AYPPF502 = 0;
          IF ((PCTPF390R - PCTPF290R) < GTPF290R) OR ((PCTPF390M - PCTPF290M) < GTPF290M)
                    THEN AYPPF902 = 100; ELSE AYPPF902 = 0;
          IF ((PCTPF310R - PCTPF210R) < GTPF210R) OR ((PCTPF310M - PCTPF210M) < GTPF210M)
                    THEN AYPPF102 = 100; ELSE AYPPF102 = 0;
      TWOYRSSS = AYPSS1 * AYPSS2/100;
          TWOYRSIND = AYPIND1 * AYPIND2/100;
          TWOYRSPF25 = AYPPF251*AYPPF252/100;
      TWOYRSPF50 = AYPPF501*AYPPF502/100;
      TWOYRSPF90 = AYPPF901*AYPPF902/100;
      TWOYRSPF10 = AYPPF101*AYPPF102/100;
          OUTPUT;
    END;
PROC FREQ;
    TABLES TWOYRSSS--TWOYRSPF10;

RUN;
```

Study 14 - Draw random samples of 63,000 records numbered 1 through 63,000 with replacement. Use record numbers generated from sample base file to draw student records from the original student-level data file.

```
DATA SAMPLE;

 STUDENTS=63000;
 DO FILENO = 1 to 3;
   DO I= 1 to STUDENTS;
     A = UNIFORM(0);
     RECORD = ROUND((A*X) + .5);
     SAMPLE = FILENO;
     OUTPUT;
     END;
 END;

 PROC SORT;
   BY RECORD;
 RUN;

DATA ORIGINAL;
 SET SASUSER.STUDENT01;
 IF _N_=1 THEN RECORD=0;
 IF GRADE=4 AND ELA_PL^=' ' AND MATHLEVEL^=' ' THEN RECORD+1;
    ELSE DELETE;
 KEEP SCHCODE GRADE RACE SWD LEP GENDER FREE ELA_SCALE MATH_SCALE ELA_PL
      MATH_PL RECORD ;
RUN;


***Merge files to create a data set containing the three sample files;
DATA SASUSER.SAMP3;
 MERGE ORIGINAL SAMPLE (IN=CHK);
  BY RECORD;
  IF CHK;

 KEEP SCHCODE GRADE RACE SWD LEP GENDER ECONOMIC ELA_SCALE MATH_SCALE
      ELA_PL MATH_PL RECORD SAMPLE;
```

Study 15 – Compute school scores from sample student files. Compare three sample files to make AYP computations

```
DATA SAMPLE;
SET SASUSER.SAMP3 ;
PROC SORT;
 BY SCHCODE;
RUN;

DATA BASEMEAN;
 SET SAMPLE;
  IF SAMPLE=1;
PROC MEANS NOPRINT;
 VAR MATH_SCALE ELA_SCALE;
 BY SCHCODE;
 OUTPUT OUT=BASEMN_YR1 MEAN=MATHSS_1 ELASS_1 VAR=MATHSS_V1 ELASS_V1 N=NSTUDENT;
RUN;

DATA STUDENTS;
 MERGE SAMPLE BASEMN_YR1;
 BY SCHCODE;

 IF SAMPLE=2 THEN MATH_SCALE = MATH_SCALE + (400 - MATHSS_1)/12; ELSE
 IF SAMPLE=3 THEN MATH_SCALE = MATH_SCALE + 2*((400-MATHSS_1)/12);

 IF SAMPLE=2 THEN ELA_SCALE  = ELA_SCALE + (380 - ELASS_1)/12; ELSE
 IF SAMPLE=3 THEN ELA_SCALE  = ELA_SCALE + 2*((380 - ELASS_1)/12);


ELA_SCALE = ROUND(ELA_SCALE);
MATH_SCALE = ROUND(MATH_SCALE);

IF ELA_SCALE>500 THEN ELA_SCALE=500;
IF MATH_SCALE>500 THEN MATH_SCALE=500;


***ASSIGN MATHEMATICS VALUES;
  IF MATH_SCALE<250 then MP10=0; else MP10=1;
  IF MATH_SCALE<284 then MP25=0; else MP25=1;
  IF MATH_SCALE<318 then MP50=0; else MP50=1;
  IF MATH_SCALE<378 then MP90=0; else MP90=1;

  IF 100<=MATH_SCALE<=281 then MATH_PL='BEL'; ELSE
  IF 282<= MATH_SCALE <=314 then MATH_PL='BAS'; ELSE
  IF 315<= MATH_SCALE <=369 then MATH_PL='PRO'; ELSE
  IF 370<= MATH_SCALE <=418 then MATH_PL='ADV’; ELSE
  IF 419<= MATH_SCALE <=500 then MATH_PL='DIS';


  IF MATH_PL ='DIS' THEN MATH_IND=5; else
  IF MATH_PL ='ADV' THEN MATH_IND=4; else
  IF MATH_PL ='PRO' then MATH_IND=3; else
  IF MATH_PL ='BAS' then MATH_IND=2; else
  IF MATH_PL ='BEL' then MATH_IND=1;

***ASSIGN ENGLISH LANGUAGE ARTS VALUES;
  IF ELA_SCALE<242 then EP10=0; else EP10=1;
  IF ELA_SCALE<278 then EP25=0; else EP25=1;
  IF ELA_SCALE<312 then EP50=0; else EP50=1;
  IF ELA_SCALE<364 then EP90=0; else EP90=1;

  IF 100<=ELA_SCALE<=262 then ELA_PL='UNS'; ELSE
  IF 263<=ELA_SCALE<=300 then ELA_PL='APP'; ELSE
  IF 301<=ELA_SCALE<=353 then ELA_PL='BAS'; ELSE
  IF 354<=ELA_SCALE<=407 then ELA_PL='PRO'; ELSE
  IF 408<=ELA_SCALE<=500 then ELA_PL='ADV';
  IF ELA_PL='ADV' THEN ELA_IND=5; else
  IF ELA_PL='PRO' THEN ELA_IND=4; else
  IF ELA_PL='BAS' then ELA_IND=3; else
  IF ELA_PL='APP' then ELA_IND=2; else
  IF ELA_PL='UNS' then ELA_IND=1;


KEEP SCHCODE SAMPLE MP10 MP25 MP50 MP90 MATH_IND MATH_SCALE
                    EP10 EP25 EP50 EP90 ELA_IND ELA_SCALE

PROC MEANS;
   VAR MP10 MP25 MP50 MP90 MATH_IND MATH_SCALE
       EP10 EP25 EP50 EP90 ELA_IND  ELA_SCALE
   CLASS SAMPLE;
   RUN;


DATA YEAR1;
 SET STUDENTS;
 IF SAMPLE=1;
  PROC MEANS NOPRINT;
   VAR MP10 MP25 MP50 MP90 MATH_IND MATH_SCALE
       EP10 EP25 EP50 EP90 ELA_IND  ELA_SCALE;
  CLASS SCHCODE;
  OUTPUT OUT = SCH_YR1  MEAN= MP10_1 MP25_1 MP50_1 MP90_1 MATH_IND_1 MPASS_1 MATHSS_1
            EP10_1 EP25_1 EP50_1 EP90_1 ELA_IND_1 EPASS_1 ELASS_1 N = NSTUD_1;
  RUN;

DATA YEAR2;
 SET STUDENTS;
 IF SAMPLE=2;
   PROC MEANS NOPRINT;
    VAR MP10 MP25 MP50 MP90 MATH_IND MATH_SCALE
        EP10 EP25 EP50 EP90 ELA_IND  ELA_SCALE;
    CLASS SCHCODE;
    OUTPUT OUT = SCH_YR2  MEAN= MP10_2 MP25_2 MP50_2 MP90_2 MATH_IND_2 MPASS_2 MATHSS_2
              EP10_2 EP25_2 EP50_2 EP90_2 ELA_IND_2 EPASS_2 ELASS_2  N = NSTUD_2;

RUN;
DATA YEAR3;
 SET STUDENTS;
  IF SAMPLE=3;
PROC MEANS NOPRINT;
   VAR MP10 MP25 MP50 MP90 MATH_IND MATH_SCALE
       EP10 EP25 EP50 EP90 ELA_IND  ELA_SCALE;
  CLASS SCHCODE;
  OUTPUT OUT = SCH_YR3  MEAN= MP10_3 MP25_3 MP50_3 MP90_3 MATH_IND_3 MPASS_3 MATHSS_3
           EP10_3 EP25_3 EP50_3 EP90_3 ELA_IND_3 EPASS_3 ELASS_3 N = NSTUD_3;
 RUN;

DATA SASUSER.OVERALL;
 MERGE SCH_YR1 SCH_YR2 SCH_YR3;
 BY SCHCODE;
 IF SCHCODE^=' ';

**** At this point, school scores in reading and mathematics have been generated for each
reporting statistic in years 1, 2, and 3.  The remainder of the program to determine whether
schools have made adequate improvement, met the status requirement, or made AYP each year is
identical to the rule applied in the programs for Study 9, Study 12, or Study 13.;
```

Study 16-19: Computing the probability of classification errors over two years for two content areas, including subgroups: Part 4 – When the status and improvement are considered

The previous program generated 24 yes/no improvement and status decisions for each school across the 6 reporting statistics, 2 content areas, and 2 years. Those decisions were stored as two 24-element arrays under the following structure

Array Elements 1-6: Math Year 1 to Year 2 on Scaled Score (1), Index (2), pi=.10 (3), pi=.25 (4), pi=.50 (5), pi=.90 (6)

Array Elements 7-12: Math Year 2 to Year 3 on Scaled Score (7), Index (8), pi=.10 (9), pi=.25 (10), pi=.50 (11), pi=.90 (12)

Array Elements 13-18: Reading Year 1 to Year 2 on Scaled Score (13), Index (14), pi=.10 (15), pi=.25 (16), pi=.50 (17), pi=.90 (18)

Array Elements 19-24: Reading Year 2 to Year 3 on Scaled Score (19), Index (20), pi=.10 (21), pi=.25 (22), pi=.50 (23), pi=.90 (24)

The current program combines the overall school and subgroup data sets and addresses the series of pass/fail questions discussed in Study 19. The program can be modified easily to accommodate different minimum sample sizes. The program can also be modified to examine improvement or status separately by overriding the answers to the appropriate pass/fail questions.

```
DATA OVERALL;
  SET SASUSER.OVERALL;
  ARRAY IMP{24};
  ARRAY STATUS{24};
  ARRAY OVERALL_IMP{24};
  ARRAY OVERALL_STAT{24};
    DO I = 1 TO 24;
        OVERALL_IMP{I} = IMP{I};
        OVERALL_STAT{I} = STATUS{I};
    END;
  OVERN=NSTUD_2;
DATA SWD;
  SET SASUSER.SWD;
  ARRAY IMP{24};
  ARRAY STATUS{24};
  ARRAY SWD_IMP{24};
  ARRAY SWD_STAT{24};
    DO I = 1 TO 24;
        SWD_IMP{I} = IMP{I};
        SWD_STATUS = STATUS{I};
    END;
  SWDN = NSTUD_2;
DATA ECON;
  SET SASUSER.ECON;
  ARRAY IMP{24};
  ARRAY STATUS{24};
  ARRAY ECON_IMP{24};
  ARRAY ECON_STAT{24};
    DO I = 1 TO 24;
        ECON_IMP{I} = IMP{I};
        ECON_STAT{I} = STATUS{I};
    END;
  ECONN = NSTUD_2;
DATA BLACK;
  SET SASUSER.BLACK;
  ARRAY IMP{24};
  ARRAY STATUS{24};
  ARRAY BLACK_IMP{24};
  ARRAY BLACK_STAT{24};
    DO I = 1 TO 24;
        BLACK_IMP{I} = IMP{I};
        BLACK_STAT_{I} = STATUS{I};
    END;
  BLACKN = NSTUD_2;
DATA WHITE;
```

```
  SET SASUSER.WHITE;
  ARRAY IMP{24};
  ARRAY STATUS{24};
  ARRAY WHITE_IMP{24};
  ARRAY WHITE_STAT{24};
    DO I = 1 TO 24;
        WHITE_IMP{I} = IMP{I};
        WHITE_STAT{I} = STATUS{I};
    END;
  WHITEN = NSTUD_2;


DATA SUBGROUPS;
 MERGE  OVERALL SWD ECON BLACK WHITE;
  BY  SCHCODE;


***Scaled Score
***Year 1;
***Math;
IF OVER_STAT1=1 THEN m1=1; ELSE m1=0;
IF OVER_IMP1=1 THEN m2=1; ELSE m2=0;
*IF OVERN<30 THEN m3=1; ELSE m3=0;

IF SWD_STAT1=1 THEN m4=1; ELSE m4=0;
IF SWD_IMP1=1 THEN m5=1; ELSE m5=0;
IF SWDN<30 THEN m6=1; ELSE m6=0;

IF ECON_STAT1=1 THEN m7=1; ELSE m7=0;
IF ECON_IMP1=1 THEN m8=1; ELSE m8=0;
IF ECONN<30 THEN m9=1; ELSE m9=0;

IF BLACK_STAT1=1 THEN m10=1; ELSE m10=0;
IF BLACK_IMP1=1 THEN m11=1; ELSE m11=0;
IF BLACKN<30 THEN m12=1; ELSE m12=0;

IF WHITE_STAT1=1 THEN m13=1; ELSE m13=0;
IF WHITE_IMP1=1 THEN m14=1; ELSE m14=0;
IF WHITEN<30 THEN m15=1; ELSE m15=0;

***English;
IF OVER_STAT7=1 THEN m16=1; ELSE m16=0;
IF OVER_IMP7=1 THEN m17=1; ELSE m17=0;
*IF OVERN<30 THEN m18=1; ELSE m18=0;

IF SWD_STAT7=1 THEN m19=1; ELSE m19=0;
IF SWD_IMP7=1 THEN m20=1; ELSE m20=0;
IF SWDN<30 THEN m21=1; ELSE m21=0;

IF ECON_STAT7=1 THEN m22=1; ELSE m22=0;
IF ECON_IMP7=1 THEN m23=1; ELSE m23=0;
IF ECONN<30 THEN m24=1; ELSE m24=0;

IF BLACK_STAT7=1 THEN m25=1; ELSE m25=0;
IF BLACK_IMP7=1 THEN m26=1; ELSE m26=0;
IF BLACKN<30 THEN m27=1; ELSE m27=0;

IF WHITE_STAT7=1 THEN m28=1; ELSE m28=0;
IF WHITE_IMP7=1 THEN m29=1; ELSE m29=0;
IF WHITEN<30 THEN m30=1; ELSE m30=0;


IF SUM(m1,m2,m3)>0 THEN a1=1; ELSE a1=0;
IF SUM(m4,m5,m6)>0 THEN a2=1; ELSE a2=0;
IF SUM(m7,m8,m9)>0 THEN a3=1; ELSE a3=0;
IF SUM(m10,m11,m12)>0 THEN a4=1; ELSE a4=0;
IF SUM(m13,m14,m15)>0 THEN a5=1; ELSE a5=0;
IF SUM(m16,m17,m18)>0 THEN a6=1; ELSE a6=0;
IF SUM(m19,m20,m21)>0 THEN a7=1; ELSE a7=0;
IF SUM(m22,m23,m24)>0 THEN a8=1; ELSE a8=0;
IF SUM(m25,m26,m27)>0 THEN a9=1; ELSE a9=0;
IF SUM(m28,m29,m30)>0 THEN a10=1; ELSE a10=0;

IF SUM(of a1-a10)=10 THEN PASSYR1=1 ; ELSE PASSYR1=0;
```

```
***Year 2;
***Math;
IF OVER_STAT13=1 THEN mm1=1; ELSE mm1=0;
IF OVER_IMP13=1 THEN mm2=1; ELSE mm2=0;
*IF OVERN<30 THEN mm3=1; ELSE mm3=0;

IF SWD_STAT13=1 THEN mm4=1; ELSE mm4=0;
IF SWD_IMP13=1 THEN mm5=1; ELSE mm5=0;
IF SWDN<30 THEN mm6=1; ELSE mm6=0;

IF ECON_STAT13=1 THEN mm7=1; ELSE mm7=0;
IF ECON_IMP13=1) THEN mm8=1; ELSE mm8=0;
IF ECONN<30 THEN mm9=1; ELSE mm9=0;

IF BLACKSTAT13=1 THEN mm10=1; ELSE mm10=0;
IF BLACK_IMP13=1 THEN mm11=1; ELSE mm11=0;
IF BLACKN<30 THEN mm12=1; ELSE mm12=0;

IF WHITE_STAT13=1 THEN mm13=1; ELSE mm13=0;
IF WHITE_IMP13=1 THEN mm14=1; ELSE mm14=0;
IF WHITEN<30 THEN mm15=1; ELSE mm15=0;

***English;
IF OVER_STAT19=1 THEN mm16=1; ELSE mm16=0;
IF OVER_IMP19=1 THEN mm17=1; ELSE mm17=0;
*IF OVERN<30 THEN mm18=1; ELSE mm18=0;

IF SWD_STAT19=1 THEN mm19=1; ELSE mm19=0;
IF SWD_IMP19=1) THEN mm20=1; ELSE mm20=0;
IF SWDN<30 THEN mm21=1; ELSE mm21=0;

IF ECON_STAT19=1 THEN mm22=1; ELSE mm22=0;
IF ECON19_IMP19=1 THEN mm23=1; ELSE mm23=0;
IF ECONN<30 THEN mm24=1; ELSE mm24=0;

IF BLACK_STAT19=1 THEN mm25=1; ELSE mm25=0;
IF BLACK_IMP19=1) THEN mm26=1; ELSE mm26=0;
IF BLACKN<30 THEN mm27=1; ELSE mm27=0;

IF WHITESTAT19=1 THEN mm28=1; ELSE mm28=0;
IF WHITE_IMP19=1) THEN mm29=1; ELSE mm29=0;
IF WHITEN<30 THEN mm30=1; ELSE mm30=0;


IF SUM(mm1,mm2,mm3)>0 THEN b1=1; ELSE b1=0;
IF SUM(mm4,mm5,mm6)>0 THEN b2=1; ELSE b2=0;
IF SUM(mm7,mm8,mm9)>0 THEN b3=1; ELSE b3=0;
IF SUM(mm10,mm11,mm12)>0 THEN b4=1; ELSE b4=0;
IF SUM(mm13,mm14,mm15)>0 THEN b5=1; ELSE b5=0;
IF SUM(mm16,mm17,mm18)>0 THEN b6=1; ELSE b6=0;
IF SUM(mm19,mm20,mm21)>0 THEN b7=1; ELSE b7=0;
IF SUM(mm22,mm23,mm24)>0 THEN b8=1; ELSE b8=0;
IF SUM(mm25,mm26,mm27)>0 THEN b9=1; ELSE b9=0;
IF SUM(mm28,mm29,mm30)>0 THEN b10=1; ELSE b10=0;

IF SUM(of b1-b10)=10 THEN PASSYR2=1; ELSE PASSYR2=0;

IF PASSYR1=1 or PASSYR2=1 THEN IDTWICE=0; ELSE
 IF PASSYR1=0 and PASSYR2=0 THEN IDTWICE=1; ELSE
  IDTWICE=.;

PROC FREQ;
  TABLES IDTWICE ;
RUN;
```

Identical code can be used to determine AYP for each of the remaining reporting statistics.