Reliability of No Child Left Behind Accountability Designs

Richard K. Hill and Charles A. DePascale
The National Center for the Improvement of Educational Assessment, Inc.

February 7, 2003

Abstract

The No Child Left Behind Act of 2001 requires states to establish accountability systems that are both valid and reliable. If one follows the language of the law literally, there is no design that will meet both requirements. If one interprets the law more flexibly, it is possible to create such a design. States will need to approach the problem carefully if they are going to appropriately balance the various probabilities of making incorrect decisions about schools.

Background

Each state must create an accountability system in response to the requirements of the No Child Left Behind Act of 2001 (NCLB). Through these accountability systems, states will determine whether schools have made "adequate yearly progress (AYP)." These decisions are required to be both "statistically valid and reliable." [20 USC 6311 §1111(b)(2)(C)(ii)]

The first thing to note is that *assessment* systems are not *accountability* systems. Assessment refers to the process of collecting information; accountability refers to the process of making decisions and applying consequences based on the information collected.

An accountability system cannot be valid without valid assessment data. If the assessment data are off-target, one cannot create a set of decision rules that will lead to accurate judgments about schools. But having valid assessment data is not sufficient to ensure a valid accountability system. For example, if schools are supposed to be judged on the amount students have learned during a year, but decisions are based solely on the scores of students at the end of the year without regard for what they knew at the beginning of the year, the accountability design will be invalid even if the assessments are perfectly valid. A fairly complete discussion of the issues associated with the validity of accountability systems is provided in Chapter 2 of "Making Valid and Reliable Decisions in the Determination of Adequate Yearly Progress" (Marion, White, et al., 2002). This paper therefore will not delve further into the validity of accountability systems, but focus instead on reliability.

Sources of Unreliability

School scores vary from year to year. Even if the curriculum and instruction provided to students across years is identical and even if the community from which students come remains constant across years, 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[1]. 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 later in 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 based largely on random error.

Some authors have questioned whether sampling error should be taken into account when considering school mean scores. Their logic is that since all the students within the school are being tested, there is no student sampling. Linn, Baker and Betebenner (2002) have responded to this position in detail, but the following points also are germane:

1. When the results are reported, they are not attributed to a particular group of students, but to the school as a whole. 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. 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 is tested 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 population to which the inference was being limited would never be the same across those years[2].

2. It would be inconsistent logic to take measurement error into account and not sampling error. That is, if one is going to take measurement error into account in determining a school's classification, 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]?

[1] 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] Thanks to Dale Carlson for originally raising this issue.

[3] Thanks to Bob Linn for originally raising this issue.

3. Perhaps the strongest argument for taking sampling error into account is not based in theory or logic, however, but by simply observing the fact that including sampling error still 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 categories and then subtract out "persistent effects." The remaining variance, the "non-persistent" variance, is more than would be expected from sampling error and measurement error alone. We conducted a comparable analysis of data from a New England state and found the same result; even though the statewide results did not change (implying that persistent effects were minimal) the variation in school mean scores from year to year was greater than one would predict from sampling and measurement error alone.

That issue leads us to the final point of this section. Because the effectiveness of instruction *does* vary in schools from year to year, and it varies in different, *unknown* amounts, the actual reliability of an accountability design cannot be determined by simply looking at the actual changes in test scores across years. Such an analysis will understate the true reliability of an accountability design. To correctly determine the reliability of school means over time, one must work from a model of school performance, but not the actual data across years.

Reliability of Student-Level Results vs. Reliability of School-Level Results

While accountability systems cannot be valid if the assessment data are not valid, it is interesting to note that an accountability system *can* be reliable even if the assessment results for individual students have only modest reliability. Table 1 is duplicated from "Examining the Reliability of Accountability Systems." (Hill, 2002) These results are not derived from any particular state, but are results that might be expected from a typical state (detail in that paper provides specifics about the assumptions made in creating Table 1). The first column provides the number of students tested in a school; the second column lists the reliability of individual student scores; and the last column yields the reliability of the school mean. As can be readily seen from the table, the number of students in the school has, by far, more impact on the reliability of the school mean than does the reliability of the individual student results. For example, if a school has 50 students, the reliability of the school mean will be higher if the reliability of individual student results is only .60 than when the school has 25 students and the reliability of individual student results is .90.

The implication of this for the design of accountability systems is substantial. Suppose, for example, one had a limited budget for testing, and had to make the choice of giving a short test to several grades of students or giving one long test to students at just one grade. If the purpose of the program is to estimate the school mean, it is clear that giving the shorter test to more students would be the better choice. Similarly, suppose one had to choose between two tests with different administration costs. Suppose the first test has high validity but also high costs, so that only a few items can be given to each student, and as a result, the student-level reliability of the test is low. In contrast, the second test has lower validity, but because its costs are low, students can take sufficient items to yield high student-level reliability. Again, if the purpose of the program is to estimate the school mean, it is clear that the better choice, by far, is the test with higher validity and lower student-level reliability.

Table 1

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

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

Determining Decision Consistency of NCLB Designs

Although we have used the term "reliability" to this point, reliability coefficients are not the actual statistic of interest. What is important is the probability of classification errors. If a school is making Adequate Yearly Progress, it should be labeled as such; the converse is true as well. The probability that these decisions are being made accurately is what is truly meant when the term "reliability" is used loosely.

There are at least four methods that can be used to estimate the probability that schools will be correctly classified:

1. Direct computation. Calculating exact probabilities when the distribution of errors can be determined through standard equations.
2. Split-half. Dividing the available data for each school into equivalent halves through random assignment and calculating the percentage of times a consistent decision is made on the two halves.
3. Random draws with replacement. Simulating the effects of random sampling by making repeated draws with replacement from the available data set.
4. Monte Carlo. Modeling the distribution of scores and creating samples through the use of random number generators.

Each of these methods is described in some detail in "Determining the Reliability of School Scores," (Hill and DePascale, 2002). Each method has strengths and weaknesses. The Direct Computation method, for example, is most desirable, since it provides exact results; but the calculations required are often too complex to make the method practical. The Split-half method is easy to implement, but can give misleading results if certain assumptions are violated. Random-draws with replacement is a fairly easy method to implement, but can give misleading results for very small schools. The Monte Carlo method requires realistic modeling of underlying distributions

and a powerful computer (although the latest generation of personal computers has been more than capable of handling all our analysis needs to date).

Because there are a large number of places where one can go wrong in implementing these studies, it is advisable to employ at least two of the methods when doing a study. Having the results for a second method confirm the findings of the first greatly decreases the likelihood of making an error.

NCLB requires states to include certain elements in their accountability design. While the actual designs will vary from state to state, each is supposed to determine whether a school as a whole and all subgroups within the school either have achieved a particular percentage of students at the proficient level or higher (met the "status" requirement) or have improved their percentage of students achieving at the proficient level or higher over the prior year's level (met the "improvement" requirement, also often referred to as "safe harbor"). If the school or any of its subgroups fails both those tests, the school fails to make Adequate Yearly Progress (AYP), and certain actions are taken against the school. Results for subgroups are not required to be included "in a case in which the number of students in a category is insufficient to yield statistically reliable information…" [20 USC 6311 §1111(b)(2)(C)(v)(II)] States are left to determine what that number might be.

One issue to be addressed is how low reliability can go before it is "insufficient." If the stakes are low, a fairly low level of reliability might be acceptable. If the stakes are high, however, one would want to be fairly certain that a school had been correctly classified before applying the prescribed consequences to the school. In NCLB, annual judgments are made about whether a school has made AYP. If a school fails to make AYP two years in a row, a series of rather drastic consequences begin. So, unquestionably, one would want the decision about whether a school had failed to make AYP two years in a row to be highly reliable. But being identified as a "failing" school even for one year could have serious negative consequences for a school, so a reasonable argument can be constructed for wanting a reliable decision to be made every year for every school.

Classifications of schools on the basis of status can be done fairly reliably, even with small numbers of students. Hill and DePascale (2002) showed that even for schools (or subgroups) with as few as 20 students, schools were correctly classified around 85 percent of the time. That is, if a school's true score (the percentage of proficient students if an infinite number of students from the school's catchment area were tested) was above the required amount, the observed score for the school usually was above that amount, too. The exact result varies depending upon the stringency of the standard established by the state: a standard that requires about 50 percent of the students to be passing produces a higher percentage of correct dassifications than a standard that is very high or very low (for much the same reason that a test consisting of items of moderate difficulty will be more reliable than a test with all easy or all hard items). When the number of students increased to 50, about 90 percent of the schools were correctly classified. If one has a very high standard for decision accuracy, these results might not be satisfactory. Too, it is worthwhile noting that decision accuracy will be lower for schools with a percentage of proficient students near the required score than those further away. Nonetheless, by most standards, fairly reliable decisions can be made about the status of schools even with fairly modest numbers of students per school (or per subgroup).

Unfortunately, the situation for measuring improvement is quite different. Sufficient improvement is defined by NCLB as reducing the percentage of students that are not proficient by 10 percent. So, for example, if a subgroup has 20 percent of its students proficient one year, the

subgroup makes AYP the next year if 28 percent or more of the students are proficient (reducing the non-proficient rate from 80 to 72 percent).

The expected value of a school's observed score is its true score; that is, we assume there is no systematic bias in sampling. So if a school's true score is 20 percent proficient, the expected value of the school's observed score is 20 percent. However, different samples of students will vary from that 20 percent figure—some will be higher, and some lower. The issue is, how much higher and how much lower, and with what frequency will those variations occur.

Let a school's true scores (expressed in percentages of students proficient) in Year 1 and Year 2 be p_1 and p_2 respectively, and the school's observed scores be p_1 and p_2. If there are N students tested in the school *each year*, the standard deviation of the difference scores will be

$$s_{p_2-p_1} = \sqrt{\frac{p_1(100-p_1)+p_2(100-p_2)}{N}} \qquad (1)$$

[4]See footnote

Now, suppose p_2 is higher enough than p_1 that the school truly did meet the AYP requirement for improvement. One possible rule for determining whether schools make AYP would be to require the school's observed percentage of non-proficient students decrease by 10 percent. But if a school's true score goes up a certain amount, the change in observed scores will be greater than that about half the time and less than that the other half. So, a rule that required observed scores to increase by the NCLB-required amount would, by definition, misclassify about half the schools that truly made that required amount of improvement. That appears, on its face, to be far too high an error rate.

Therefore, we must create a rule for which the likelihood of misclassification is less than 50/50. One possibility would be to say that a school (or subgroup) made AYP if its percentage of observed scores is simply higher in Year 2 than in Year 1 (not by 10 percent, but just any improvement whatsoever). Now our question would be, "If schools truly improved by the AYP required amount, what percentage of schools would be misidentified as having failed to make AYP if all a subgroup needed to do to make AYP was increase its observed score from one year to the next?" Restated, we will start by positing that every school (or subgroup) truly improves by the amount required by NCLB, and use a rule that a school fails AYP if its observed results in Year 2 are no higher than they were for Year 1. Under those conditions, what percentage of schools would be falsely labeled as not having made AYP?

[4] Equation 1 is accurate for one test and one content area, and under the assumption that the samples for the two years are independent. Under NCLB, a school has to meet the improvement criteria for both reading and mathematics; the probability of misclassifying on the basis of two tests will be somewhat higher than the probability for one test (just how much higher will be explored in a later section of this paper). If a school has multiple subgroups, the probability of misclassification might be considerably higher that for the calculations provided in this section, since the errors for each of the tests will accumulate. If a state has tests at adjacent grades, the likelihood of misclassification will *decrease*, since the sampling errors across years will tend to balance each other out (e.g., if the third graders this year are unusually high achieving, they likely will be again next year when they are included in the school average as fourth graders). However, if there is not testing at adjacent grades, the formula above is accurate. Beginning in 2006, when states are required to have tests at every grade from 3 to 8, there will be testing in adjacent grades; until then, however, many states have tests only at selected grades, and therefore will find the above equation accurate for one content area for one group within the school.

Table 2 provides the answer for schools of various sizes and starting points. In each case, the true score for the school in Year 2 has 10 percent fewer non-proficient students than in Year 1. Note that the probabilities take into account the discrete nature of the data with which we are dealing; if a school has 50 students for example, possible observed percentages of proficient students are 0, 2, 4, etc. Such a school cannot have 1.2 percent proficient students, for example. Therefore, the first column of results provides the probability that a school's observed score will not increase; the second column provides the probability that a school's observed score will be exactly the same, despite the improvement the school has made in its educational program. If a state decided to make a rule that a school's results were acceptable if the results in Year 2 were *the same as or greater than* the results for Year 1, the probability of misclassifying truly improving schools would be the difference of those two columns.

Table 2

Percent of Times Observed Scores Will Not Increase from Year 1 to Year 2 If the True Scores for a School or Subgroup Increase by the NCLB-Required Amount

True Score Year 1	True Score Year 2	Number of Students Tested	Percent of Times Observed Score in Year 2 will be Not be Greater than the Observed Score in Year 1	Percent of Times Observed Score in Year 2 will be Equal to the Observed Score in Year 1
10	19	50	12.6	5.0
		100	4.3	1.5
		200	0.6	0.2
20	28	50	20.5	6.0
		100	10.6	2.7
		200	3.4	0.8
30	37	50	26.2	6.4
		100	16.4	3.4
		200	7.6	1.4
40	46	50	30.7	6.7
		100	21.6	3.9
		200	12.2	1.9
50	55	50	34.4	7.0
		100	26.2	4.4
		200	17.0	2.4

From Table 2, we see that if a school (or subgroup) has 50 students per year, and the school's true score improves from 10 percent proficient in Year 1 to 19 percent in Year 2, observed scores still would either be the same or drop 12.6 percent of the time. Since they would stay the same 5.0 percent of the time, we know that they would actually drop 7.6 percent of the time, despite the improvement of the school.

If a school has a higher percentage of students passing in Year 1, the improvement goal is smaller, which means that the gain is smaller relative to the standard error of the difference scores.

As a result, the probability that a school will be misclassified increases as the Year 1 true score increases. If schools' true scores improved from 40 percent passing to 46, observed scores in Year 2 would be no higher than they were than Year 1 for over three-tenths of the schools with 50 students per grade.

Linn (2002) approached this same issue from a slightly different point of view. He showed that for a situation that might be typical (50 students tested in a subgroup each year, 50 percent of the students proficient), the standard error of the gain scores is *twice* the required amount of improvement. A minor extension of his calculations shows that in order to have the standard error equal the required gain, 200 students per year would need to be in the subgroup. To have a 95 percent confidence interval equal the required amount of improvement, *541* students would need to be in the subgroup each year.

Confirming These Results and Adding a Second Content Area[5]

Earlier in this paper, we suggested that researchers use at least two different methods for estimating reliability, since it is possible to make erroneous assumptions when doing studies of this nature. To confirm the results in Table 2 (which were generated through the Direct Computation Method), we replicated the study using the Monte Carlo method.

It was straightforward to adapt the computer programs used in an earlier study to replicate the analyses in this study. All we did was create a true score for a school so that the percentage of students proficient ("proficient" defined as scoring at the 50th percentile or above) would be equal to the amount posited for the school. That allowed us to posit a distribution of scores for each school that was normal, had a mean equal to the true score obtained above, and had a standard deviation equal to the standard deviation of students within school (the square root of the variance of students across schools minus the variance of school means). From each school's hypothesized distribution, we then drew a random sample of students of size *N* for that school, and made the appropriate calculations. This was done 10,000 times, so that the final result we had would not be likely to vary substantially under another replication of the study. However, by the very nature of a Monte Carlo study, the results will vary somewhat each time the study is run, since the scores are chosen at random. Since the assumptions behind drawing the reading and math scores were identical, the reader can get some idea of what the variability would be by comparing the last two columns of Table 3. Those are, in essence, two runs of the same study.

Table 3 provides the results of the Monte Carlo study. The results given in Table 2 are incorporated into Table 3 so that results of the "direct calculation" method (provided in Table 2) can be easily compared to the results of the Monte Carlo study.

As can be seen from Table 3, the results of the Monte Carlo method matched those of the direct computation method very closely. This is evidence that the results in Table 2 are not some statistical artifact, but reflect the amount of variation that really occurs in school scores.

A second reason for running the Monte Carlo study is that it permits us to easily calculate the probability of misclassifying a school when we require schools to improve both their reading and math scores. To keep the study simple, we presumed that a school's true score was the same in

[5] Readers wanting more detail on the issues discussed in this section are encouraged to reference "Determining the Reliability of School Scores" (Hill and DePascale, 2002).

reading and math, but when drawing samples of students, we allowed their scores to vary as they might typically do. That is, when we drew a particular random student for a school, we chose two correlated random variables, so that the correlation between reading and math scores at the student level in this study is the same as one would typical find in a state. That meant, as would be the case in real life, that sometimes we drew a student who passed reading and not math, and sometimes the reverse. As a result, although the true scores in reading and math for the school both increased by the same amount, sometimes a school would have its scores improve in reading but not math, and sometimes the reverse. To the extent this happens, the percentage of misclassified schools increases if we require that a school improve in both content areas (which is a requirement of NCLB).

Table 3

Percent of Times Observed Scores Will Not Increase from Year 1 to Year 2 If the True Scores for a School Increase by the NCLB-Required Amount, Comparing Direct Calculation with Monte Carlo

True Score Year 1	True Score Year 2	Number of Students Tested	Percent of Times Observed Score in Year 2 will Not be Greater than the Observed Score in Year 1		
			Direct Calculation	Monte Carlo—Reading	Monte Carlo--Math
10	19	50	12.6	12.8	12.3
		100	4.3	4.2	4.3
		200	0.6	0.6	0.7
20	28	50	20.5	19.7	20.0
		100	10.6	10.7	10.7
		200	3.4	3.1	3.4
30	37	50	26.2	25.6	26.1
		100	16.4	16.2	16.7
		200	7.6	7.3	7.6
40	46	50	30.7	31.0	30.3
		100	21.6	22.1	22.3
		200	12.2	12.2	12.7
50	55	50	34.4	34.1	33.8
		100	26.2	27.2	26.9
		200	17.0	16.9	17.1

The question of interest, of course, is how often that occurs. We already have seen that some proportion of schools that truly improve do not see their observed scores increase from one year to the next, and that for some situations, that proportion is uncomfortably high. The results in Table 3 provide those proportions for just one test—either reading or math. The results in Table 4 compare those results to those when we require that a school have increasing scores in both reading *and* mathematics.

The results in Table 4 are not encouraging. They show that probability of misclassifying a school, using our given decision rules, increases substantially when we require the school to have increasing scores in both reading and mathematics. To be dramatic, let's look at the worst case (the third row from the bottom in Table 4). Suppose a state had a group of schools that all had 50 students in one grade, had a true percentage-passing rate of 50 percent, and improved their true score

from Year 1 to Year 2 by the NCLB-required amount (5 percentage points). Suppose further that the state required schools to increase their observed percentage of students passing from one year to the next in *both* reading and mathematics in order to make AYP. In that case, even though every school *did* improve its true score by the amount required by NCLB, almost *half* would be labeled as failing. Even if the rule were applied only to schools with *200* students per grade, over one-fourth would be incorrectly labeled as not improving.

Table 4

Percent of Times Observed Scores Will Not Increase from Year 1 to Year 2 If the True Scores for a School or Subgroup Increase by the NCLB-Required Amount, Comparing One Test to Two

True Score Year 1	True Score Year 2	Number of Students Tested	Percent of Times Observed Score in Year 2 will not be Greater than the Observed Score in Year 1		
			Reading Only	Math Only	Reading *and* Math
10	19	50	12.8	12.3	20.8
		100	4.2	4.3	7.5
		200	0.6	0.7	1.2
20	28	50	19.7	20.0	31.7
		100	10.7	10.7	17.9
		200	3.1	3.4	5.9
30	37	50	25.6	26.1	39.0
		100	16.2	16.7	25.7
		200	7.3	7.6	12.6
40	46	50	31.0	30.3	44.5
		100	22.1	22.3	34.0
		200	12.2	12.7	20.4
50	55	50	34.1	33.8	48.8
		100	27.2	26.9	40.1
		200	16.9	17.1	26.9

NCLB requires schools (and each subgroup) to reach a standard in both reading and mathematics. Unsatisfactory performance in either content area will cause a school to not make AYP. This is called a "conjunctive" decision model. In the past, many states have averaged scores together across two or more content areas to arrive at a total score for a school, and then made a decision about the school on the basis of that overall score. That is a "compensatory" decision model. Compensatory decisions are more reliable than the best component, whereas conjunctive decisions are no more reliable than the worst component. Conjunctive decisions would have the error rate in the last column of Table 4; compensatory decisions would have error rates that are better than the ones in the "Reading Only" and "Math Only" columns. In short, the error rates for a conjunctive model will often be twice that for a compensatory model.

These misclassification rates will probably strike most people as unacceptably high. That is, for most of the entries in Tables 2 and 4, the probability of misclassifying a school that has truly improved the required amount, *even if our criterion for identification is simply requiring higher*

observed scores in Year 2 than Year 1, is greater than the standard rates of 5 or 1 percent that are typically applied in tests of statistical significance in educational research.

"Reasonable" Error Rate and Multiple Comparisons

There is legitimate debate over what a reasonable error rate is for identifying schools. Choosing an error rate that decreases the likelihood that schools will be incorrectly identified as not having made AYP when they truly have simultaneously increases the likelihood that schools that have not truly made AYP will not be identified. Therefore, one must accept the fact that errors in both directions will be made—the only question is how high an error rate is acceptable. In an earlier section, we referred to decisions with an error rate of 15 percent being "fairly reliable." Each state might have a different judgment about the accuracy of that label, depending on the consequences it assigns to identified schools and their tolerance for error.

Whatever error rate one might accept as tolerable, it needs to be noted that all the discussion so far references one judgment being made about a school. In fact, NCLB requires that several judgments are made about each school (one for reading and one for mathematics for each of several subgroups), and a school is identified as not having made AYP if any subgroup has inadequate performance in either content area. The error rate for a school across all these judgments is substantially higher than the error rate reported for any one decision. That issue must be taken into account when deciding what a reasonable error rate is for a school as a whole.

Selecting a Fixed N

Many states are taking the approach of requiring that a subgroup have a particular number of students in order to be included, regardless of the performance of the subgroup. A recent survey by the National Association of State Directors of Special Education (Markowitz, 2002) showed that 12 states had selected a minimum N at which they felt results for measuring improvement would be sufficiently reliable for measuring improvement; the choices ranged between 10 and 75, with a median of 30.

The idea of selecting one fixed number for deciding whether a subgroup's results are sufficiently reliable appears to be an approach that will not work well for either measuring status or improvement. If a certain fixed number is chosen, schools will not be directly accountable for subgroups with fewer than that number (those subgroups will be included in the school's total score, but the performance of that subgroup by itself will not be looked at). No matter how small a number is chosen, this will exclude many subgroups, leading to an incomplete look at the performance of the school. Thus, one could argue that a number like 30 is *far* too large a number—a requirement that subgroups meet this minimum N will eliminate the vast majority of subgroups in most schools.

On the other hand, the results for subgroups are supposed to be "statistically reliable." That would mean, at a minimum, that if a subgroup causes a school to fail AYP, another sample of students in that subgroup drawn for that school would be likely to produce the same result. While reasonably modest numbers of students often (but not always) can be used to reliably determine whether a subgroup has met the status requirement, it takes hundreds of students, as shown in the previous section, to reliably detect whether a school has made sufficient improvement. Note that both the status and improvement decisions for a school must be made correctly for the overall judgment about the school to be correct. But since the number of students required to make a reliable judgment about improvement is substantially higher than the number required to make

reliable judgments about status for most schools, it is the decisions about improvement that drive the reliability of the system as a whole, not the decisions about status.

So, a state should pick a fairly small N for purposes of validity (say, certainly something no larger than 10), but it would need a very high N for purposes of reliability (say, 300 or more). Obviously, a value that provides reasonable validity is wholly inadequate for reliability purposes; a value that provides reasonable reliability is wholly inadequate for validity purposes. A figure between those two is largely inadequate for *both* purposes. This is the reason states are having such a hard time choosing a fixed value for minimum N. At first blush, it appears that the problem is choosing a modest fixed N that is a reasonable compromise between reliability and validity; a careful look tells us that choosing any value is wholly inadequate for at least one of the two concerns, if not both. In short, there isn't a reasonable answer to this dilemma. One is not faced with a reasonable balancing of concerns over reliability and validity; any answer will be clearly wrong for at least one of the two.

Given that one cannot have validity without reliability, it would be justifiable for a state to select a minimum N of 300. Granted, an N of this size will eliminate virtually every subgroup in most schools, essentially eliminating this aspect of NCLB. But such an N would at least ensure that decisions would be sufficiently reliable.

An Alternative Approach to Measuring Status

An alternative to selecting a fixed N is to run a test of statistical significance. That way, subgroups that are far from the standard do not need to have a large N for a reliable decision to be made. For example, suppose the standard for a state is 50 percent proficient. If no students in a subgroup are proficient, a reliable decision (one that has less than a 1 percent probability of misclassifying the subgroup) that the subgroup fails the status test can be made if there are just seven students in the subgroup. That is, if 50 percent of the students in a subgroup are proficient, there is less than 1 chance out of 100 that no students a sample of seven would be proficient. Thus, in cases where results are extremely low, the inadequate performance of the subgroup can be reliably detected even with small Ns[6]. On the other hand, if 499 out of 1000 students were proficient, one would not be certain that another sample of students from that same subgroup wouldn't have at least 50 percent proficient. So, this system will select a group that is far away from the standard even if the group is small, but will not select a group that is close to the standard even if the group is quite large. Not only is this a better application of statistics than the fixed N approach, it also is more fair and valid. Certainly, one would want to identify and target resources to very low-achieving subgroups before doing the same to subgroups that are very close to the state's standard.

Alternative Approaches to Measuring Improvement

The numbers of students required to measure whether a school has improved, on the other hand, probably are considerably higher than most people would have expected before looking at Tables 2-4. Certainly, they are impractical. For most schools and subgroups, there is sufficient uncertainty around their scores that one rarely could state with assurance whether improvement truly took place. If a state decided to judge school improvement scores as acceptable in all cases except

[6] Note that a state would always need to set some minimum N (typically, 3-10) for reporting the results of a subgroup, to ensure the confidentiality of individual student results. The suggested test of statistical significance would be run only on those subgroups of sufficient size to meet the state's reporting requirements.

when there was clear evidence that improvement did not take place, virtually every school would be placed in that category. That would be the case if the "statistical significance" approach outlined for measuring status were applied to improvement. Few schools or subgroups would fail to have the null hypothesis rejected that they had not improved—the amount of uncertainty around improvement scores is simply too large. As a result, it would be possible for there to be minimal improvement year after year, but with the state still not identifying any schools as not having made AYP. Such a policy clearly would be circumventing the intent of NCLB. Therefore, this section offers some alternatives to consider.

Use a different alpha level. One approach would be to use a different alpha level in running the test of statistical significance on improvement. Whereas a state might be willing to use an alpha of .05 or .01 on a status test, it might want to use an alpha of .25 or even .50 for improvement. There might be two justifications for doing so:

1. Every decision has a certain probability of error. When running tests of statistical significance, there is some likelihood that the null hypothesis is true, and another that it is false. When we observe results, we make a decision whether to reject the null hypothesis or not. If the null hypothesis is true, there is some probability we will reject it. That is the alpha level of our significance test, and is referred to as Type I error. If the null hypothesis is false, there is some probability we will fail to reject it. The probability of failing to reject the null hypothesis when it is false is called Type II error.

   Both Type I and Type II errors have costs. If a school truly has met the AYP requirements of NCLB but gets identified because of the bad luck of a poor sample, that is Type I error. There are consequences applied to the school, and those are the Type I costs. If a school truly has not met the AYP requirements but still is judged as acceptable, not because of real improvement but because there is too much sampling error to detect that the improvement hasn't taken place, that is Type II error. Type II costs include students not being given the resources or options that they would have received had the school been (correctly) identified as not having met AYP.

   A good design estimates the costs of these errors and strikes a balance so that the total cost of errors is minimized. A design that uses too stringent an alpha level lowers Type I errors (and therefore Type I costs), but also increases Type II costs. The opposite is true if an alpha level is used that permits too many Type I errors.

   For a status design, decisions are reliable enough that a fairly stringent alpha level can be chosen without increasing Type II errors too much. But because there is so much uncertainty around improvement scores, an alpha level might be needed that limits the percentage of Type II errors to a reasonable amount (at the cost of making more Type I errors).

2. All the calculations provided in this paper assume unconditional probabilities. Suppose, however, that despite all the improvements expected in student performance from NCLB, the average performance across all schools in a state is the same one year as the previous year. If that were the case, it would be highly likely that few, if any, schools made any improvement, much less the 10 percent reduction in percent non-proficient that NCLB calls for. In that case, one could establish a set of prior probabilities about school improvement and compute Bayesian (conditional) probabilities of school improvement. If the prior probabilities were strong that improvement had not taken place, most of the schools in the state would be

judged as non-improving. If that were the case, one could arrive at the same point, using far less complex statistical reasoning, by simply setting a low alpha level.

Examine changes over a longer period of time. However, an even better approach might be to look at improvement over longer periods of time than one year. If NCLB is interpreted literally, a school is accountable for its change only from one year to the next. As has been noted repeatedly in this paper, the amount of improvement expected over the course of just one year is undetectable except for very large groups. However, if we examine improvement over a longer period of time, we can detect whether or not change has occurred for substantially smaller schools and subgroups.

Two alternatives are to look at improvement over two or three years, or to average data for one 2-year period and compare it to another. In this section, we outline those approaches and compare the results to those for the one-year-to-one-year approach that is seemingly required by NCLB and were provided in an earlier section on this paper.

(a) Measure improvement over two or three years. If a school is expected to reduce its percentage of non-proficient students by 10 percent a year, it is expected to reduce the percent of non-proficient students by 19 percent over two years, or 27.1 percent over three years. Detecting changes of that magnitude are considerably easier than a change of 10 percent. Tables 5 and 6 are based on the same premises as those for Table 2; the only difference is that Table 5 looks at the probability of misclassification if two years' of improvement is required, and Table 6 does the same for three years'.

As can be seen from the tables, the probability of misclassification decreases dramatically if there are two or three years of improvement expected. When we were trying to detect change over one year (the results in Table 2), large numbers of students were required to state with reasonable certainty that a school had not improved. In Table 5, but especially in Table 6, we see that a school that has not increased its scores can be determined to have not improved with reasonable certainty, often even when the number of students is fairly small. The bottom line is that detecting changes of 5 or even 10 percent cannot be done reliably—but detecting changes of 20 or 30 percent can. So the key to effective evaluation of schools is to establish situations where the required amount of improvement approaches those levels—and under NCLB, that is a period of two or three years for most schools, not one.

Table 5

Percent of Times Observed Scores Will Not Increase from Year 1 to Year 3 If the True Scores for a School or Subgroup Increase by the NCLB-Required Amount

True Score Year 1	True Score Year 2	Number of Students Tested	Percent of Times Observed Score in Year 2 will be Less than the Observed Score in Year 1	Percent of Times Observed Score in Year 2 will be Equal to the Observed Score in Year 1
10	27.1	50	1.7	0.8
		100	<0.1	<0.1
		200	<.0.1	<0.1
20	35.2	50	5.4	2.1
		100	0.9	0.3
		200	<0.1	<0.1
30	43.3	50	9.9	3.2
		100	2.9	0.9
		200	0.3	<0.1
40	51.4	50	4.7	4.2
		100	6.0	1.5
		200	1.2	0.2
50	59.5	50	19.6	5.1
		100	9.9	2.3
		200	3.1	0.6

Table 6

Percent of Times Observed Scores Will Not Increase from Year 1 to Year 4 If the True Scores for a School or Subgroup Increase by the NCLB-Required Amount

True Score Year 1	True Score Year 4	Number of Students Tested	Percent of Times Observed Score in Year 3 will be Less than the Observed Score in Year 1	Percent of Times Observed Score in Year 3 will be Equal to the Observed Score in Year 1
10	34.39	50	0.2	<0.1
		100	<0.1	<0.1
		200	<0.1	<0.1
20	41.68	50	1.1	0.5
		100	<0.1	<0.1
		200	<0.1	<0.1
30	48.97	50	3.0	1.2
		100	0.3	0.1
		200	<0.1	<0.1
40	56.26	50	6.1	2.1
		100	1.2	0.4
		200	<0.1	<0.1
50	63.55	50	10.1	3.1
		100	3.0	0.9
		200	0.3	<0.1

(b) Compare two-year averages. An alternative to the above approach is to average results together for two years for a school to establish a baseline, and then compare that result to another two-year average, requiring a 19 percent reduction in the percentage of non-proficient students (since the improvement takes place over a two-year period). This basic accountability system is already employed by several states. It reduces the error over the one-year-to-one-year system in two ways: (1) it requires that the amount of improvement be 19 percent, rather than 10 percent, which, as we saw in the above section, is substantially more reliable, and (2) doubles the number of students included both the baseline and the growth score. The results for this approach are the same as those in Table 5 except that the number of students is divided in two. So, for example, the probability that a school's observed score would not rise if its true percentage of proficient students went from 10 to 27.1 is .017 if there are 25 students per year in the school, rather than 50.

Use a different reporting statistic. In other papers (Hill {2002}, Hill and DePascale (2002}), we have shown that use of an index produces decisions of somewhat higher reliability than percentage of students passing. Much of the language of NCLB suggests that the reporting statistic should be the percentage of students at or above the state's level of proficiency. Yet it also prescribes setting more than one cutscore and says that AYP should be determined by "the progress" of students [20 USC 6311 §1111(b)(2)(C)(iv)]. This suggests that a more reliable reporting statistic than percentage of students passing might be acceptable.

Use a compensatory model for decision-making. As shown in an earlier section, a compensatory model is far more reliable than the conjunctive model specified by NCLB. A school that is relatively weak in both reading and mathematics might be within a reasonable margin of error in both reading and mathematics, and therefore would not be identified if a conjunctive model were used, but identified as not having made AYP under a compensatory model were used. States could examine the numbers and types of schools accurately identified under both types of models and make a decision to choose a compensatory model if that led to a more accurate system. This would more likely occur in states with large numbers of very small schools.

Summary and Conclusions

There are several key points to consider in the design of an accountability system for NCLB.

1. Make sure the overall design of the system is valid, and that it is implemented in a way that leads to valid results. Consider all the threats to validity outlined by Marion, White, et al. (2002) and eliminate those as a first step.
2. Use valid assessments, even if higher assessment validity comes at the cost of lower student-level reliability. Lower student-level reliability will have only a minor impact on the reliability of school scores.
3. Create a system that leads to reliable judgments. That requires some flexibility in the interpretation of NCLB language, but it is possible to make reliable judgments about both the status and improvement of most schools if results are examined over a period of years.
4. Do not set a fixed N for subgroups. There is no choice of fixed N that will provide for both valid and reliable results, and most choices provide for neither.
5. Do not attempt to measure school improvement over one year. The uncertainty is too great relative to the amount of improvement expected to make reliable judgments for most schools.
6. Carefully consider the real costs of Type I and Type II errors and attempt to reach a reasonable balance between the two. Do not assume that the costs associated with either type of error is insignificant.
7. Recent decisions from the U.S. Department of Education, particularly the approval of the initial five states in January, 2003, indicate a willingness on their part to interpret NCLB flexibly. States should carefully consider their choices in the creation of a design, select one that is as valid and reliable as possible for them, and then make a case for its acceptability based on that information.

References

Hill, R. K. (2002). *Examining the reliability of accountability systems*. Paper presented at 2002 Annual Conference of the American Educational Research Association. Portsmouth, NH: The National Center for the Improvement of Educational Assessment, Inc.

Hill, R. K., & DePascale, C. A. (2002). *Determining the reliability of school scores*. Portsmouth, NH: The National Center for the Improvement of Educational Assessment, Inc.

Kane, T.J., & Staiger, D.O. (2002). *Volatility in School Test Scores: Implications for Test-Based Accountability Systems*. Brookings Papers on Education Policy, 2002. Washington, DC: Brookings Institute.

Linn, R.L., Baker, E. L. & Herman, J. L. (2002, Fall). Minimum group size for measuring adequate yearly progress. *The CRESST Line*, 1, 4-5. (Newsletter of the National Center for Research on Evaluation, Standards, and Student Testing [CRESST], University of California, Los Angeles)

Linn, R.L., Baker, E. L. & Betebenner, D. W., (2002). Accountability Systems: Implications of Requirements of the No Child Left Behind Act of 2001. *Educational Researcher, 31*(6), 3-16.

Marion, S., White, C., Carlson, D., Erpenbach, W., Hill, R., Rabinowitz, S. & Sheinker, J. (2002). *Making valid and reliable decisions in the determination of Adequate Yearly Progress.* Washington, DC: Council of Chief State School Officers.

Markowitz, J. (2002, December). Subgroup size—Confidentiality and Statistical Reliability. *Project Forum: Quick Turnaround.* Alexandria, VA: The National Association of State Directors of Special Education.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425 (2002).