

Adequate Yearly Progress Under NCLB: Reliability Considerations

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

Paper presented at the 2003 Annual Meeting of the
National Council on Measurement in Education

Background

Each state must create an accountability system in response to the requirements of the No Child Left Behind Act (NCLB). Among the requirements is the determination of whether a school and each identified subgroup within the school either has achieved a particular percentage of students at the proficient level or higher (met the “status” requirement) or has improved its percentage of students achieving at the proficient level or higher over the prior year’s level (met the “improvement” requirement, often referred to as “safe harbor”). If the school or any subgroup within the school fails both those tests, the school fails to make Adequate Yearly Progress (AYP) and certain actions are taken against it. Results for subgroups are not required to be evaluated “in a case in which the number of students in a category is insufficient to yield statistically reliable information.” Each state is left to determine what “statistically reliable” means.

States have a wide range of choices to create their NCLB accountability design. One of the considerations of the design should be its reliability. The purpose of this paper is to outline some of those choices and discuss the impact that the choices will have on reliability.

Defining “Reliability”

While the term “reliability” is used in NCLB, the statistic of interest is not a reliability coefficient, but the probability of misclassifying a school. Reliability is dependent on the variance of the statistic of interest and basically tells us whether schools are ranked in the same order over two occasions. That is not what we are concerned about here; in this case, we want to know if a school is correctly or consistently classified, not if its position relative to other schools is unchanged. If schools are highly variable, many incorrect decisions might be made about schools even though reliability is high; and conversely, a system might make mostly correct decisions about schools despite low reliability if the variance is small. Therefore, throughout this paper, references to “reliability” will mean “probability of a correct decision.”

Variation in School Scores

The first point to note is that 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. Said another way, even if the “true score” for a school remains constant over time, its “observed score” 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 error¹. 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 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 years².
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 fixed³?

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.

Factors that Have a Strong Effect on Reliability

Some factors strongly influence the probability that a school will be correctly classified, while others affect it hardly at all. The major influences include the number of students tested, the number of decisions to be made, the number of grades being tested and the number of years of data used before making a decision. The minor influences include the reporting statistic and the student-level reliability of the test being used. We will discuss that first set of factors in this section and then discuss the remaining two in the next.

Size Matters

Of all the factors influencing the probability of correct decisions, size is by far the most important—and in each of these various meanings:

1. The number of students in each group
2. The number of independent judgments to be made for each school
3. The distance the true score for the subgroup is from the target

We will discuss each of these “size” variables in turn.

The number of students. The number of students largely determines the standard error of the percentage proficient. The error due to sampling students, which is by far the largest factor in the standard error, is a function of the square root of the number of students in the group. Every time the number of students is increased by a factor of 4, for example, the standard error is cut in half.

NCLB says that subgroup data “shall not be required in a case in which the number of students in a category is insufficient to yield statistically reliable information.” Some states have interpreted this to mean they should select a minimum number of students that should be in each subgroup before looking at the subgroup; other states have interpreted it to mean that the results for a subgroup should be statistically significant before making a decision.

The number of independent judgments. For most states, there are eight potential subgroups within each school. Each subgroup, and the school as a whole, must meet specified criterion for either status or improvement for both reading and mathematics. This means that a school must successfully pass 18 tests—each test having some degree of error associated with it—to not fail making AYP. In addition, the school also must pass tests—each of them also containing some degree of error—in participation rates and an “additional indicator.”

To get a sense of how bad things can be, let's take a worse-case scenario: a school that has students in all subgroups, with membership in every subgroup independent of every other subgroup, with every subgroup falling below the status bar but improving exactly the 10 percent required of it by NCLB. In this case, every decision about the school would have an error rate of 50 percent (if the true improvement was exactly equal the required improvement, half the time the observed improvement would be more than the required amount and half the time it would be less). Looking only at the 18 tests for reading and mathematics, the probability that this school would incorrectly be identified is $1 - .5^{18}$, or .9999962. The good news is that this is an extreme case, and the likelihood that any given school would have such a high probability of error is small; the bad news is that for many schools, the error rate, while not this extreme, is still fairly high, particularly if it has been improving.

One key reason why the error rate usually is not that high is that there is great dependence among the subgroups. Often, for example, minority students comprise a substantial portion of the "economically disadvantaged" students. As a result, the errors are correlated, thereby reducing the error rate of multiple tests. As an extreme example, suppose all the Hispanics students in a school were the only economically disadvantaged students. In that case, the error rate for the two subgroups would be the same as the error rate for either group—the groups are the same. When that happens, the number of effective tests is considerably smaller than 18.

Even when there is not overlap among the subgroups, reading and math scores are highly correlated. Again, the consequence of this is to reduce the effective number of tests. The "effective" number of tests is dependent upon the number of subgroups within schools and the intercorrelation among them. In order to get a feel for what the effective number might be, we did the following data analysis.

We took the student-level data file for a state and drew a sample with replacement. Then, we drew another sample along with a randomly drawn value from a uniform distribution that ranged from 0 to 1. If the student drawn in the second sample was a failing student, we changed that failing score to passing if the value of the random uniform number was less than .10. By this process, we created a second sample whose true score had 10 percent fewer failing students than the original sample. Said another way, we drew a sample from a population that met the requirement for improvement set by NCLB—that the percentage of students failing was reduced by 10 percent from one year to the next. To be correctly classified, these schools should never be labeled as failing AYP—every subpopulation in them met the necessary standard for improvement.

We then computed the standard error of the difference as follows:

Let P = the proportion of students in the subgroup passing in Year 1,
 $Q = 1 - P$,
 $R = P + 0.1 * (1-P)$
 $S = 1 - R$

Thus, P is the observed result for Year 1 and R is the target result for Year 2. Then,

$$s_{x_2-x_1} = \sqrt{\frac{PQ}{N_1} + \frac{RS}{N_2}}$$

We then computed the difference between the observed change and R-P, and then divided that difference by the standard error. If the distribution were normal, we would expect that 5 percent of the time, the value of the ratio would be less than -1.645. Given the small numbers of students in many of the subgroups in these schools, that was not the case. We found that using a *z*-score of -1.23 gave us about a 5 percent error rate for larger units, and something less than that for smaller units.

Thus, the error rate for testing each subgroup in a school was .05 or less. The results are shown in Table 1. In 19 percent of the schools, at least one subgroup failed reading; similarly, in 19 percent of the schools, at least one subgroup failed mathematics. In almost 32 percent of the schools, at least one subgroup failed the improvement test in at least reading or mathematics. Thus, while the error rate for any individual subgroup was set to be fairly low (.05 or less), almost one-third of the schools failed AYP. Remember, this is true despite the fact that the data were established so that *no* school should fail AYP—every subgroup in every school improved (in its population) by the amount required by NCLB. Indeed, when we ran the tests of statistical significance on the population (comparing Year 1 to Year 2 without drawing a new sample), every school made AYP.

We repeated the analysis, but this time used an *alpha* level of .01. In this case, 4.5 percent of the schools had at least one failing subgroup in reading. Eight percent had at least one failing subgroup in either reading or mathematics.

Table 1
Probability of a School Not Making AYP Even if Every Subgroup Has Improved the Required Amount

Statistic	Nominal alpha rate = .05	Nominal alpha rate = .01
Actual alpha rate (average across all subgroups)	.033	.0067
Within one content area, some subgroup fails	.19	.05
Across both content areas, some subgroup fails	.32	.08

Thus, despite the fact that these judgments were not independent, it turned out that the percentage of schools incorrectly classified as failing even though they had truly improved was considerably higher than the error rate for the judgments on each individual subgroup within the school. Using an error rate of .01 on each decision made within a school leads to an overall (“school-wise”) error rate of approximately 8 percent. Using an error rate of .05 on individual groups leads to a school-wise error rate of over 30 percent.

Distance from the target. Big differences are easy to detect, even with modest numbers of students, while small differences may be undetectable even with very large numbers. If a state’s status requirement is 50 percent proficient, for example, a subgroup with no students passing out of 10 is highly likely not to have a true score at or above the required amount. On the other hand, a subgroup with 499 students passing out of 1,000 may very well have a true score at that level. Similarly, it is easier to detect whether a subgroup has made its required gain if the amount is 15 percent instead of 5 percent.

Given that the performance of minority subgroups often is well below the status bar established by NCLB, one can detect with a fair degree of accuracy whether a subgroup has failed to meet the status bar for a large number of subgroups. But since the amount of improvement called for by NCLB is often a fairly small amount, detecting that change accurately can be problematic.

Status vs. Improvement

To fail AYP, a subgroup must fail two tests: status and improvement. It fails the status test if its percentage of proficient students is below the state's requirement amount. It fails the improvement test if the percentage of non-proficient students is not reduced by at least 10 percent from the previous year.

Status tests, in general, are far more reliable than improvement tests, for two primary reasons: (1) the distance from the target is generally much larger for status than it is for improvement, and (2) status tests involve just one sample (and one set of sampling error) while improvement tests involve two. In many states, the status requirement is anywhere from 30 percent passing on up. Given that many subgroups in many schools have small percentages of proficient students, it is not unusual to find subgroups that are 10, 20 or 30 percentage points away from the required level. In contrast, the largest improvement requirement possible is 10 percent (if no students currently are passing, the subgroup must improve to 10 percent). Combined with the fact that the standard error of the difference for two independent samples is the square root of 2 times the standard error of one sample, improvement is measured far less accurately in most cases than status.

Number of Grades and Independence across Years

Many states currently are testing only in selected grades. By the 2005-06 school year, they must have testing in reading and math in all grades 3-8 and at least one grade in 10-12. At that point, many of the design parameters will change.

For example, with testing at every grade level, the standard error of difference scores across years will shrink. First of all, the numbers of tested students will increase substantially. At the same time, many of the students tested one year will be retested the next, thereby reducing the sampling error involved in measuring gain. Suppose a K-5 school has 50 students per grade with 50 percent of the students passing. If just one grade is tested, the standard error of the difference across two years of testing is $\sqrt{2 * 50 * 50 / 50}$, or 10.0. If three grades are tested, 50 percent of the students return from Year 1 to be tested in Year 2, and the correlation of student scores across years is .7, then the standard error of the difference is $\sqrt{(2 * 50 * 50 - 2 * .7 * .5 * 50 * 50) / 150}$, or 4.67. That is a very significant reduction from what it was with just one grade tested—the equivalent of *quadrupling* the number of students tested in each subgroup.

Looking at Results across Years

One option to increase the reliability of accountability systems is to combine data across years. This has the effect of increasing the number of students in a subgroup, thereby reducing standard errors.

Note also, however, that one could also use the idea of increased years to require increased improvement. If a subgroup is expected to reduce its percentage of non-proficient students by 10 percent each year, it should reduce that percentage by 19 percent over two years and 27.1 percent over three. As noted above, larger targets lead to fewer classification errors, so a system that expects 27.1 percent improvement over three years will have a far lower misclassification rate than one that expects a 10 percent improvement in one.

Perhaps more importantly, however, schools start accruing severe sanctions only after they have been identified as failing AYP two consecutive years. This appears to be a significant element in accountability design that has been largely ignored. Perhaps states have decided that mislabeling a school even for one year is unacceptable, and therefore designed systems to label schools as accurately as possible each year. However, if one takes the position an error in any one year is moderately acceptable, so long as few errors are made two consecutive years, one can approach the problem of accountability design very differently.

Factors that Have a Small Effect on Reliability

There are at least two factors that one would think would have a strong effect on reliability, but turn out not to: the reporting statistic and the reliability of the tests used.

No Child Left Behind calls for pass/fail judgments to be used. Mean scaled scores and indices (giving students more points for scoring at higher performance levels) *are* more reliable reporting statistics, but given the variety of rules applied in NCLB, it turns out that decision classification is only marginally improved if the same rules (or parallel rules) are employed in an NCLB design using these more reliable reporting statistics.

The reliability of student-level scores has an even smaller impact on the reliability of school-level decisions. In several of the papers we have produced on school-level reliability, we have shown that reliability of student-level data has a very minor impact on the reliability of school-level data.

NCLB's Strengths and Weaknesses (from a Reliability Point of View)

In some ways, the provisions of NCLB mandate an accountability design that cannot be highly reliable. The number of decisions that need to be made about each school is a primary cause for this. Dividing schools into subgroups (which, by definition, have a smaller number of students than the school as a whole) and then requiring that each of the subgroups passes tests in both reading and mathematics makes it highly likely (as shown above) that an incorrect decision will be made about a school. If only status decisions need to be made, the probability that a subgroup will be incorrectly classified is low, so the school-wise error rate, made jointly across all the decisions for a school, will be low. However, the school-wise error rate for improvement decisions is likely to be quite high for schools that have truly improved.

On the other hand, there is an element of NCLB that allows for much higher correct classification rates—the ability to require that improvements be examined over two years rather than one. The consequences for a school that fails to make AYP two years in a row are considerably higher than those for failing to make AYP in any given year. If NCLB designs took advantage of this, they could be made considerably more reliable.

In the lingo of NCLB, a school that fails one year “fails to make AYP.” A school that fails two consecutive years is “INOI (In Need of Improvement).” There are several possible decision rules for determining whether a school should be classified as INOI:

1. The school or any subgroup within the school fails either reading or math one year, and then fails either reading or math the second year (“Any Subgroup, Either Test”).
2. The school or any subgroup within the school fails reading two years in a row or fails math two years in a row (“Any Subgroup, Same Test”).
3. The same subgroup fails the same content area two years in a row (“Same Subgroup, Same Test”).

The Relative Error Rates of Alternative Decision Rules

To look at the relative error rates of these various possible decision rules, we extended the data analysis we described in the earlier section where we discussed the impact of the number of judgments on the error rate. We drew a third sample for which the probability that a previously failing student was reduced another 10 percent. Thus, we modeled a state for which no school should be identified either year—the percentage of students failing the second year was 10 percent less than the first, and the percent failing the third year was 10 percent less than the second. We then applied these decision rules to the observed results.

The results are shown in Table 2. As would be expected, the more constrained the decision rule, the smaller the proportion of misclassified schools.

Table 2

Probability of a School Being Identified as INOI Even if Every Subgroup Has Improved the Required Amount Two Consecutive Years

Decision Rule	Nominal alpha rate = .05	Nominal alpha rate = .01
Any Subgroup, Either Test	.12	.016
Any Subgroup, Same Test	.08	.012
Same Subgroup, Same Test	.04	.002

As noted, all these schools improved and therefore should not have been identified. To make the analysis more meaningful, we ran the same analyses, but on schools that made no improvement. In that second analysis, every school *should* be identified as not making AYP in one year and as being INOI after two years. The most reliable accountability system would identify a small percentage of improving schools and a large percentage of not improving schools. That is, any accountability system could be designed to correctly identify improving schools—all it would have to do is make the identification rules so stringent that no schools got identified, whether improving or not. Similarly, any accountability system could be designed to correctly identify non-improving schools—it simply would have to identify all schools. What is most important is not the percentage of schools identified or not identified, but the ability of the system to accurately distinguish between those that have improved and those that have not.

Table 3

**Probability of a School Not Making AYP or Being Identified as INOI,
Reported by Whether the School Has Improved**

Statistic	Nominal alpha rate = .05		Nominal alpha rate = .01	
	Improving	Not Improving	Improving	Not Improving
One Year—Within one content area, some subgroup fails	.19	.40	.05	.17
One Year—Across both content areas, some subgroup fails	.32	.52	.08	.26
Two Years—Any Subgroup, Either Test	.12	.38	.016	.13
Two Years—Any Subgroup, Same Test	.08	.34	.012	.11
Two Years—Same Subgroup, Same Test	.04	.26	.002	.08

With just one year of data, it is difficult to discriminate between the two groups. To illustrate, suppose our state had 100 schools, and 50 had improved while 50 had not. If we used a .05 alpha rate in our tests of statistical significance and identified as not making AYP any school having a failing subgroup in either content area, we would identify 16 of the 50 improving schools (32 percent) and 26 of the not-improving schools, or a total of 42 schools. Of those 42, 16 of them (the 16 that actually improved), or 38 percent, would be incorrectly identified. We would not identify 58 schools. Of those, 24 *should* have been identified—the error rate is 41 percent. So the total error rate for this example is around 40 percent.

If we extend our example across two years and apply the “Any Subgroup, Either Test” rule at the .05 level, we would identify 6 improving schools and 19 not-improving schools. The error rate for identification would be 24 percent—a substantial improvement over the one-year error rate. The error rate for non-identification would be 31/75, or 41 percent—about the error rate of the one-year rule. Thus, the two-year decision rule allows us to identify improvers more accurately, but not non-improvers.

Now, suppose we apply the second two-year rule—Any Subgroup, Same Test. The error rate for identification is 4/21, or 19 percent. The error rate for non-identification is 42 percent. Fewer schools are identified, and we more accurately judge those who have improved as having done so.

Finally, suppose we apply the “Same Subgroup, Same Test” rule. We identify fewer schools. Again, the error rate for identifying improved schools drops—this time 13 percent of the identified schools would be those that truly improved and therefore are incorrectly chosen. The error rate for non-identified schools would be 44 percent—a small increase over the other decision rules, but arguably the improvement in the error rate for truly improving schools sufficiently compensates for that.

The above discussion assumes that half the schools have improved (and improved within every subgroup) by the required amount and half have not. If a larger percentage of schools have truly improved, the error rates will be lower the more restrictive the decision rule chosen (i.e., the statistics favoring “Same Subgroup, Both Years” will increase faster than the rates for the other decision rules). If a smaller percentage of schools have truly improved, a decision rule that identifies more schools will yield a lower misclassification rate.

Ideally, a state would wait to see statewide results and look at the amount of improvement being made before chosen a decision rule. The more statewide improvement shown, the more likely it is that more schools have made improvement, and therefore the state should use a more restrictive rule that identifies fewer schools. Given that decisions must be made in advance about what decision rules will be applied two years from now, states must make a decision about whether their schools will be improving or not. The more optimistic the state, the more appropriate it is they choose a “Two Years, Same Subgroup” rule.