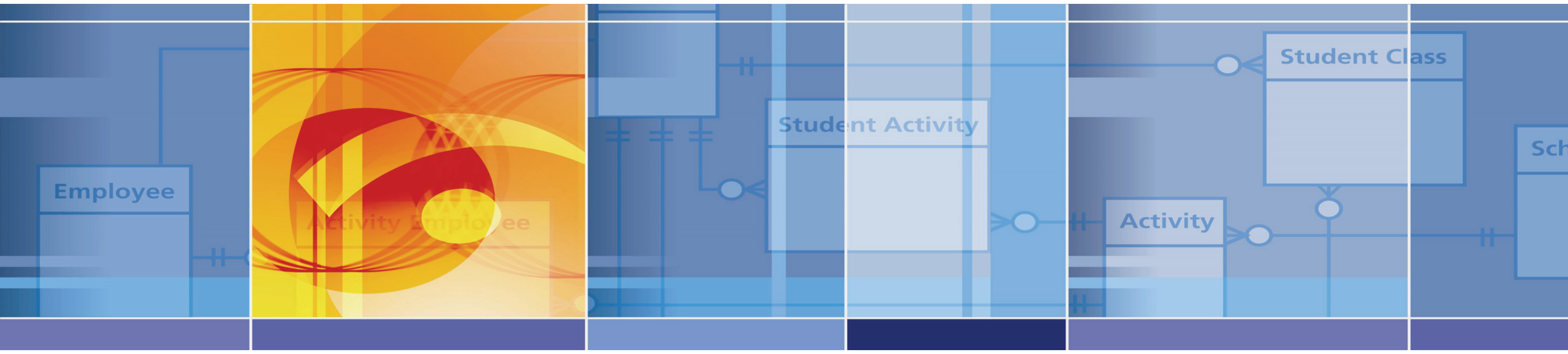


The Optimal Reference Guide:

Confidentiality and Reliability Rules for Reporting Education Data

A guide for establishing decision rules for disaggregating & reporting assessment results, and other indicators

Extraordinary insight™ into today's education information topics

By Glynn D. Ligon, Ph.D. and Barbara S. Clements, Ph.D.

ESP Solutions Group

Table of Contents

Foreword

By Glynn D. Ligon

Districts and states are responsible for protecting the confidentiality of personally reliable information about individuals whenever data are reported publicly. They are also charged both professionally and legally with determining the reliability of the data published. This is not new; FERPA has been around since 1974. Hays published the second edition of his statistics textbook the year before. The No Child Left Behind Act dusted them off and moved them to the top of everyone's "must read" list.

This resource guide talks a lot about the No Child Left Behind Act and adequate yearly progress, but every time a district or state reports data, these same issues apply. So, please do not think this document is only for AYP reporting; however, the time is here to revisit AYP decisions made related to confidentiality and reliability. Real data are in hand now to evaluate decision rules.

We are struggling with the tension between masking data that reveal personally identifiable information and preserving the integrity of our accountability systems by including all students. There is also the tension to preserve the integrity of our accountability systems by reporting only statistically reliable data. Now enter the statisticians with textbooks in hand and arcane ideas of how to apply statistics to today's accountability reports. What we need is thoughtful politimetrics to replace traditional psychometrics and statistics. The new politimetrics will help us implement accountability systems that work for today's schools and students. Politimetrics will merge the political mandates and realities with the appropriate statistical methodologies.

ESP Insight
The new politimetrics will help us implement accountability systems that work for today's schools and students.

Using the context of today's schools rather than a research university, this resource guide pushes back on traditional sampling theory as the best way to determine reliability. This guide also proposes alternative reporting methods for protecting the confidentiality of individuals in small groups without losing all the information we have about those groups.

I began studying confidentiality and reliability issues without a full appreciation for their complexities. I thought education agencies would be able to select two reasonable numbers, say 5 for confidentiality and 30 for reliability, and move on to other priorities. Now I know that 5 as a minimum for confidentiality may work well, but any single number for statistical reliability has problems. Some proponents of sampling-based methods think my recommendation to use standard error of measurement (SEM) in a significance test for reliability is off-base. I think the case for SEM with the No Child Left Behind Act is compelling. This publication should help you form or reinforce your own conclusion.

These are significant issues for the success of an accountability plan. Please feel welcome to contact us for additional help and advice related to confidentiality, reliability, or other related issues.

The following persons also contributed to the contents of this document:
Vince Paredes, Ph.D., Judy Jennings, Ph.D., and Evangelina Mangino, Ph.D.

The concepts discussed in this paper are as important now as they were in 2005 when the paper was published. Reporting student data requires careful consideration of groupings of students such that individual students cannot be identified, as required by the Family Educational Rights and Privacy Act (FERPA). The U.S. Department of Education (USED) published proposed revisions for implementation of FERPA in March 2008. These proposed revisions reinforced the importance of ensuring that policies and procedures are in place to prevent individually identifiable information about students from being disclosed to inappropriate people. The proposed revisions discuss de-identification of student records and other methods used to ensure non-disclosure. This paper provides specific guidelines for meeting the requirements of FERPA.

Introduction

Each state set a minimum number of students before disaggregating subgroups in response to accountability and reporting requirements of No Child Left Behind. School districts have also established local rules before publishing reports about student performance. These minimums ensure that no individual student's information will be revealed publicly and that no disaggregated subgroup will be too small for their results to be statistically reliable.

ESP Insight
The confidentiality conundrum: If two's company, and three's a crowd, can we pick a student out of a crowd?

> ***n*** **= number: *In this publication, n is used to designate a number selected as the minimum for confidentiality or statistical reliability.***

No Child Left Behind does not require a traditional accountability system. Quite to the contrary. The system detailed by No Child Left Behind was not already implemented by any state (including Texas). So should statisticians be applying traditional methods to No Child Left Behind issues? No.

- No Child Left Behind does not allow schools to meet adequate yearly progress (AYP) objectives by averaging student scores in reading, mathematics, and science. A great score by one student cannot average out a poor score by another.
- No Child Left Behind looks at every single student. The requirement is that every single student reaches proficiency. Improving America's Schools Act, the predecessor to No Child Left Behind, introduced this perspective.
- Even the uniform averaging procedures of combining across grade levels or across school years is mathematically equivalent to combining students into a common pool to be counted together.
- However, this requirement to count individual students rather than to average scores across students contrasted with many currently implemented state accountability systems that did use averages.

ESP Insight
The statistical reliability conundrum: How many students must be tested for us to know anything significant about them?

No Child Left Behind created a new accountability system for all of the nation's schools. The *Improving America's Schools Act* provided a system of accountability only for Title I schools. No Child Left Behind expands accountability to be a unified system for all schools. Each state's rules were revisited, and the interpretations of them updated. Each state determined how to achieve continuity and transition from their old to their new system. The central concept of AYP is precisely described and formulated in the law.

Simple answers based upon adopted minimum numbers of students appear to be described by No Child Left Behind. "The 95% requirement… disaggregation… inclusion in an annual report card… shall not be required in a case in which the *number* of students in a category is insufficient to yield statistically reliable information or the results would reveal personally identifiable information about an individual student."

Statisticians and policy decision makers must explore all the ramifications, assumptions, and implications beyond simple numbers for every alternative. The implementation of No Child Left Behind required the psychometricians and statisticians to work with the politicians and policy makers to devise and adopt methods that are both theoretically sound and politically practical. In discussing these issues, we "invented" the word politimetrics, only to find that the concept as well as the actual term has been in use since at least the 1970's. Politimetrics (Gurr, 1972; Alker, 1975; Hilton, 1976), the accommodation of psychometrics to the political world, is required by No Child Left Behind. Politimetrics describes the dynamics when politicians must rely upon statisticians to devise defensible methodologies. The statisticians are then challenged to present and explain their solutions in a political context. What works in educational research for a professional journal may make little sense in the real world of schools, state government, and federal mandates. The classic example is the impossibility of randomly assigning students to schools for a research study of school effectiveness. This Byzantine example of random assignment is similar to what some statistical techniques assume happens in the real world.

ESP Insight

Politimetrics, the accommodation of psychometrics to the political world, is required by No Child Left Behind.

No Child Left Behind and the realities of its 12-year accountability cycle depart from traditional psychometric and statistical techniques. Principals think of their individual schools as unique. With about 90,000 schools, potentially over 7 million subgroups being measured, and all this over a 12-year cycle (that is 84 million subgroups in all), just about every conceivable combination of statistics and unique distributions of assessment scores may be observed. That is why states must test theories both in simulations and with their actual data from the first years of AYP. States must understand the implications for alternative decisions on a wide range of score distributions. States must be cautious in accepting methods that neatly accommodate 99% of the schools and subgroups. After all, that other 1% would be 900 schools and 840,000 subgroups as exceptions.

AYP as required by No Child Left Behind is **NOT**:

- A value-added model that measures how much a school has accomplished based upon a student's starting point, demographics, or resources assigned.
- The measurement of AYP is not a regression formula that accounts for these variables to predict an expected performance level.
- A comparable-schools model that identifies schools of similar characteristics and compares relative performance.

- A best-practice model that finds high-performing schools and uses them as benchmarks for measuring others.
- A normative model that ranks schools and divides them into quartiles based upon performance.
- A gains model that measures a school's improvement in performance across years.
- A gains model that measures an individual student's improvement.
- A gains model that compares one year's performance in a grade to the next year's performance in the next higher grade level.
- A level-playing field model that adjusts each school's objectives according to the students or resources available.

In short, No Child Left Behind's AYP model is not any of these. Simply put, AYP requires every subgroup in a school to meet an annual objective set for each of multiple indicators. The objective is the same for all schools in a state. Specifically, No Child Left Behind requires that the objective be the same for Title I and all other schools within a state. This is a true standards-based model. Each SEA defines its standards; establishes performance benchmarks for them; measures current student performance; and sets incremental annual objectives that require each school, district, and state to "make progress" toward reaching 100% of the students in each subgroup performing at or above the performance benchmark (e.g., "proficient" on an assessment). "Make progress" is somewhat of a misnomer. A high performing school may already exceed the annual objectives set for the next five years.

No real progress would be required to meet the annual AYP standard—until that sixth year. The concept of progress is in the characteristic of the annual objectives, which continue to rise until they reach 100%. So a school that meets the annual objective right on the money each year would be making steady progress. However, the annual objective is stated in terms of an absolute performance in each year, not a gain in performance from the prior year.

Two issues must be resolved by each state in order to implement their accountability plan. These same two issues face districts whenever performance reports are published.

1. When are there too few students in a subgroup to allow disaggregation that will not reveal personally identifiable information for individual students?
2. When are there enough students in a subgroup to yield a statistically reliable measure of the subgroup's performance toward meeting the annual objective established for adequate yearly progress?

Neither of these issues is simple. The suppressed disaggregated values for one subgroup might be derived from the values published for other subgroups. A larger subgroup's value may be more statistically reliable than a smaller subgroup's. Therefore, these complexities are weighed in the guidance provided in this publication.

ESP Insight

The suppressed disaggregated values for one subgroup might be derived from the values published for other subgroups. However, a larger subgroup's value may be more statistically reliable than a smaller subgroup's. Therefore, these complexities are weighed in the guidance provided in this publication.

The Law – No Child Left Behind

No Child Left Behind says:
(C) DEFINITION-Adequate Yearly Progress shall be defined by a State in a manner that—

> (i) applies the same high standards of academic achievement to all public elementary and secondary school students in the State;
> (ii) is statistically valid and reliable;
> (iii) results in continuous and substantial academic improvement for all students;
> (iv) measures the progress of public elementary schools, secondary schools and local education agencies and the State based primarily on the academic assessments described in paragraph (3);
> (v) includes separate measurable annual objectives for continuous and substantial improvement for each of the following:
>
> > (I) The achievement of all public elementary school and secondary school students.
> > (II) The achievement of—
> >
> > > (aa) economically disadvantaged students;
> > > (bb) students from major racial and ethnic groups;
> > > (cc) students with disabilities; and
> > > (dd) students with limited English proficiency; except that disaggregation of data under subclause
> > >
> > > > (II) shall not be required in a case in which the number of students in a category is insufficient to yield statistically reliable information or the results would reveal personally identifiable information about an individual student

States must establish rules for:

- The smallest number of students in a subgroup (e.g., category) that can be disaggregated *without revealing personally identifiable information about an individual student*
- The smallest number of students in a subgroup (e.g., category) that can be disaggregated *and yield statistically reliable information*

Definitions

- **Adequate Yearly Progress (AYP):** Meeting the annual objectives for each indicator, grade level, and subgroup in a year

- **Annual Objective (AO):** The percent of students that must meet the criterion for proficiency on an assessment (or additional indicator); increases annually from the starting point to 100% in 12 years for assessments

- **Confidentiality:** Inability to determine from the subgroup values reported how an individual student performed on an indicator

- **Cut-Point:** The score that divides proficiency levels, e.g., the lowest assessment score that classifies a student as proficient rather than basic

- **Error:** The amount that a score or a measure derived from a group of scores varies across either measurement times or samples (Statisticians use error to refer to the imprecision of their tests and statistics. Error in this context does not equate to mistakes)

- **Indicator:** An assessment, graduation rate, or other measure of academic achievement

- ***n*****:** The number of students in a subgroup

- **Null Hypothesis:** For a subgroup that does not meet the annual objective, there is no difference between the observed distribution of scores and a distribution that would meet the annual objective; therefore, the subgroup's results are unreliable

- **One-Tailed Test:** Directional hypothesis that tests that one value is larger than another, contrasted with a two-tailed test that tests whether the two values are different in either direction

- ***P*****-Value:** The probability that the hypothesis is true (e.g., p = .05 means that the probability that the null hypothesis is true is only 5%; or that the probability that a directional hypothesis is true is 95%)

- **Standard Error of Measurement (SEM):** Range in which a student's score might vary if tested multiple times; plus or minus one SEM represents the range within which a student's observed score would vary around the student's true score 68% of the time; 32% of the time, the student's observed score would be farther away from the true score
    - **Test - Retest:** The SEM is determined by testing the same individuals multiple times
    - **Internal Consistency:** The SEM is determined by correlating individual item values with other items and derived scores

- **Starting Point:** The percent of students meeting the criterion for proficiency in the first year of No Child Left Behind

- **Statistical Reliability:** The degree of confidence associated with the decision of whether or not enough students in a subgroup performed above the cut point for proficiency to meet the annual objective

- **Subgroup:** A category of students as defined by No Child Left Behind for AYP (e.g., each race/ethnic group, economically disadvantaged, limited-English proficient, and children with disabilities with an IEP) or annual report cards (gender and migrant)

- **Type 1 Error:** Rejecting the hypothesis when it is really true (i.e., the subgroup is considered to have not met the annual objective when it actually has)

- **Type 2 Error:** Accepting the hypothesis when it is really false (i.e., the subgroup is considered to be statistically unreliable when it actually did not meet the annual objective)

Dynamics

Several AYP dynamics are evident among the factors related to decision rules for confidentiality and statistical reliability.

- The more subgroups that are disaggregated, the more subgroups that fail to meet an annual objective, and the more schools that fail to make adequate yearly progress (AYP) and are classified as in need of improvement (INOI).
- The higher the minimum *n* for confidentiality is set, the fewer subgroups that are disaggregated.
- The higher the minimum *n* for statistical reliability is set, the fewer subgroups that are disaggregated.
- If a statistical test is used for groups above the minimum n, fewer subgroups will be disaggregated because more will be classified as statistically unreliable.
- A one-tailed statistical test will identify more subgroups as statistically reliable than a two-tailed test with the same *P*-value.
- The smaller the p-value required for statistical significance, the fewer subgroups will be identified as statistically reliable.

ESP Insight
Without the same decision rules being used, comparability across states will be elusive.

Another way to think about these dynamics is to consider the decisions that would minimize the number of subgroups that are disaggregated—resulting in fewer schools being identified as in INOI.

Fewer schools are identified as INOI when:

- A larger *n* is adopted for confidentiality.
- A larger *n* is adopted for statistical reliability.
- A statistical significance test is used to determine the probability that a subgroup met an annual objective.
- The statistical test uses a two-tailed (non-directional) hypothesis.
- A lower *p*-value (e.g., .01 rather than .1) is used.

Two contrasting sets of rules illustrate these dynamics.

- **The maximum number of schools is identified as INOI when these are adapted:**
  a. Small minimum *n* for confidentiality
  b. Small minimum *n* for statistical reliability
  c. No statistical significance test

- **The minimal number of schools is identified as INOI when these are adapted:**
  a. Large minimum *n* for confidentiality
  b. Large minimum *n* for statistical reliability
  c. Statistical significance test for subgroups larger than the minimum *n*
    - Two-tailed test
    - *p* < .01

Hypotheses and Tails

The dynamics described above use the following hypotheses and significance tests.

> **Annual Objective:** A percent of students that must perform at or above the proficient level for the subgroup to meet AYP.
>
> ***p*-Value:** The probability that the subgroup’s percent proficient and advanced and the annual objective are the same (null hypothesis) or that the annual objective is greater than the subgroup’s percent proficient and advanced (directional hypothesis). If p = .01, then the probability that the hypothesis is true is 1%.

Statistical Significance

The table that follows clarifies how the type of hypothesis and one- or two-tailed test aligns with the wording of the question being answered. The decision to accept or reject the hypothesis is matched with the conditions for acceptance or rejection and the meaning of that decision.

Table 1: Hypothesis Wording Alignment

Type	Wording	Decision	Conditions	Meaning
Null Hypothesis Two-Tailed Test	For Subgroup Status: Met or Not Met	**Accept**	The calculated *p*-value is greater than the criterion *p*-value. (Example: *p* = .45; criterion = .05)	The subgroup's percent proficient and advanced is probably the same as the annual objective. The subgroup's performance is statistically unreliable.
	The subgroup's percent proficient and advanced is the same as the annual objective (i.e., equal to or higher).	**Reject**	The calculated *p*-value is less than the criterion *p*-value. (Example: *p* = .04; criterion = .05)	The subgroup's percent proficient and advanced is probably different from the annual objective. The subgroup's performance is statistically reliable.
Directional Hypothesis One-Tailed Test	For Subgroup Status: Met	**Accept**	The calculated *p*-value is less than the criterion *p*-value. (Example: *p* = .04; criterion = .05)	The subgroup's percent proficient and advanced is probably higher than the annual objective. The subgroup's performance is met, statistically reliable.
	The subgroup's percent proficient and advanced is equal to or greater than the annual objective.	**Reject**	The calculated *p*-value is greater than the criterion *p*-value. (Example: *p* = .45; criterion = .05)	The subgroup's percent proficient and advanced is probably not higher than the annual objective. The subgroup's performance is statistically unreliable.
	For Subgroup Status: Not Met	**Accept**	The calculated *p*-value is greater than the criterion *p*-value. (Example: *p* = .45; criterion = .05)	The subgroup's percent proficient and advanced is probably not lower than the annual objective. The subgroup's performance is statistically unreliable.
	The subgroup's percent proficient and advanced is equal to or greater than the annual objective.	**Reject**	The calculated *p*-value is less than the criterion *p*-value. (Example: *p* = .04; criterion = .05)	The subgroup's percent proficient and advanced is probably lower than the annual objective. The subgroup's performance is not met, statistically reliable.

The *p*-value represents the level of confidence the state requires for its determination of met or not met related to a subgroup's performance on an annual objective. Where the *p*-value is set makes a difference in the risk of making type 1 or type 2 errors. In other words, an unacceptable number of failing subgroups can be unreported for being statistically unreliable if the probability (*p*-value) required to reject the null hypothesis or accept a directional hypothesis is very low. Researchers often use .01, .05, or .1. For a state, the decision related to poorly performing subgroups is whether to:

- Select a low p-value such as .01 and risk excluding from AYP too many subgroups because their percent below proficient is determined to be statistically unreliable (type 1 error),

  OR

- Select a high p-value such as .1 and risk including in AYP too many subgroups because their percent below proficient is determined to be statistically reliable (type 2 error).

Generally the discussion among states has been that the second risk is the one to avoid. States would prefer to identify fewer low-performing schools than identify some that may not be low-performing. Therefore, selecting a lower p-value would be desirable. States must find a balance between protecting the schools from unfair identification and protecting the students in schools in need of improvement.

ESP Insight
States would prefer to identify fewer low-performing schools than identify some that may not be low-performing.

This is an important decision. The nature of statistical significance tests is such that the selection of the p-value could impact the number of subgroups designated as statistically unreliable just as much or more than the setting of a minimum n for reporting or setting of the *n*SEM as described in the alternatives provided.

Which Decision Rule Should a State Adopt First?

Should a state determine one of these two decision rules first? There does not appear to be a necessary sequence. Even though one may override the other in practice, both the confidentiality and reliability decisions must be made.

At What Level Do We Count Students for Confidentiality?

The minimum subgroup size to protect confidentiality should be applied to the whole subgroup, not to the number of students performing at each proficiency level.

What happens when a subgroup has enough students to meet the state's criterion for confidentiality, but when the students' performance is reported, there is a level (e.g., basic, or proficient/advanced) that contains fewer students than the criterion for confidentiality?

For example, there are 10 economically disadvantaged students with third-grade math scores. The state's minimum n for reporting is five, so the subgroup gets reported—or does it? What if there are only two students at the basic level? Does this subgroup then get eliminated from the AYP calculations? No. The subgroup should be disaggregated for AYP, but the distribution of the scores by proficiency level would be masked in public reporting. Reporting them publicly is part of the annual report card requirements. The report card reporting requirements specify what has to be disaggregated publicly. So a group can be included in AYP calculations but reported in the annual report card using the method for reporting ranges rather than actual values as described in the section titled **"Confidentiality *n*."**

If this were not the case, then self-defeating conditions would apply. For example, whenever a subgroup has fewer than n students performing either at the basic level or proficient/advanced, the group would be eliminated from consideration for AYP. Thus, subgroups that approach 100% proficiency would not be included as fewer than *n* students are left at the basic level. Poorly performing schools could escape inclusion as long as fewer than n students reached proficiency.

Related Issues

A. **Higher Standards for Small Schools and Subgroups:** Reality is that a small subgroup must have 100% of its students at or above proficient in order to meet lower annual objectives. In other words, a group of five students must have 100% or all five students at or above proficiency to meet an annual objective of 81%. Table 2 shows the points at which small groups of various sizes reach the 100% level.

ESP Insight

Reality is that a small subgroup must have 100% of its students at or above proficient in order to meet lower annual objectives.

Table 2: Point at Which 100% Becomes the Criterion for Meeting Annual Objectives for Small Subgroups

Students in Subgroup	3	4	5	6	7	8	9	10	15	20	25-33	34-49	50-99	>=100
Annual Objective that Requires 100% Proficient or Advanced	67%	76%	81%	84%	86%	88%	89%	91%	94%	96%	97%	98%	99%	100%

B. **Impact of 95% Inclusion Rule on Small Schools and Subgroups:** AYP requires that 95% of the eligible students within a subgroup be included in an assessment. In support of 20 as a minimum for statistical reliability, a subgroup of fewer than 20 students must have 100% inclusion on an assessment in order to meet the 95% standard. This places schools with smaller subgroups at risk of failing to meet AYP by virtue of a less than perfect participation rate. Although not technically a reliability issue, if subgroups smaller than 20 students are held to the 95% standard for participation, a single student untested will cause the subgroup and the entire school to fail to meet AYP.

ESP Insight

A subgroup of fewer than 20 students must have 100% inclusion on an assessment in order to meet the 95% standard.

C. **Sequence for Applying 95% and *n* Rules:** The sequence for applying the various rules is crucial. (This is different from the sequence for adopting the rules.) Logically, they should be in this order.

1. Determine that the minimum n for confidentiality is met. If this criterion is not met, then the subgroup is excluded from any disaggregation.
2. Determine that the results for the subgroup are statistically reliable.
3. Determine that 95% of each subgroup was included in the assessment. If this criterion is not met, then the subgroup

fails the annual objective—regardless of the percent of students performing at or above proficiency.

Why does the 95% rule come last? Consider a subgroup of 20 students. If only 18 (90%) are assessed, then the subgroup would fail. However, if this subgroup with 18 scores is first compared to a minimum for statistical reliability of 20, then it would be categorized as unreliable—not as failing. In another example, if four out of five students are assessed (only 80%) and the subgroup is first compared to a minimum for confidentiality of 5, then the subgroup would be suppressed rather than failed for not including 95% of the students.

D. **Students Not Tested:** What is to be done with the less than 5% of the students who are not tested? What is the denominator for this calculation? Do these students get included in the determination of the starting point and subsequently in the calculation of AYP each year? Some states consider these students as not being proficient or advanced. In the short-term, counting these students as not having met proficiency has little impact other than on an individual school with a preponderance of students not tested. In the long-term, for a school to meet the 100% annual objective in the 12th year, not only would all students tested have to perform at the proficient level, but there could be no students untested.

E. **Uniform Averaging Procedure—Across Years:** No Child Left Behind allows for combining across the most recent two or three years to determine if the annual objectives are met. There is no specification whether this is to be a weighted average (adjusted for the number of students in the subgroup each year) or whether each year counts equally. The most logical approach may be to combine the counts of all students over the years into one group and calculate the percent proficient or advanced. This is not an average, but is equivalent to a weighted average. The uniform averaging procedure across years would increase the number of students in a subgroup and may allow for use of more subgroups in AYP. Could a state include in its plan the use of averaging across years as a method only for increasing the size of subgroups?

ESP Insight

As a general process for all schools, averaging across years does not make much sense.

As a general process for all schools, averaging across years does not make much sense. Assuming that schools are improving across time, this averaging would always have a depressing effect on a school's percent proficient and advanced. In fact, in order to meet the year 12 objective of 100%, a school would need to have already been at 100% for two or three years. Therefore this provision in effect shortens the number of years in which a school must reach 100%.

This provision makes sense as a safe harbor. As such, minor changes in a school's student population or other factors that might lower its performance slightly for a year might not cause the school to be in need of improvement. The inclusion of this option under the uniform averaging procedure section appears to indicate that this would have to be adopted by a state as the general process used for AYP—rather than using this option as another safe harbor provision and applying it only to schools already determined not to have met an annual objective. However, a state might propose this as a safe harbor provision and seek approval of the U.S. Department of Education for this interpretation.

If averaging across years is used as a safe harbor, then it would apply only to schools already meeting the minimums for confidentiality and statistical reliability, because those would be prerequisites to not meeting an annual objective. Thus, prior years' students would not be counted toward either minimum.

F. **Decrease in the Percent Performing at the Basic Level:** No Child Left Behind allows for calculating the decrease in the percent of students in a subgroup who perform at the basic level if the annual objective is not met in the most recent year. If this percent decreases by 10%, then AYP is met for this subgroup. This raises the issue of whether the prior year's percent must meet both the minimums for confidentiality and statistical reliability to be used. A state's plan should describe the process to be used.

G. **Report Cards versus AYP:** Might a state adopt different rules for confidentiality and statistical reliability for the annual report cards and for AYP? For example, might a state disaggregate on the report card a subgroup that is large enough to protect the confidentiality of the students but has been determined to be statistically unreliable for AYP? No Child Left Behind uses the same wording for AYP and the annual report cards. Even though statistically unreliable for AYP purposes, the value of a subgroup is actual and may be of interest to someone in the public. In reference to report cards and public reporting, No Child Left Behind says that "disaggregation shall not be required in a case in which the number of students in a category is insufficient to yield statistically reliable information or the results would reveal personally identifiable information about an individual student." Although not required, a state's plan apparently could provide for the reporting of subgroups in the annual report card that were considered statistically unreliable for AYP. Alternatively, there does not appear to be a requirement that the rules for confidentiality and statistical reliability be the same for AYP and the report cards.

ESP Insight
A state's plan could provide for the reporting of subgroups in the annual report card that were considered statistically unreliable for AYP.

ESP Insight
There does not appear to be a requirement that the rules for confidentiality and statistical reliability be the same for AYP and the report cards.

H. **Alternative Academic Indicators:** Do the rules for confidentiality and statistical reliability apply equally to the alternative indicators that are not assessments? These indicators may include, but not be limited to, graduation rate; decreases in grade-to-grade retention rates; attendance rates; and changes in the percentage of students completing gifted and talented, advanced placement, and college reparatory courses. No Child Left Behind does not address this issue directly, but only references the assessment indicators when discussing statistical reliability and the 95% inclusion rule. These are counts of real events without the same type of measurement error associated with assessments. Errors in counting or reporting these statistics are certainly not sampling errors or measurement errors with a known variance. They are mistakes of unknown frequency or probability. We do not offer a rationale for selecting a method of calculating statistical reliability for these counts other than making a professional judgment with defensible face validity. One such judgment would be to apply the same confidentiality minimum *n*, and to accept the logic that these are counts of clear events/students; therefore, no statistical reliability calculation is required. Guidance from the U.S. Department of Education has indicated agreement with this perspective.

I. **Rounding:** If the annual objective is 75%, has a school with 74.5% proficient or advanced met the annual objective? In science, a threshold is the point at which something occurs. Nothing occurs below that point, and something always occurs above that point. There is no rounding. The point here is that the threshold for meeting the annual objective must be clearly defined by the state—whether that point is 74.5% or 75%. Then there should be no additional rounding.

The same applies to the determination of a 10% reduction in the percent of students at the basic level when that safe harbor provision is used. The state must determine exactly whether the reduction must be 9.5% or 10.0%. Ten percent may not be a whole student. Out of 22 students, 10% is 2.2. Will the state accept a reduction of 2 students, which is 9.09%? Because the number of students may vary from one year to the next, the percent is the reasonable number to use rather than a count of students.

A second rounding issue is whether the annual objectives will be expressed as whole numbers or will be expressed with decimal places.

J. **Highly Qualified Teachers:** Highly qualified teachers are not part of AYP. However, the issue arises as to whether or not the confidentiality of teachers should be protected the same as that of students. This is most likely an issue that must be answered by each state based upon the applicable state laws. As public employees, certain aspects of teachers' qualifications and assignments may be public. Statistical reliability does not apply to the reporting of highly qualified teacher data.

ESP Insight
The most successful students are removed from the LEP and IEP subgroups each yuear and replaced with less successful students.

K. **LEP and IEP Catch 22:** As students develop their English skills, they leave LEP status. Students in some disability categories exit services as they improve. The impact is that the most successful students are removed from these subgroups each year and are replaced by others that are less successful. Therefore, the "Catch 22" becomes that schools that are successful with these subgroups are denied the inclusion of the scores for their most successful students. Could states consider including the scores of IEP and LEP students for at least one year after they exit from services?

L. **Students in Multiple Subgroups:** An artifact of AYP is that every student is in at least one subgroup and the total group, and some students are in many more. For example, an economically disadvantaged, Hispanic, LEP, special education student is in four subgroups plus the total group. This may not affect decisions about minimum subgroup sizes, but it does influence how states, districts, and schools think about individual students. If the student in this example performs at the basic level, then the status of meeting the annual objective is at risk for four subgroups and the total. If this student does not participate in the assessment, then all groups risk falling below the 95% participation criterion. These multiple subgroup students can be of considerable influence in a small school or in a school in which several subgroups are small.

Because AYP treats the subgroups separately—each receiving an independent determination of meeting the annual objective—the determinations of confidentiality and statistical reliability appear to be independent for the subgroups even though they share the same students.

M. **Attention to the Reliability of State Assessments:** The statistical reliability demands of AYP place considerable attention on the SEM/reliability of a state's assessments. A well-designed criterion-referenced assessment will have a preponderance of items with a 50/50 difficulty level (p = .5) for students performing near the criterion. This maximizes the precision of measurement at the critical cut point. Unfortunately, it also can lessen the precision of decisions at other cut points such as between proficient and advanced. The solution is to also have a large number of items around that cut point. The overall length of the assessment then becomes an issue. State assessments will be exposed to more scrutiny of these issues than may have been directed at them in the past.

The Future Gets Even More Intriguing.
We should begin to imagine what will happen as we approach year 12 in the long-term cycle of No Child Left Behind.

- Schools that meet the final annual objective of 100% of their students performing at or above the proficient level would not be permitted under FERPA's most restrictive interpretations to publish that success.
- Schools that approach 100% proficiency may find that statistical reliability will be very difficult to achieve.

In other words, a conundrum emerges. When schools meet their goals, we may not be able to credit them with that success. As statisticians, when the schools match their goal, we will have met our match as well. Recent FERPA guidance has softened on the issue of suppressing "good news." If 100% do well, that may be reported.

ESP Insight
If FERPA is applied too rigorously, then when schools meet their goals, we may not be able to credit them with that success.

Small *n* Decision Rubric

Setting the Minimum n for Confidentiality & Reliability

What criteria should a state use when selecting a minimum for confidentiality and statistical reliability?

The "Small *n* Decision Rubric" is a wizard-like flow chart designed to guide a state in establishing the decision rules for protecting confidentiality and establishing statistical reliability. The rubric should be used with the "Decision Template" that outlines the impact and considerations associated with n's of various sizes. The sections titled "Confidentiality *n*" and "Reliability *n*" provide details about these issues.

ESP Insight

These 13 steps will produce a defensible decision rule for confidentiality and reliability.

1. What is the minimum n that protects the confidentiality of individual students? *Three students have degrees of freedom of 2; five students protect against someone who knows up to three students' scores.*

2. What number represents the point above which there is little benefit in reliability from adding more students? In other words, at what point does the gain in reliability from having more students in a subgroup decelerate or start to level off? *Generally, the rule of thumb in statistics has been that the probability tables begin to flatten at about 30 subjects.*

3. What number is so high that an unacceptable number of subgroups would be excluded from AYP? *The higher the minimum n, the fewer subgroups will be disaggregated and reported. The fewer the subgroups, the fewer the schools that will be classified as in need of improvement (INOI). However, the validity of the accountability system is jeopardized if too many subgroups and too many schools are excluded because they are too small.*

4. What number is fair to small groups having to meet the 100% participation rate for assessments? *The 95% participation rate requirement becomes 100% for a subgroup smaller than 20 students.*

5. What number is fair to small groups when the annual objectives reach 80%, 90%, or higher? *An annual objective of 95% becomes 100% for a subgroup smaller than 20; 90% becomes 100% for a subgroup smaller than 10.*

6. At what point is a small subgroup unlikely to achieve statistical reliability regardless of its performance? *Other than subgroups with 100% performance at the same level, subgroups around five are unlikely to be statistically reliable.*

7. Is there a number below which a subgroup should not be judged even if all its students perform at the basic level? This is a judgment call based upon political or community consensus.

8. Is one of these issues so important that it should determine the final number?

9. What is the minimum number below which a subgroup should be excluded from AYP—based upon the considerations above?

10. Should subgroups larger than this minimum number be tested to ensure their results are statistically reliable? In other words, should a statistical test be run to establish the actual probability that a subgroup's performance is different from the annual objective, or should the results for all subgroups above the minimum number of students be accepted as statistically reliable?

11. Is there a size above which no test should be run? Is there a number of students that is sufficient to provide statistical reliability without a test being run?

12. If the answer to #10 is "Yes," what test of statistical reliability will be used? See "Reliability *n*" for details describing alternatives.

13. What level of confidence will be accepted for statistical reliability (e.g., p = .1)? Is this a directional hypothesis (for instance, yes, e.g., the observed value is greater than the annual objective = one-tailed test; no, e.g., the observed value and the annual objective are equal = two-tailed test)? See "Reliability *n*" for details describing alternatives.

The resulting decision rules can be summarized by filling in the following statements.

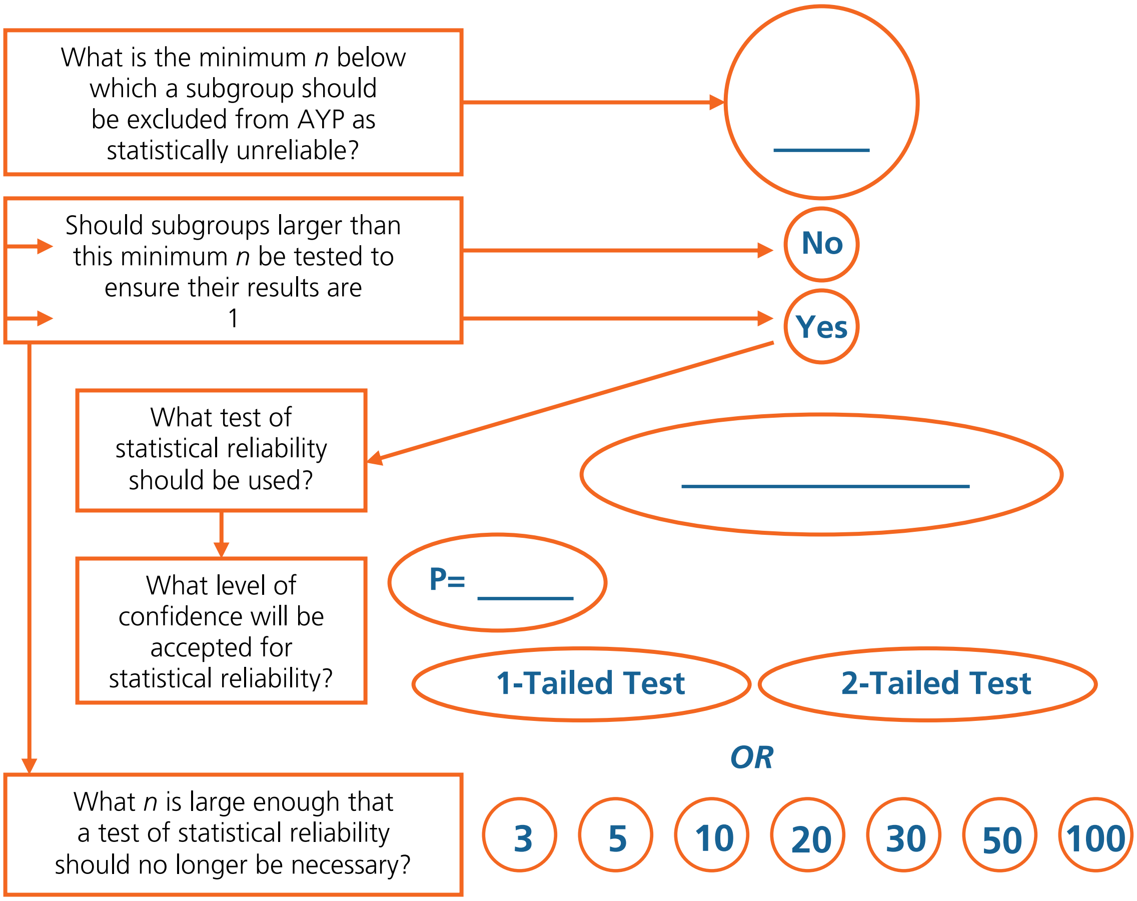
Decision Rules:

1. Confidentiality: Do not disaggregate subgroups smaller than ______.
2. Statistical Reliability:
3. Do not disaggregate subgroups smaller than ______.
    a. Use (name of statistical test), p = (level of probability),
    b. (one- or two-) tailed test to determine reliability of larger subgroups.
    c. Do not test for reliability of subgroups larger than ______.

Decision Rubric

Figures 1 & 2: Decision Rubric

1	What is the minimum *n* for protecting confidentiality?	3	5	10	20	30	50	100
2	What *n* is at the point where the gain in reliability from having more students levels off?	3	5	10	20	30	50	100
3	What *n* is so high that an unacceptable number of subgroups will be excluded from AYP?	3	5	10	20	30	50	100
4	What *n* is fair to small subgroups in meeting the 95% participation requirement?	3	5	10	20	30	50	100
5	What *n* is fair to small subgroups when the annual objectives reach 80% or higher?	3	5	10	20	30	50	100
6	At what *n* is a small subgroup unlikely to achieve statistical reliability regardless of its performance?	3	5	10	20	30	50	100
7	Is there an *n* below which the state does not wish to judge a subgroup regardless of its performance?	3	5	10	20	30	50	100
Is one of these issues so important that it should determine the minimum *n*?		1	2	3	4	5	6	7



What is the minimum *n* below which a subgroup should be excluded from AYP as statistically unreliable?

______

Should subgroups larger than this minimum *n* be tested to ensure their results are 1

No

Yes

What test of statistical reliability should be used?

____

What level of confidence will be accepted for statistical reliability?

P= ______

1-Tailed Test

2-Tailed Test

OR

What *n* is large enough that a test of statistical reliability should no longer be necessary?

3 5 10 20 30 50 100

Decision Rules:

1. Confidentiality: Do not disaggregate subgroups smaller than ______.

2. Statistical Reliability:

- **a. Do not disaggregate subgroups smaller than ______.**
- **b. Use , p= , -tailed test to determined reliability of larger subgroups.**
- **c. Do not test reliability of subgroups larger than ______.**

Decision Template

Number of Students in a Subgroup	Confidentiality		Number of Students in a Subgroup	Statistical Reliability
	Impact	Comment		Impact
3	The maximum number of subgroups will be disaggregated.	Degrees of freedom = 2; statistically 3 is the smallest number that protects individual identities.	3	Decisions will have a high degree of unreliability. Exception: if all students scored more than *n*SEM from the cut point.
5	Exclusion of subgroups from disaggregation remains minimal; more are included in AYP.	Protects against someone knowing more than one student in a subgroup (e.g., twins, triplets, friends).	5	
10	Exclusion of subgroups from disaggregation increases significantly as this number increases.	All sizes above 5 add decreasing additional protection at the expense of sacrificing the inclusion of subgroups in AYP.	10	Each student counts as 10 percentage points in the subgroup's performance. Example: 100% of a subgroup of 10 must be proficient to meet an annual objective of 91%.
20			20	Each student counts as 5 percentage points in the subgroup's performance. Example: 100% of a subgroup of 20 must be proficient to meet an annual objective of 96%.
30			30	Each student counts as 3.3 percentage points.
50			50	Individual students impact the subgroup's percent less as the number of students increases.
100			100	

Statistical Reliability	Number of Students in a Subgroup	Statistical Reliability
Comment		Impact
Textbooks caution against trying to interpret the significance of groups less than 10. Exception: if all students scored more than *n*SEM from the cut point.	**3**	The minimum for statistical reliability may be set higher than the minimum for confidentiality. A state may set a minimum below which subgroup results will be considered unreliable without calculating any statistical test.
	5	
Minimum number cited as acceptable for use of statistical tests of reliability.	**10**	If the minimum for confidentiality is lower than the minimum for reliability, then highlighting statistically unreliable subgroups among those reported in annual report cards must be considered. *If a subgroup meets the minimum, then the state may elect to calculate a statistical test to determine reliability.* *
Below 20 students, 100% must be included in an assessment to meet the 95% participation requirement.	**20**	At this point, the tension between the benefits to students of identifying schools in need of improvement (INOI) and protecting schools from inappropriate identification as INOI arises. The higher the minimum number goes, the fewer subgroups that will be disaggregated. When fewer subgroups are disaggregated, Fewer schools are identified as INOI.
The National Center for Education Statistics uses 30.	**30**	
Statistical tests can provide a reasonable probability estimate for groups this size.	**50**	*If a subgroup meets the minimum, then the state may elect to calculate a statistical test to determine reliability. A reliability test will most likely reduce the number of subgroups disaggregated, resulting in fewer subgroups failing the annual objective and fewer schools identified as INOI.*
	100	

Confidentiality *n*

When are there too few students in a subgroup to allow disaggregating that will not reveal personally identifiable information for individual students? The intent in No Child Left Behind is to remove the possibility that this accountability system would require states to violate the established federal protection of student privacy as mandated under section 444 (b) of the General Education Provisions Act (Family Educational Rights and Privacy Act (FERPA) of 1974). Thus, if a subgroup is so small that publishing the percent proficient would reveal how an individual student scored, the state is not required to disaggregate the subgroup, and the school is neither responsible for reporting on this subgroup, nor responsible for this subgroup's meeting the annual objectives.

The majority of the content in this section is drawn from two prior papers.

- Ligon, G. D. (1998). Small Cells and Their Cons *(Confidentiality Issues)*: NCES Summer Data Conference.
- Ligon, G. D., Clements, B. S., & Paredes, V. (2000). *Why a Small n is surrounded by Confidentiality: Ensuring Confidentiality and Reliability in Microdatabases and Summary Tables.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans.

ESP Insight
No Child Left Behind and FERPA are aligned.

This discussion makes several assumptions that are necessary to implement the confidentiality intent and methodology of No Child Left Behind.

- The Family Educational Rights and Privacy Act (FERPA) is the primary federal mandate to be followed.
- The values for subgroups with too few students should be suppressed in all public reports.
- Suppressed values should not be recoverable through calculations using other published statistics, e.g., the values of other subgroups or values published in separate documents.
- The existence of a suppressed subgroup should not require the suppression of other sufficiently large subgroups to satisfy the previous assumption.
- The same minimum number of students should apply to all schools, districts, and the state in the calculation of AYP. (This is not specified in the law or regulations, but is an equity issue and a control to avoid manipulation of the rules to benefit individual schools, districts, or states.)

Data collected by governmental agencies must remain confidential in order to protect the privacy of individuals. For the Census Bureau, that information may be related to geographic region, such that information reported for a sparsely populated area can easily be tracked to the few individuals who live in that area. For the Internal Revenue Service, it may be related to income, in that certain income levels are only attained by a few individuals. For educators, it can be information about test scores, disabilities, or socioeconomic status that must be reported in a way that does not reveal information about individual students.

If, for instance, there are two Asian students in the fourth grade of a school and the percent proficient for Asian fourth graders is 50%, the parents of each of those students, knowing their own child's proficiency level, can easily figure the other child's. Alternatively, if there are 100 Hispanic students in the fourth grade, and the percent proficient for Hispanic fourth graders is 100%, then it can be easily determined that each Hispanic student scored at the proficient level. However, important information on subgroups must be reported. Certainly the taxpayers of a school district want to know if students of one gender or ethnicity lag behind others in test achievement. The task becomes finding a way to report enough information while still protecting the privacy of individuals.

Evans, Zayatz, and Slanta (1996) address data confidentiality issues faced by the Bureau of the Census. As in education, "The disclosure limitation problem is to prevent data users from being able to recover any respondent's reported values using values appearing in the published tables" (Evans, et al., 1996). They note that cell suppression is a choice, but while suppressing individual cells can be done relatively easily, suppressing those cells in associated documents can be overwhelming. In this case, if the number of subjects in any cell is less than a certain number, that cell is suppressed from any data presented to the public. While this is fairly simple, it becomes more complicated because those cells may be carried over onto other data tables, and must be suppressed there, as well. In addition, revealing any cells which could lead to the exposure of the values in a small cell must also be suppressed. It is conceivable that this situation could lead to the loss of information for all subgroups. As noted earlier, it is unacceptable in an accountability system to lose information unnecessarily.

ESP Insight

Cell suppression is a choice, but while suppressing individual cells can be done relatively easily, suppressing those cells in associated documents can be overwhelming.
(Evans, et al., 1996)

Adding noise to data tables is suggested as an alternative by Evans, et al. (1996). This means multiplying the data from each establishment by a noise factor before tabulating the data. Over all establishments, the number of positive (>1) and negative (<1) multipliers would be equal, so that they would cancel each other out in the end. Cells which appear in more than one data table would carry the same value to all tables. Zayatz, Moore, and Evans point out, however, that if the number in a cell is too small (1 or 2) it can still be possible to discern a unique contributing entity. Winkler (1997) observes that introducing enough noise to prevent re-identification of records may also make the files analytically invalid.

The method of choice for protecting the confidentiality of student records has been cell suppression. According to numbers reported on state agency web sites, North Carolina and Texas do not report cell sizes fewer than 5, Oregon and Wisconsin fewer than 6, Pennsylvania, Washington, Wyoming, Michigan, Florida fewer than 10, and Colorado and Delaware fewer than 16. As noted above, however, this can lead to problems with suppression of the same cells in other forms of data, or with suppression of other cells which could reveal the information in the sensitive cells.

Moore (1996) identifies three other methods used by the Census Bureau. They are (1) release of data for only a sample of the population, (2) limitation of detail, and (3) top/bottom-coding. Because of the requirements of No Child Left Behind, the first is not practical for the field of education. Information released must be based upon all students in all schools. The second, limitation of detail, is practical and useful in education. The Bureau restricts release of information which would be restricted to a subgroup less than 100,000. Educators use a much smaller limit, but as mentioned above they do, in fact, restrict release of information about subgroups which do not meet a certain size. The third method, top/bottom-coding, is very appropriate to the field of education. The Census Bureau limits reported levels of income because they might identify individuals. So incomes above a certain level, which might lead to identification of individuals, are reported as "over $100,000."

Numbers of students in a subgroup can be reported in a similar way. The following is an example of a way to report information about the percent of students who passed an assessment with a score of "proficient" using limitation of detail. See Table 3.

Table 3: Limitation of Detail Using Ranges for Number of Students

	Total Students	African American	Hispanic	White	Asian	American Indian
% Proficient or Above	77.39	90	85	70	80	*
Number of Students in Group	115	5 to 15	26 to 35	51 to 60	16 to 25	<5

ESP Insight

The use of ranges is an alternative to suppressing small cell values.

For all of the above subgroups except American Indian, the number of students in the group is more than five. Therefore, the percent proficient or above is reported. Because there are fewer than five American Indian students, the percent proficient or above is not reported. In addition, the actual number of students is not reported. In this way, it becomes far more difficult to deduce the percent or number of American Indian students scoring proficient or above. If actual numbers of students in each subgroup were reported, it might become possible, using numbers in groups and percentages, to discern confidential information. In that situation, more cells would have to be suppressed. This method allows for the maximum amount of information to be reported while still protecting the privacy of individuals.

Assessment scores can also be reported using top/bottom coding. Here, the issue is reporting information about how well a subgroup performed without revealing the exact scores of that group. If a range is reported rather than specific score levels the purpose (how the group did on the test) is met, but individual scores cannot be determined. Note that this is especially important at the top and bottom of the scale (scores of zero or 100). See Table 4.

Table 4: Top/Bottom Coding

	Total Students	Score Range				
		>94	75-94	50-74	25-49	<25
Percent of Total	100	13	35	26	22	4
Number of Students in Subgroup	115	15	40	30	25	5

As noted earlier, if this particular subgroup were small, and the average score were 100, it would be obvious that all students earned a score of 100. If, however, a score level of >94 was reported, even if all subgroup students scored in that category, it would be impossible to determine an individual's score.

The reported score range or number of students reported in a group range would depend upon the total number of students in the group. The following could be considered for implementation of the above rules if six or more were used as the number of students in a subgroup for confidentiality purposes. See Table 5.

Table 5: Recommended Ranges for Obfuscating Actual Values

If Total Number of Students is…	Use Percent Above Cut-Point Intervals of…	Use Ranges of Number of Students of…
<6	None	None
6-20	10	25
21-33	5	20
>33	3	5

ESP Insight

A minimum cell size of five will meet the requirements of No Child Left Behind, exceed the statistical minimum of three, and provide states a comfort zone above that minimum.

These statements have been summarized from the review of methodologies used by statistical agencies for masking or suppressing the values of small groups and their relevance to education.

1. From a pure and simple statistical perspective, a minimum subgroup size of three protects the identity of the subgroup's members (degrees of freedom = 2). For example, knowing the value for one member of the subgroup still leaves two values unknown, so the value of any one of the other two cannot be determined. An example of a situation that contradicts the use of three as a minimum is a subgroup containing twins. The family of these two students would know the values for two rather than just one student.

No state uses three as a minimum for reporting their public assessment results (AIR 2002).

2. Most state education agencies, school districts, and other types of agencies exceed this minimum "to be cautious." This protects against someone knowing the values of more than one student in a subgroup.

3. A minimum cell size of five will meet the requirements of No Child Left Behind, exceed the statistical minimum of three, and provide states a comfort zone above that minimum. See Table 6. *Fourteen states use 5 or 6 as a minimum to report their assessment data (AIR 2002).*

4. Minimum cell sizes above five may inappropriately reduce the number of subgroups for which a school is responsible. Excessively high minimums will violate the intent of No Child Left Behind by excluding subgroups and the individual students in them from the accountability mandates of the law. *Twenty-one states use 10 or 11 as the minimum for reporting assessment data; four states have higher minimums up to 31 (AIR 2002).*

Table 6: Minimum Subgroup Size of Five (5) for Confidentiality

GROUP:	All Students	White	African American	Hispanic	Asian Pacific Islander	American Indian	LEP	IEP	Economically Disadvantaged
% Proficient or Advanced	68%	20%	80%	60%	100%	100%	0%	33%	25%
Number Assessed	22	5	5	5	2	5	4	6	8
Met 75% Annual Objective?	No	No	Yes	No	Yes	Yes	No	No	No
Reported Status	Not met	Not Met	Met	Not Met	Too Few to Report	Met	Too Few to Report	Not Met	Not Met
NOTE: This table is irrespective of statistical reliability decisions.					Statistics Not Reported Publicly				

5. For reporting, if a small *n* is present, blanking out that cell in a table may not be an adequate solution. The cell value may be restorable based upon the values of other cells that are reported. See Table 7.

Table 7: Reconstituting Suppressed Cell Values

GROUP:	All Students	White	African American	Hispanic	Asian Pacific Islander	American Indian	LEP	IEP	Economically Disadvantaged
% Proficient or Advanced	68%	20%	80%	60%	***100%***	100%	0%	33%	25%
Number Assessed	22	5	5	5	***2***	5	4	6	8
Met 75% Annual Objective?	No	No	Yes	No	***Yes***	Yes	No	No	No
Reported Status	Not met	Not Met	Met	Not Met	Too Few to Report	Met	Too Few to Report	Not Met	Not Met
NOTE: This table is irrespective of statistical reliability decisions.					Statistics Not Reported Publicly			***Values That Can be Calculated***	

6. If a school has a small subgroup, blanking out that subgroup and all others that might be used to derive that subgroup's value could result in the loss of all subgroups. This should be unacceptable in an accountability system. See Table 8.

Table 8: Loss of Valid Cells to Avoid Disclosing Suppressed Cell Values

GROUP:	All Students	White	African American	Hispanic	Asian Pacific Islander	American Indian	LEP	IEP	Economically Disadvantaged
% Proficient or Advanced	68%	***20%***	***80%***	***60%***	***100%***	***100%***	0%	33%	25%
Number Assessed	22	***5***	***5***	***5***	***2***	***5***	4	6	8
Met 75% Annual Objective?	No	***No***	***Yes***	***No***	***Yes***	***Yes***	No	No	No
Reported Status	Not met	Not Met	Met	Not Met	Too Few to Report	Met	Too Few to Report	Not Met	Not Met
NOTE: This table is irrespective of statistical reliability decisions.					Statistics Not Reported Publicly			***Values That Can be Calculated***	
Values Suppressed to Avoid Calculation of Suppressed Values									

7. As an alternative to blanking out all subgroups when one is too small to report, the values can be reported in ranges (with ranges for the n's as well) that obfuscate the actual values enough to prevent calculations. See Table 9.

Table 9: Loss of Valid Cells to Avoid Disclosing Suppressed Cell Values

GROUP:	All Students	White	African American	Hispanic	Asian Pacific Islander	American Indian	LEP	IEP	Economically Disadvantaged
% Proficient or Advanced	68%	***0 to 20%***	***80 to 100%***	***40 to 60%***	***100%***	***80 to 100%***	0%	33%	25%
Number Assessed	22	***5 to 20***	***5 to 20***	***5 to 20***	***2***	***5 to 20***	4	6	8
Met 75% Annual Objective?	No	***No***	***Yes***	***No***	***Yes***	***Yes***	No	No	No
Reported Status	Not met	Not Met	Met	Not Met	Too Few to Report	Met	Too Few to Report	Not Met	Not Met
NOTE: This table is irrespective of statistical reliability decisions.					Statistics Not Reported Publicly			***Values That Can No Longer be Calculated***	
Values Suppressed to Avoid Calculation of Suppressed Values									

Reliability *n*

No Child Left Behind specifically excludes from disaggregation for adequate yearly progress or annual report cards…

"…a case in which the number of students in a category is insufficient to yield statistically reliable information…"

When are there enough students in a subgroup to yield a statistically reliable measure of a subgroup's performance toward meeting the annual objective established for adequate yearly progress? The regulations are explicit in leaving to the discretion of each state how to determine statistical reliability.

A general expectation has been that a single minimum number, as is the case for confidentiality, could be selected to determine statistical reliability. However, as this publication discusses in detail:

- the number of students,
- the distribution of their scores, and
- the confidence level adopted by the state

determine reliability. States should consider several alternatives. They can rely upon a minimum number of students or perform a statistical test—or both. See Table 10. The intent of No Child Left Behind appears to be to avoid labeling a school as in need of improvement based upon a subgroup with a small number of students. Therefore, a state could choose to establish a minimum number of students, as established for confidentiality, below which a subgroup is eliminated—regardless of how poorly that subgroup performed. This publication explores the added process of examining even larger groups to identify those that may meet a minimum number of students if one is set, but fail to pass a test for statistical reliability.

ESP Insight
There are alternatives to selecting a single minimum cell size for reliability.

This section on reliability is partially based on, with excerpts from, content that originally appeared in a currently unpublished background paper written for the CCSSO CAS/SCASS on Accountability (Ligon, Jennings, and Clements, 2002).

Table 10: Approaches for Statistical Reliability

Approaches for Statistical Reliability		Statistical Test	
		Yes	**No**
Minimum Number	**Yes**	Run a test only if the subgroup already had a minimum number of members	Select a number that ensures enough students to be reliable
	No	Run a test to determine the probability that the subgroup really passed or failed	**Not an Available Option**

ESP Insight

Test measurement error (SEM) can be determined by either a test-retest or an internal consistency formula. Test publishers rarely use test-retest because of the costs, so states and districts only have internal consistency measures available for use.

Standards for Reliability

At least four contrasting methods are available for establishing the statistical reliability of the determination of whether or not a subgroup has met an annual objective. This is not a determination as to whether or not the overall AYP decision for a school is valid or reliable. The four methods are summarized below in Table 11.

Table 11: Methods for Determining a Subgroup's Annual Objective

METHOD	DESCRIPTION
Minimum n	With all factors considered there is a minimum number of students that a subgroup should have before being included in an accountability system. Therefore, a single minimum is selected below which a subgroup's results are considered to be based upon too few students to be reliable. No Child Left Behind and the regulations refer to a minimum number of students for statistical reliability without reference to the consideration of their distribution of scores or a confidence level.
Student Sampling Error	The students who are tested in a school (i.e., subgroup) in a year are a sample of the students who might have been tested or who will be tested over the years. Therefore, reliability is based upon how much the results for a year would vary with different samples of students drawn from the same population. Sue Rigney, U.S. Department of Education, has described this perspective as, "AYP spans 12 years and requires continuous improvement for schools that are below the target for a given year. Genuine school level gains on the assessment over time are confounded with the cohort effect. Sampling error is widely recognized as the appropriate error term in this case."
Test Measurement Error	The students tested in a subgroup in a year are the population of students for that subgroup—not a sample chosen from a larger population. Therefore, reliability is based upon how much those students' individual scores would vary if they were retested multiple times. The students tested in a subgroup in a given year are the complete set of students (i.e., population) used for determining AYP. Sampling error can only be estimated because a population beyond these students cannot be measured, so reliability can best be based upon measurement error.
School Measurement Error	The distribution of the percent of students performing at or above proficiency across all schools represents the performance of a population from which each school's students are drawn each year. Therefore, reliability is based upon a confidence interval established around each school's percent. The actual distribution of school level results is the best basis for establishing how much a school's percent might vary across years.

Cronbach et al. (1995) considers a subgroup's scores as either a sample from an infinite population or as the population.

A traditional analysis treats pupils as randomly sampled from an infinite population. In the present context (estimating school-level error) an infinite population would be assumed to exist for each school, and the pupils tested would be conceived of as a random sample from the population associated with the school. The infinite population of pupils associated with the school is obviously hypothetical. Alternatively, the population may be limited to the actual student body, the MN (M = number of classes in a school, and N = number of pupils tested in every class) is the school and grade this year.

ESP Insight
Since all eligible students were tested, rather than a sample of students, this population value will not fluctuate statistically. (Jaeger and Tucker, 1998)

Jaeger and Tucker (1998) state that:

> Figures that define results for an entire population of individuals are regarded by statisticians as immutable. Since all eligible students were tested, rather than a sample of students, this population value will not fluctuate statistically.
>
> To a statistician, an average score on an achievement test, computed for, say, every student in a particular racial or ethnic group in a school district, would be considered a population parameter. But to a measurement specialist, such an average would be a statistic that estimated what the students' true average score would be, were it possible to administer an infinite number of different forms of the achievement test to the population of students on an infinite number of occasions, provided the students' true achievement did not vary across test forms or occasions. The difference in perspectives between the statistician and the measurement specialist is that the statistician only considers sampling fluctuations across samples of students to be a source of error in trying to estimate a population parameter. The measurement specialist also considers measurement error across test forms and testing occasions, regarding a single administration of one form of a test to be a sample of students' performance across all possible forms and occasions that leave the students' true performances intact.

ESP Insight
A straightforward interpretation of AYP is that the students in a subgroup are a finite population representing all those taught and tested in a given year by the school.

They also state that "from another perspective, one could argue that the … result was not only a consequence of the quality of education provided … but occurred in part because of the particular students who happened to be (enrolled)… in part to the differences between the backgrounds of students who happened to be enrolled … during the two school years. If the (current students) are considered to be a sample drawn from a larger population … who might be enrolled across the years,… the percent … would be regarded as a sample statistic rather than a population parameter."

With which part of the experts' statements will each state align? A straightforward interpretation of AYP is that the students in a subgroup are a finite population representing all those taught and tested in a given year by the school. A broader interpretation is that one year's cohort of students is one sample from all the cohorts which will be passing through the school during the 12-year span of No Child Left Behind.

What is the question that is being addressed? Here each method differs.

METHOD	QUESTION ADDRESSED
Minimum n	Are there sufficient students in this subgroup to meet the minimum standard for reliability?
Student Sampling Error	How would this school perform with multiple samples of students?
Test Measurement Error	How would these students perform if retested multiple times?
School Measurement Error	How would this school perform over multiple test administrations?

Because the samples cannot vary less than the error already present in the measurement itself, sampling error is typically larger than measurement error. Methods based upon student sampling error assume that the measurement error is reflected in the sampling variance. Methods based upon measurement error assume that sampling error is not relevant. Methods based upon school measurement error assume both student sampling error and measurement error are reflected in the variance observed across schools.

No Child Left Behind asks, "Are there enough students in the subgroup to ensure that we would classify a subgroup's performance on an annual objective the same (i.e., avoid declaring a school as failing when the school has an acceptable probability of passing) if…?" The "if" varies depending upon the method applied.

METHOD	Are there enough students in the subgroup to ensure that we would classify a subgroup's performance on an annual objective the same if…
Minimum n	More students had been tested?
Student Sampling Error	A different sample of students had been drawn from the same population?
Test Measurement Error	The same students were retested?
School Sampling Error	The school is measured again at another time?

The following statements provide the basis for establishing reliability.

1. The lowest level question for a state to answer as posed by No Child Left Behind is simply:

   Did the subgroup meet the annual objective? (Of course there are multiple annual objectives for multiple indicators, and each subgroup must meet the annual objective for each one individually.)

2. This accountability question translates to whether or not an equal or greater percent of the subgroup's students performed at the proficient level or higher on the indicator compared to the percent established by the state as the annual objective.

3. The denominator for this percent is the number of students who:
   a. Have a valid assessment score or other determination of their proficiency level (excluding those not tested),
   b. Were enrolled for a full school year as defined by the state, and
   c. Belong to the subgroup (or total) being measured.

4. For a school to disaggregate a subgroup, there must also be enough students in the subgroup to protect the individual identity of each student when the results are reported. If there are too few, then a determination of whether or not this subgroup met the annual objective is unnecessary because it will not be disaggregated. (This may differ if a state adopts different minimum numbers for AYP and annual report cards.)

5. For a school to disaggregate a subgroup, there must be enough students in the subgroup to ensure statistical reliability, i.e., there is a reasonable level of confidence that the decision made about the subgroup is the right one. Alternatively, if the state does not set a minimum number for statistical reliability, then the rule adopted for establishing reliability would be applied.

6. Sequentially, a determination of whether the subgroup is large enough to protect individual students' identities should be first. Then if this standard is met, a determination of statistical reliability should be made.

ESP Insight
Reliability is getting the same result and drawing the same conclusion from more than one observation. Yet reliability is not validity. Invalid measure might be very reliable.

The Authors' Perspective on Error

After carefully thinking through all of these issues, we determined that SEM is the best error estimate for AYP. However, because other perspectives have been published extensively, they are also presented here. We would expect that a state could present to the USED an acceptable rationale for any of the four methods described above. Certainly the law and the regulations do not exclude any justifiable method.

ESP Insight
SEM is the best error estimate for AYP.

We have not recommended traditional reliability statistics (sampling error) because they depend upon some basic assumptions about students, schools, and scores that are not generally true. Ask principals if the students in their schools are randomly drawn from a population. They know they are not. Ask principals if they understand that test scores can vary just by retesting the same students. They know that to be true.

ESP Insight
We have not recommended traditional reliability statistics (sampling error) because they depend upon some basic assumptions about students, schools, and scores that are not generally true.

- Districts and schools do not draw their students randomly from a designated population of students. (See Figure 3.) The political process of drawing district and school boundaries is not random.

ESP Insight

The reliability question is "how confident are we that 75% of the students perform at the proficiency level?" Not "how confident are we that 75% of a different sample of students would also perform at the proficiency level?"

- The students tested one year are not necessarily the same as those tested the next. The subgroups and the students in them may vary considerably from one year to the next.
- The assessment scores and the student performance levels derived from them are not likely to be normally distributed.
- School statistics are not likely to be normally distributed, e.g., the percent of students above the criterion for proficiency is not a statistic that is normally distributed across schools.
- The variance of scores around a school's mean is irrelevant to the determination of whether or not a school meets its annual objective.

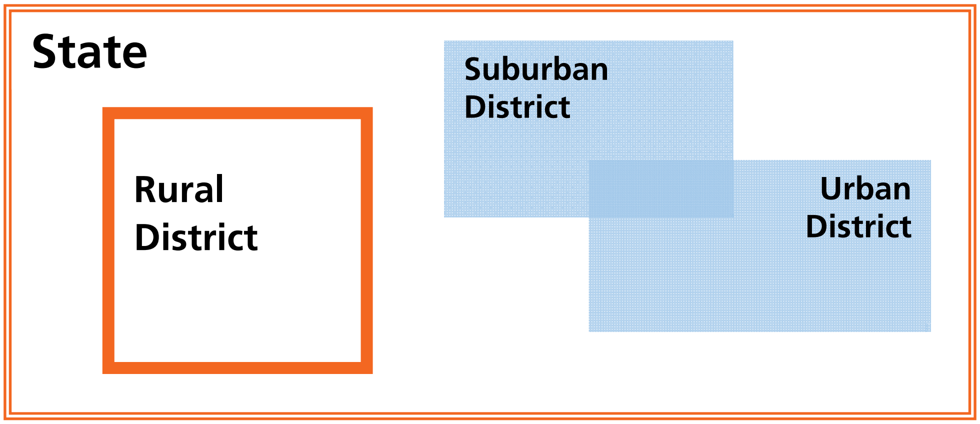
Six Degrees of Separation

Harvard Sociologist Stanley Milgram in 1967 found that only five or fewer connections between people were needed to link strangers. Using 100 friends, it is possible to "know" the whole planet within 5 steps. Then Strogatz and Watts from Cornell pointed out that all these connections are not mutually exclusive—our friends know the same friends as we do. Strogatz says, "We are very much constrained by our socioeconomic status, geographical location, background, education, professions, interests, and hobbies. All these things make our circle of acquaintances highly nonrandom."

The same applies to students in schools. They are not there by any probabilistic order or randomness. They are there because a political body drew boundary lines, because public housing is available, because private schools are too expensive for them, because the family moved. They do not enroll as random samples from an attendance area. **The students in a school each year are their own population. See Figure 3.**

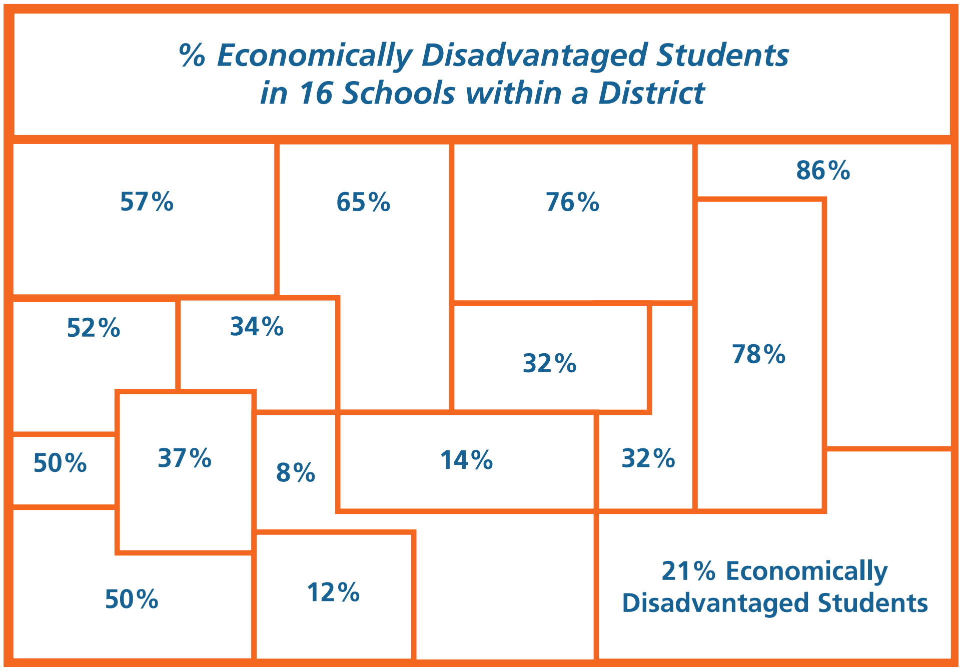
With these conditions, the assumptions required to use our favorite statistical tests are violated. The tests rendered inappropriate include multiple-linear regression, analysis of variance or covariance, and other popular parametric analyses. Statistical tests that rely upon the standard error, variance, or deviation are all based upon assumptions about and the use of means—either student score means or school means. Even if the means are of the percent of students performing at or above the criterion for proficiency across all schools (which are the means used to establish a traditional confidence interval), these confidence intervals result in illogical and unacceptable conclusions about schools. For example, a school with only five students, but all of them making perfect scores, may be designated as statistically unreliable—even though all five scored so high above the criterion for proficiency that not a single one of their scores is in doubt. No matter how many times these students are tested, the status of this small group's success must be considered statistically reliable.

Figure 3: Are Districts Randomly Sampled from State Populations?



	Proficient & Advanced	White	African American	Hispanic	Asian, PI	American Indian	Econ. Disadv.	LEP	IEP
State	75%	50%	12%	12%	1%	4%	45%	10%	12%
Rural District	80%	75%	10%	15%	2%	5%	25%	12%	12%
Urban District	65%	25%	20%	10%	8%	1%	65%	6%	16%
Suburban District	95%	80%	2%	5%	8%	0%	7%	1%	9%

Figure 4: Are Students Randomly Sampled from Districts for Schools?


% Economically Disadvantaged Students
in 16 Schools within a District
57%
65%
76%
86%
52%
34%
32%
78%
50%
37%
8%
14%
32%
50%
12%
21% Economically Disadvantaged Students

States that choose to adopt a methodology based upon sampling error will follow the lead of many educational researchers who apply statistical techniques to assessment data even though all of the assumptions underlying the analyses are not met. They will be able to say that they are comfortable with the degree to which the assumptions are met. States that choose to adopt a methodology based upon measurement error will do so based upon the conclusion that the appropriate statistic is one that describes how reliable each student's score is, and makes a determination of the probability that enough of the students whose individual scores are in doubt performed at the proficient level to allow the school to meet its annual objective.

ESP Insight
A small subgroup can be more reliable than a large subgroup.

There is one basic assumption that all statistical tests must meet. The assumption is that enough students are measured to make sense. Making sense for No Child Left Behind is simply, "Are there enough students in the subgroup to ensure that we avoid declaring a school as failing when the school has an acceptable probability of passing if the students were retested?" In other words, measurement error users do not want to declare a school as low performing if too many of a subgroup's basic-level students scored within the test's standard error of measurement or if the school or subgroup scored close enough to the annual objective to leave room for doubt. Sampling error users do not want to declare a school as low performing if another sample would likely have scored above the annual objective. Both determinations are at least partially determined based upon the number of students are included in the analysis.

This changes the focus from simply how many students are in the subgroup to how many students in the subgroup are performing so close to the criterion that their proficiency level is in doubt (measurement error) or do samples of this size vary enough to allow this sample to cross over the annual objective (sampling error). ***Read this closely.*** **A small subgroup can be more reliable than a large subgroup.**

Now we can move on to the more controversial statements. A strict interpretation of AYP is that statistical reliability is required for a single subgroup, for the given year being measured. After all, the decision being made is whether or not to include that subgroup as part of a school's AYP determination. This interpretation would not demand acceptance of Cronbach's definition of an effective school. The bottom line here is that many people are challenging AYP as a methodology, attempting to measure its validity and reliability. That is not the goal of this publication. Here the intent is to describe methodologies that comply with AYP. Not to validate or challenge its basic tenants.

AYP is an annual status. Measures of AYP in different years are not independent observations of a school at the same point in time. The assumptions that must be made to use a stability analysis for AYP are not likely to be met for an individual school (e.g., same students each year, students randomly assigned from a larger population each year, no change in intervening variables, etc.). Therefore, it is at least practical, if not logical, to consider the annual determinations of a school's AYP status as reflecting different years (and different groups of students, faculty, resources, etc.). Thus, the estimates of statistical reliability should focus on each year's measures with the realization that a school's status may change back and forth if that school is performing close to the cut point of the annual objectives.

1. Statistical reliability for subgroups is not a stability measure across years, but a determination of the confidence we have in the current year's status of the school.

Students may be assigned to a school from a population, but that assignment is far from random. An individual school cannot be assumed to have a random or representative sample of students from the district's population of students. Even within the school's attendance zone, which students enroll is not random or representative. So, analyses that rely upon sample statistics do not fit well for a school that is actually its own population. AYP is an individual school determination that is not at all based upon the norms or averages of a district from which the school's students are drawn. So our second statement is:

2. A school's students are not randomly or representatively drawn from a district's population, but make up the school's own population. Therefore sample statistics do not apply to the determination of statistical reliability.

No Child Left Behind by its name makes it clear that we are to look at individual students, not averages. The AYP methodology counts students and determines percents rather than using any mean scores to determine AYP. So our third statement is:

3. Parametric statistics do not meet all of their assumptions because we are counting students who perform in categories rather than calculating any measures of central tendency based upon continuous variables.

Because every individual student counts, the reliability of the decision of each student's performance level is the key. The standard error of measurement (SEM) is the metric that describes how reliable an individual student's score is on an assessment. So our next statement is:

4. The standard error of measurement (SEM) at the cut point for each assessment (at each grade level, in each area) estimates the reliability of a student's actual performance at the criterion for proficiency. (The SEM must be calculated at the criterion.) Students who perform within *n*SEM (whatever range is adopted) of the criterion are the ones whose statuses are in doubt.

5. Students who perform outside the confidence interval defined by the SEM can be considered to have a statistically reliable placement at the basic or proficient levels.

6. The number of students performing within the SEM range of the criterion score determines the statistical reliability of the school's status on AYP.

7. If a school has enough students scoring more than 1 SEM above the criterion score to satisfy the annual objective, then that school's status of meeting AYP can be considered to be statistically reliable—regardless of the total number of students.

ESP Insight

The real instance when statistical reliability matters is when a subgroup fails to meet the annual objective for AYP.

8. If a school has enough students scoring more than 1 SEM below the criterion score, then that school's status of not meeting AYP can be considered to be statistically reliable regardless of the total number of students.

In No Child Left Behind, there are consequences for not meeting AYP. However, meeting AYP or suppressing a subgroup because there are too few students or because the results are not statistically reliable are of equivalent consequence. In each case, the school avoids being classified as low performing. This means that the consequential decision for suppression of a subgroup's performance occurs when the subgroup does not meet the annual objective for AYP. (If a state is rewarding and recognizing schools for consistently meeting AYP standards, a reliability test for schools meeting AYP may be desirable as well.)

9. The real instance when statistical reliability matters is when a subgroup fails to meet the annual objective for AYP.

10. For schools not meeting the annual objective, if a school has enough students within *n*SEM below the criterion to place the school's failing status

in doubt, then a *Chi*- square test can be applied to determine the probability that enough of these students have true scores at or above the criterion to categorize the school as meeting AYP. In other words, if four more students passing would have met the annual objective, and the school has at least four students performing within *n*SEM of the criterion, then the status of the school can be considered to be in doubt and can be tested using Chi-square.

11. If a school's probability of having enough students below the criterion is high enough, then the school's status of not meeting AYP is statistically reliable.

Methods have been used that calculate a SEM based upon school distributions (Yen, 1997). These result in a confidence interval for a school's percent above the criterion score that is based upon the total number of students in the school. However, the resultant confidence interval is applied to every school regardless of whether or not the school has any students within 1 SEM of the criterion score. This can result in labeling a school as statistically unreliable when in fact not a single student's proficiency level is in doubt. So we need this statement.

12. A confidence interval based upon distributions of school-level percents and school size may inappropriately label some schools as unreliable.

The intent of No Child Left Behind relative to statistical reliability appears to be a fairness issue. This fairness issue is to avoid a school's being considered low-performing based upon so few students that the results might have been different just by retesting the students the next day. The wording of the law makes it clear that the intent is to find small subgroups that are unreliable. The intent does not appear to have been to declare large subgroups to be unreliable under any unanticipated circumstance. When a significant number of students score perilously close to the criterion for proficiency, the results for even a very large subgroup can be considered to be unreliable. However, a state could be considered to be complying with the intent of the law if the results for any subgroup over an established, reasonable number (e.g., 100) are declared to be statistically reliable. However, this does not mean that all subgroups under 100 are statistically unreliable.

Why the Standard Error of Measurement for Individual Student Scores at the Criterion for Passing is the Appropriate Metric for Judging the Statistical Reliability of a Subgroup's Status on Meeting an Annual Objective for Adequate Yearly Progress

Standard error of measurement (SEM) is used as an acknowledgement of the fact that a student's true ability can never be measured with absolute accuracy. Whereas height, weight, and speed can be directly observed, academic achievement or ability cannot. However, educational scientists have found ways of approximating true ability using observed measures, such as test scores. When a test is constructed, the relation between scores on that test and true abilities can be computed. Confidence bands are established around observed test scores, indicating what range of true abilities each test score represents. If, for instance, a student has a raw score of 35 on an achievement test, and the SEM is 2 items, then it can be said with 68% accuracy that the student's true ability falls within plus or minus one SEM, or a raw score between 33 and 37 on the test. As larger bands are constructed around the observed test scores, true score can be estimated with more confidence. Using the above example, it can be said with 95% accuracy that the same student's true ability would be measured at between 31 and 39 (2 SEM) on that same test. Thus, the observed score of 35 is a good approximation of the student's true ability as long as we are comfortable with the confidence interval established.

A big question arises with schools that do not have enough students scoring above the cut score to be deemed "passing," but with a large number of students scoring within one SEM below the cut score. In this situation, it could be that the students' true abilities are in fact high enough for them to have scored above the cut score, and in fact it was only measurement imprecision and errors that caused some of them to fail. If the same students were tested again on another day, there is a describable probability they would score above the cut score, based upon the SEM of the test.

If a cut score were set at 35, with an SEM of 2 points, and the student achieved a score of 40, it could be said that the judgment of "passing" was reliably made for that student because the score is more than two SEM from the cut point. Similarly, if the student had a score of 30, it could be said that the judgment of "failing" was reliably made for that student. If, however, the cut score were set at 35 and the student scored a 35, keeping in mind that the student's true ability ranges from 31 to 39, a judgment of "passing" or "failing" is much less reliable. These are the students whose status is in doubt when determining a school's status on an annual objective.

Which Subgroups Require a Reliability Determination?

The consequential decision for suppression of a subgroup's performance based upon a lack of statistical reliability occurs when the subgroup does not meet the annual objective for AYP. In No Child Left Behind, there are negative consequences for not meeting AYP. However, meeting AYP or suppressing a subgroup because there are too few students or because the results are not statistically reliable are of equivalent negative consequence. In each case, the school avoids being classified as in need of improvement. (If a state is rewarding and recognizing schools for consistently meeting AYP standards, a reliability test for schools meeting AYP may be desirable as well.)

ESP Insight

A state's plan may accept the "met" status for all subgroups without applying a statistical test for reliability. The subgroup would still need to meet the minimum n for confidentiality, and any minimum n for reliability if one has been set by the state.

- The real instance when statistical reliability matters is when a subgroup fails to meet the annual objective for AYP.

A state's plan may accept the "met" status for all subgroups without applying a statistical test for reliability. The subgroup would still need to meet the minimum *n* for confidentiality, and any minimum *n* for reliability if one has been set by the state.

The intent of No Child Left Behind relative to statistical reliability appears to be a fairness issue. This fairness issue is to avoid a school's being considered low-performing based upon so few students that the results might have been different just by retesting the students the next day. The wording of the law clearly recognizes that too few students in a subgroup impacts reliability, but does not address declaring large subgroups to be unreliable under any unanticipated circumstance. When a significant number of students score perilously close to the criterion for proficiency, the results for even a very large subgroup can be considered to be unreliable. However, a state could be considered to be complying with the intent of the law if the results for any subgroup over an established, reasonable number (e.g., 50) are declared to be statistically reliable. However, this should not mean that all subgroups under 50 are statistically unreliable.

Alternative Methods for Establishing Statistical Reliability

Table 12: Alternative Methods for Establishing Statistical Reliability

METHOD	TEST	DESCRIPTION
Minimum n	None	State selects a number below which disaggregation is not performed.
Student Sampling Error	Binomial Test	Gravetter & Wallnau, 2000, page 623
	Test of Frequencies	Pearson Chi-Square, Hays, 1994, page 369
	(School-Level Options)	(Direct Computation, Split-Half, Random Draws with Replacement, Monte Carlo, Correlations)
Test Measurement Error	Reliable Cut-Point	State cut-point is higher than desired performance level.
	Confident Cut-Point	State accepts student performance within *n*SEM of desired level.
	SEM Test	State counts as proficient only student *n*SEM above cut-point.
	Fail-Safe Test	Pearson Chi-Square test for association, Hays, 1994, page 369, with students within *n*SEM of cut-point.
School Measurement Error	Model I, II, or III	Sliding scale established using distribution of school percent proficient across schools of various sizes.

The following descriptions have been provided to various groups studying the options for statistical reliability. Much debate has occurred, based mainly around the assumptions required for each analysis.

The Minimum *n*

Procedure: Establish a set minimum number with face validity (e.g., high enough to instill confidence, but low enough to avoid the look of attempting to exclude too many small schools). The state might defer to the rule established for confidentiality and assume that if a school has sufficient numbers of students to mask identities that there are enough students to yield a reliable measure. However, if this is five, a number that low may not have the required face validity.

The National Center for Education Statistics uses 30 as a minimum. In the textbooks, 30 is where the graphs start leveling out, meaning the benefit for going higher begins to lessen. However, 30 would eliminate substantial numbers of subgroups and even whole schools from the accountability process. Jeager and Tucker (1998) used 10 as a minimum for reporting in a sample scenario. There is a point at which an annual objective becomes a virtual 100% objective for a small school. For example, if the annual objective is 81%, a group of five students must be 100% proficient, because four students are only 80%. Examining that issue shows that a group of 20 students does not reach a 100% requirement until the annual objective is 96%. This is a leveling out point in the chart. Thus a criterion of 20 would have some face validity in that it is high enough to delay the 100% virtual standard, but low enough to include most schools with only one class per grade level.

An example:

Total Students:	8
Students Performing at the Basic Level:	0
Students Performing at the Proficient Level:	0
Students Performing at the Advanced Level:	8

If the cut point for the advanced level is more than two standard error of measurement (SEM) units above the cut point for the proficient level, the argument that this small subgroup's meeting of the annual objective (even the year 12 objective of 100% proficiency) is statistically reliable is very strong.

Another example:

Total Students:	8
Students Performing at the Basic Level*:	8 *2 SEM below the Cut Point
Students Performing at the Proficient Level:	0
Students Performing at the Advanced Level:	0

If the scores for all eight of these students fall more than two SEM below the cut point for proficiency, the argument that this subgroup's failing to meet the annual objective is not statistically reliable is very weak.

The bottom line for this alternative is that the face validity of the single-criterion minimum number for statistical reliability must be politically very strong. There will be subgroups of students that clearly appear to have failed or passed but have too few students to be counted.

Another perspective on this is to look at how influential a single student is on a subgroup's performance. In a subgroup of five students, a change in one student's performance results in a change of 20%. In a subgroup of 15 students, a change in one student's performance results in a change of 6.7%. From 15 on, the impact lessens. At 100 students, a change in one student's performance results in a change in 1%. This is important because annual objectives may increment only one or two percentage points each year. Small groups would have to make substantially larger percentage gains to meet an annual objective that might go up only one percent in a year. An example is provided in Table 13.

Table 13: Change Required for Small Groups to Improve One Percentage Point

Subgroup Size	Percent Represented by One Student	Percent Required to Meet Year One Annual Objective of 80%	Percent Required to Meet Year Two Annual Objective of 81%	Change Required
5	20.0%	80%	100.0%	20.0%
15	6.7%	80%	86.7%	6.7%
20	5.0%	80%	85.0%	5.0%
30	3.3%	80%	83.3%	3.3%
75	1.3%	80%	81.3%	1.3%
100	1.0%	80%	81%	1.0%

This illustrates that the precision at which annual objectives may be measured imposes greater performance standards on small groups than on large ones that can more precisely match an annual objective.

A minimum *n* ranging from 20 to 30 should be considered.

Student Sampling Error

Alternatives described for this methodology are based upon the assumption that students within a subgroup are sampled from a population and that if different samples are drawn, the results would vary within a range established by probability based upon either the normal curve or the binomial distribution.

The Binomial Test

Procedure: Consider that the percent proficient or above for a subgroup is one value for a sample of students and that other samples from the same population would range around that value. The probability of the actual percent observed being above the cut point for the annual objective can be established.

To illustrate this, consider a subgroup of 20 students, 7 performing at the proficient/advanced levels and 13 at the basic level. The annual objective is 40% proficient/advanced, so the 35% for this subgroup does not meet the annual objective.

To test the probability that this subgroup's true percent is 40%, (H0: p=.40) the following formula is used.

$$z = X / n - p$$

$$\sqrt{pq / n}$$

$$= .35 - .40 \qquad = -.459 \qquad \text{p=.677}$$

$$\sqrt{(.4 * .6) / 20}$$

The probability is 67% that the subgroup's true value is 40%. The null hypothesis is not rejected.

Where:

1. X/n is the proportion of individuals in the sample who are classified in category A.
2. p is the hypothesized value (from Ho) for the proportion of individuals in the population who are classified in category A.
3. pq/n is the standard error for the sampling distribution of X/n and provides a measure of the standard distance between the sample statistic (X/n) and the population parameter (p).

(Gravetter & Wallnau, 2000, page 623)

Test of Frequencies

Procedure: Consider that the percent proficient or above for a subgroup is one value for a sample of students and that other samples from the same population would range around that value. The probability that the actual distribution of students at each performance level would be above the annual objective with subsequent samples can be established.

The formula for the Pearson chi-square test is

__P= j(fj_mj)_
mj

where f_j is the obtained frequency in category j
m_j is the expected frequency in that same category j

(Hayes, 1994, page 369)

Other sampling-based reliability tests are described by Hill (2002). As stated in his publication, these are tests to determine "the reliability of an accountability system." They are described in relation to the reliability of a school-level decision, not decisions for individual subgroups.

Direct Computation

Procedure: Compute the errors around estimates using areas under the normal curve to determine the probability of a correct classification. To apply this to AYP, a state would need to determine how to calculate the appropriate mean and standard deviation relative to the percent of students performing at or above proficiency.

Split-Half

Procedure: Randomly divide a subgroup into two samples and test the difference between them. The technique is problematic with small groups because the group size is reduced to half, thus the probability of statistical reliability is reduced as well.

Random Draws with Replacement

Procedure: Draw random samples repeatedly from the subgroup and test the differences across these samples. Rogosa (1999) detailed this sampling technique. This is similar to "bootstrapping," which typically calls for 500 or more samples to be drawn for analysis (Opperdoes, 1997). Multiplying 500 times the number of subgroups in a state would produce a very large number of samples to be drawn and analyzed. Fortunately, today's computers can handle this task if someone has the time to run them.

Monte Carlo

Procedure: Using estimates of the parameters of a subgroup, draw repeated samples for the school. As with the prior method, multiple random draws are required.

Linn and Haug (2002) correlated school ratings based upon a weighted standard score computed from student proficiency levels across years. This provided a measure of stability for the ratings, but did not explore the reliability of subgroups either within or across years.

Correlations

Procedure: Compute the correlation of school (subgroup) ratings (or percents) across years to establish a typical relationship. A state's plan would need to describe how to use these correlations to establish reliability separate from real changes that would occur from improvement over time.

A necessity for all student sampling error methods is to know or estimate the population mean and variance. A state's plan would need to make clear if the same population parameters are estimated for all subgroups or if separate estimates would be made for each subgroup. The latter seems reasonable because subgroups have performed differently on state assessments over the years. Knowing the actual population parameters is unlikely. The best estimate of the population parameters is typically the sample parameters. In this case, the rationale for a sampling-based methodology is diluted because the population and the sample must have the same mean.

Test Measurement Error

The Criterion- Versus Norm-Referenced Analogy

No Child Left Behind clearly supports criterion-referenced assessments and a standards-based accountability system. The sampling-based methodologies described above are logically aligned with the norm-referenced models of assessment and accountability where continuous scales, means, and standard deviations are employed. Criterion-referenced assessments make pass-fail judgments about students. A well-designed criterion-referenced test will not provide the distribution of student scores that is assumed for parametric statistics, e.g., normal distributions. The best illustration of this is year 12 when schools are expected to be at 100% proficiency. Neither the distribution of student scores nor the distribution of school percents will be normally distributed. Both will be highly negatively skewed.

The test measurement error methods avoid the issues of normal distributions for scores.

The first two alternatives in this category are fascinating, because they provide a rationale for declaring every subgroup to be statistically reliable without the application of statistical test for reliability.

The Reliable Cut Point

Procedure: Assert that when the state's cut score was adopted for proficiency that the SEM was accounted for by raising the required cut score by *n*SEM to ensure that no student would be categorized as proficient unless that student's score was statistically reliably above the performance criterion desired by the state. With this assertion, every school's percent proficient would have statistical reliability and no subgroups would be disregarded (other than for having too few students for confidentiality). All subgroups with fewer students than required for confidentiality could be considered to be statistically unreliable as well. This is a moot point because they would not be disaggregated and would be excluded any way.

We have not found any state (yet) that has made this assertion. This simply means that the cut score is set high enough that there is little doubt that a student meets the passing standard if that score is attained. This would be a policy decision. This one negates the need for any statistical analyses.

This method would result in a lower starting point for setting annual objectives because a lower percent of students would meet the criterion. However, meeting the 100% goal in 12 years would be more difficult.

The Confident Cut Point

Procedure: Assert that every student who scores within *n*SEM of the cut score for proficiency has an acceptable statistical probability of actually being at or above the cut score; therefore, the real cut score for determining proficiency for AYP is *n*SEM lower. With this assertion, every school's percent proficient would have statistical reliability and no subgroups would be disregarded. All subgroups with fewer students than required for confidentiality could be considered to be statistically unreliable as well. This is a moot point because they would not be reported and would be excluded anyway.

This is also a policy decision that would negate the need for any statistical analyses. This might be characterized as lowering a state's standard, but in reality it is an acknowledgement that the state's test is not precise enough to fail with confidence a student who gets this close to passing.

The SEM Test

Procedure: Consider all schools and subgroups with fewer students than required for confidentiality to be statistically unreliable as well, or set a higher minimum (e.g., 20) based upon the examples presented in the minimum n alternative above. For all other schools and subgroups, calculate the percent of students scoring *n*SEM above the cut point for each group and use that percent as a statistically reliable measure of whether the school met the annual objective. The effect of this is to raise the criterion by the size of the SEM as described in alternative 1 above. Then for those schools not meeting the annual objective but with some students performing above but within *n*SEM of the criterion, test the probability that enough of the students above but within *n*SEM of the cut point would be above the cut point with multiple observations (*Chi*-square) and declare schools with sufficient numbers to be statistically reliable and to have met the annual objective. All other schools with sufficient numbers of students above the cut point but not above *n*SEM will be considered statistically unreliable.

For subgroups with too few students above the cut point, identify those with enough students more than *n*SEM below the cut point to have not met the annual objective. These subgroups did not meet the annual objective and are statistically reliable. Calculate a *Chi*-square and probability for subgroups with enough students within *n*SEM to have a probability of meeting the annual objective to determine those that will be considered statistically unreliable (i.e., enough students score close enough to establish the possibility that they could have scored higher if retested).

Figure 5 provides an overview of how Alternative 3 might work.

Figure 5: Flow Chart for Determining Statistical Reliability – SEM Test Alternative

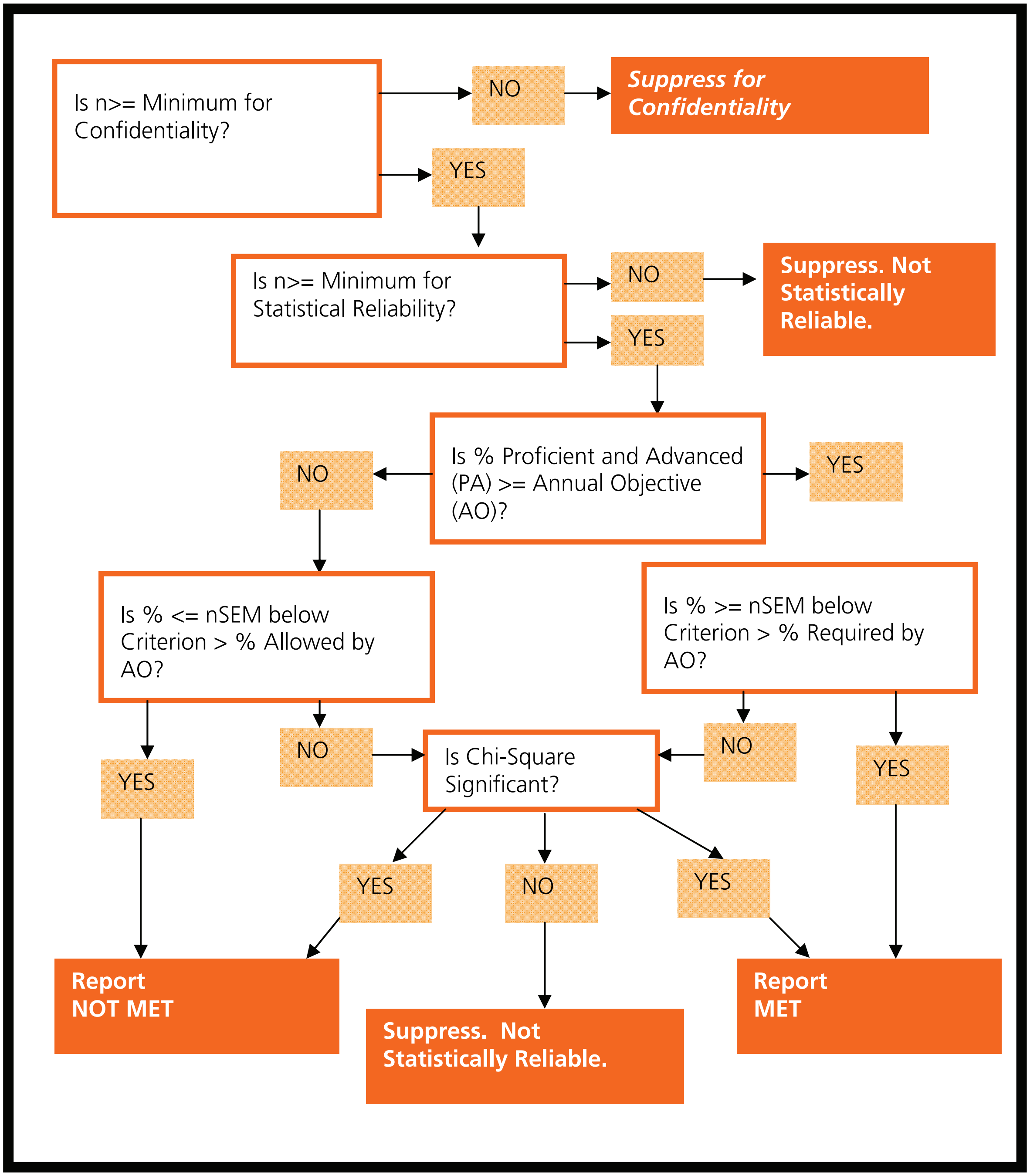

Is n>= Minimum for Confidentiality?
NO
Suppress for Confidentiality
YES
Is n>= Minimum for Statistical Reliability?
NO
Suppress. Not Statistically Reliable.
YES
Is % Proficient and Advanced (PA) >= Annual Objective (AO)?
NO
YES
Is % <= nSEM below Criterion > % Allowed by AO?
Is % >= nSEM below Criterion > % Required by AO?
YES
NO
Is Chi-Square Significant?
NO
YES
YES
NO
YES
Report NOT MET
Suppress. Not Statistically Reliable.
Report MET

Table 14 provides some sample schools using an annual objective of 75%.

Table 14: Sample Schools Using an Annual Objective of 75%

Situation	Total Number of Students	Number Needed to Pass	Passed by More than 1 SEM	Passed, but within 1 SEM	Failed, but Within 1 SEM	Failed by More than 1 SEM	Reliable/Not Reliable
A	100	75	75	3	7	15	Reliably passed
B	100	75	65	10	10	15	Not reliable
C	100	75	30	0	0	70	Reliably failed
D	10	8	8	0	2	0	Reliably passed
E	4	3	0	1	0	3	Reliably passed
F	25	19	3	10	10	2	Not reliable

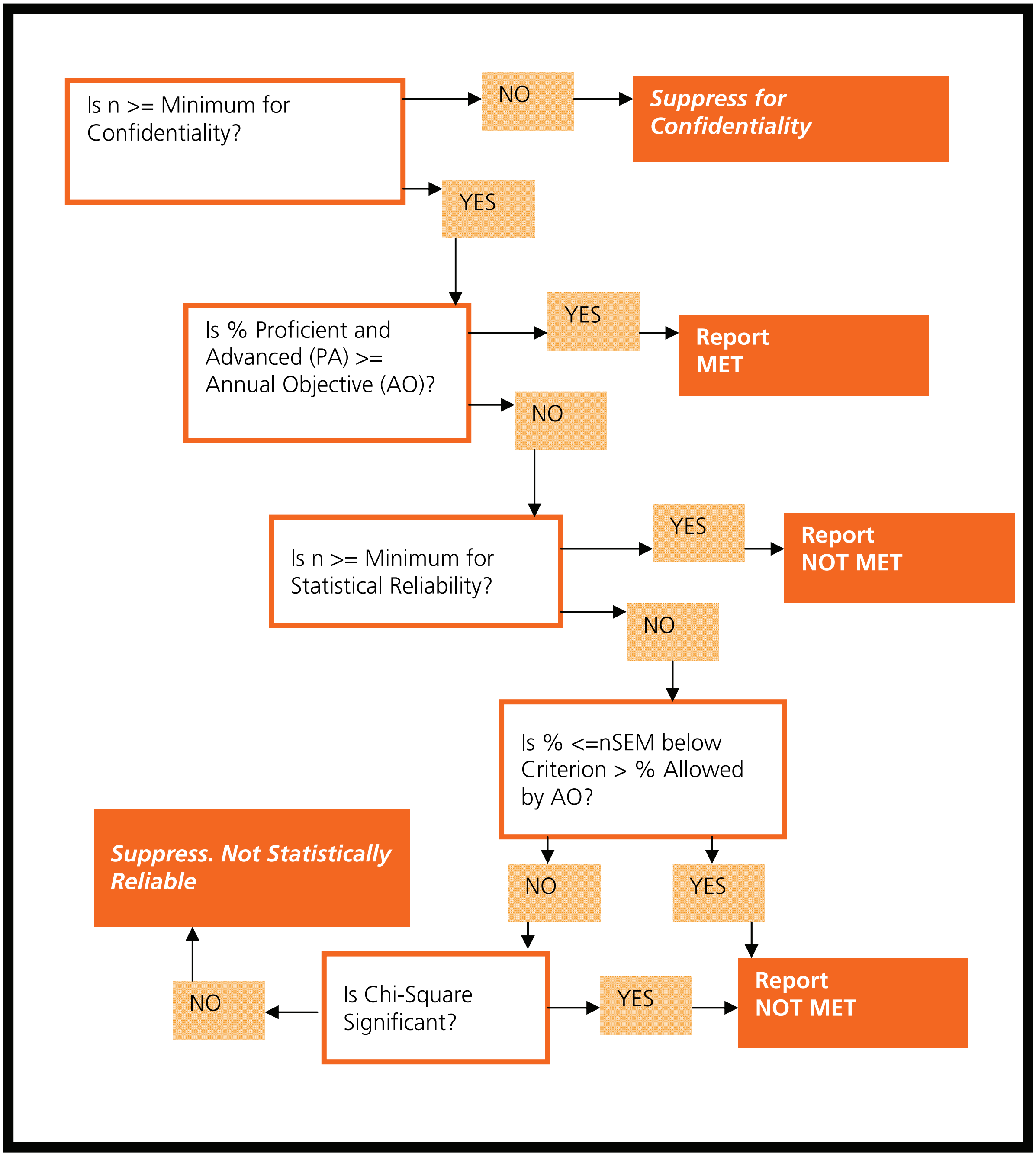
The Fail-Safe Test

Procedure: Accept the results for all subgroups that meet the annual objective regardless of size because the consequences of meeting or being declared statistically unreliable are the same.* Then apply the methodology in alternative 3 above only to those failing subgroups with sufficient numbers of students within *n*SEM of the criterion to place the status of the subgroup in doubt. For all subgroups that fail, determine if there are enough students failing by *n*SEM or less to place the results in doubt. If not, the school has failed with statistical reliability. If there are enough students in doubt, determine the probability that enough of those students below but within *n*SEM of the criterion may have scored above if retested.

**No Child Left Behind asks states to recognize schools that have performed well in meeting AYP and in closing the gap between subgroups. Alternative 4 may not be acceptable if a state does not want to reward and recognize schools that may have met annual objectives without an established degree of reliability. In which case, Alternative 3 is preferable.*

Figure 6 provides an overview of how alternative 4 might work. This process is complex, but a simpler chart can be developed that provides the single number for determining statistical reliability for each subgroup.

Figure 6: Flow Chart for Determining Statistical Reliability – Fail-Safe Alternative



Is n >= Minimum for Confidentiality?
NO
Suppress for Confidentiality
YES
Is % Proficient and Advanced (PA) >= Annual Objective (AO)?
YES
Report MET
NO
Is n >= Minimum for Statistical Reliability?
YES
Report NOT MET
NO
Is % <=nSEM below Criterion > % Allowed by AO?
Suppress. Not Statistically Reliable
NO
YES
NO
Is Chi-Square Significant?
YES
Report NOT MET

Figure 7: Statistically Unreliable Group

	Basic >nSEM below Criterion			Proficient & Advanced >nSEM above Criterion
Actual	**20**	**10**	**30**	**40**
True to Change Decision	**20**	**5**	**35**	**40**

Chi-Square = 2.05

P = .15

- **Accept Null H (Same)**
- **Group Could Have Passed**
- **Not Statistically Reliable**

Figure 7 provides an example of a subgroup that did not meet the annual objective, but has sufficient numbers of students within *n*SEM of the cut point who have a probability that they may pass if retested.

Figure 8: Statistically Reliable Group

	Basic >nSEM below Criterion			Proficient & Advanced >nSEM above Criterion
Actual	**25**	**15**	**20**	**40**
True to Change Decision	**25**	**0**	**35**	**40**

Chi-Square = 12.00

P = .00

- **Reject Null H (Same)**
- **Group Not Likely to Have Passed**
- **Statistically Reliable**

Figure 8 provides an example of a subgroup that did not meet the annual objective, and has insufficient numbers of students within *n*SEM of the cut point who have a probability that they may pass if retested.

Chi-square may be inappropriate for very small groups (fewer than 10), so Fisher's exact test (Hays, 1994) was also used to determine statistical reliability. The table below shows that the differences between the two are minimal, but when the call is close for a group, the two can differ. The annual objective is 75% in the examples in Table 15.

Table 15: Statistical Significance of Decision that Subgroups Failed to Meet an Annual Objective

Subgroup	Passed by More than 1 SEM	Passed, but Within 1 SEM	Additional Number Needed to Pass	Failed, but Within 1 SEM	Chi-Square Value Likelihood of NO Statistical Reliability		Fisher's Exact Test Likelihood of NO Statistical Reliability
A	70	4	1	6	.20	.65	.50
B	71	2	2	7	1.00	.32	.31
C	65	4	6	25	3.39	.06	.06
D	70	4	1	25	.13	.72	.50
E	70	0	5	0	N/A	N/A	N/A
F	40	25	10	25	2.38	.12	.08
G	73	0	2	15	2.14	.14	.24
H	59	10	6	20	2.44	.12	.09
I	40	10	25	30	31.75	0	0
J	40	20	15	15	12.00	0	0
K	40	30	5	10	2.05	.15	.12

In order to help determine if a decision whether a subgroup met an annual objective or not is statistically reliable, a simple Chi-square test can be computed to determine whether the observed distribution of scores in doubt is reliably worse than the distribution of scores that would allow the subgroup to pass.

In the first subgroup (A), 70 students passed the test by more than 1 SEM. The scores of those students are not in doubt. However, 4 students passed within 1 SEM and 6 students failed within 1 SEM, and 74 passing is not enough for the school to be rated passing. If just 1 of the 6 failing students had passed, the school would have been rated as passing. Given the fact that any or all of the 6 failing students might have a true ability above the cut score, how likely is it that, in fact, at least one of those 6 actually has the knowledge and skills to be at the proficient level? In statistics, a range of 0 to 1.0 is used to express probabilities. Any likelihood above .05 means that the situation is very likely. In this subgroup, a likelihood of .65 means that the two distributions (4 pass/6 fail, or 5 pass/5 fail) are very likely to be distributions representing the same real distribution for this school. There is a great deal of doubt about whether the school did, in fact, pass or fail the test. Therefore, in this subgroup, the judgment is not statistically reliable.

In the next to the last subgroup (J), 40 students passed the test by more than 1 SEM. However, 20 students passed within 1 SEM and 15 students failed within 1 SEM. In order for 75 of the students to pass the test, all 15 of the failing students would have to have passed. Statistically, a likelihood of .00 means that it is extremely unlikely that all 15 students actually have the knowledge and skills to be at the proficient level (because the probability is smaller, and therefore less likely, than .05). There is not much doubt about whether the school failed the test. In this situation, the judgment is statistically reliable.

As can be seen from the above table, in general, unless a substantial proportion of students needs to move from "failed" to "passed," the two distributions compared (i.e., the observed distribution and the next occurring distribution that changes the school's status) are statistically the same, and therefore the judgment of the school can be deemed "unreliable."

The test used in the above calculations is the Pearson Chi-square test for association. This tests whether the two distributions are from the same distribution. There are, however, assumptions made when using the Pearson Chi-square, which include:

1. Each and every observation is independent of each other observation. (Individual student scores are not dependent upon each other.)
2. Each observation qualifies for one and only one cell in the table. (Each student can either pass or fail.)
3. The sample size is large.

(Hays, 1994)

In a two-by-two table such as those used here, Hays recommends an expected frequency of 10 in each cell. In certain cases, the impact of fewer numbers in some cells should be investigated.

Can This be Simplified?

All these formulas and statistics easily become overwhelming. Even though the computers have no trouble calculating them, the processes become obscure and difficult to explain. The following look-up table was created to illustrate how to simplify the calculations described in alternative 4.

To use the table, determine how many additional students would have needed to score above the cut point for the subgroup to meet the annual objective. Add this number to the students scoring above but within *n*SEM of the cut point to get the MINIMUM. If this MINIMUM is equal to or higher than the number in the table below (Table 16), then the subgroup's result is statistically reliable—and the subgroup did not meet its annual objective.

Table 16: Look-Up Table for Statistical Reliability

One-Tailed Test at .05 Confidence Level		Number of Students Scoring ABOVE Criterion but Within *n*SEM										
		0	1	2	3	4	5	6	7	8	9	10
Number of Students Scoring ABOVE Criterion but Within *n*SEM	0											
	1											
	2	2										
	3	3	4	5	6							
	4	3	4	6	7	8	9	10	11	12	13	14
	5	3	5	6	7	8	10	11	12	13	14	15
	6	3	5	6	8	9	10	11	12	13	14	15
	7	4	5	7	8	9	10	11	12	13	14	16
	8	4	5	7	8	9	10	12	13	14	15	16
	9	4	6	7	8	9	11	12	13	14	15	16
	10	4	6	7	8	10	11	12	13	14	15	16

School Measurement Error

Sliding Scale

Procedure: Establish a sliding scale of minimums for total students in a subgroup based upon the variance of the percent proficient across all schools (or subgroups) in the state assessment and various group sizes. In other words, the larger the group, the closer the percent can be to the criterion and still be statistically reliable. Thus, a confidence interval is established for subgroups of each size. If the school's percent is above the annual objective by at least the size of the confidence level, then the school's status is statistically reliable. In the absence of a minimum subgroup size, this method can declare very large groups as unreliable if their percent proficient is very close to the criterion. This model was applied to Maryland's state test by Yen (1997). Maryland's state assessment included performance measures and item sampling strategies that may have not yielded usable individual student scores to which a standard error of measurement could be applied. Others have described methods for calculating the parameters used in this type of analysis even if they did not specifically relate their work to this application (Rogosa, 1999).

With this approach, a confidence interval is empirically calculated for each subgroup size (e.g., 2 percentage points for groups with 100 students, 8 percentage points for groups with 30 students, and 12 percentage points for groups with 10 students). These confidence intervals are based upon the distribution of all schools' (of the same size) percent of students above the criterion for proficiency. The mean and standard deviation of this distribution is used to calculate the probability that a school’s percent at or above proficiency could have been made if that school had been randomly drawn from the entire population of schools.

Model I (Schools fixed, forms random, pupils sampled from a finite population):

$SE^2(PAC) = \underline{\quad}_f/F+\underline{\quad}_{sf}/F+\underline{\quad}_w\underline{\ }R/n_T \cdot A+\underline{\quad}_w(1-R)/n_T$

$\underline{\quad}_f=\{MS(f)-A\_MS(w)\}/S\_n*$

$\underline{\quad}_{sf}=\{MS(sf)\_A\_MS(w)\}/n*$

$\underline{\quad}_w=MS(w)$

Where

F=number of forms
n*=number of pupils per form per school in the ANOVA
n_T=the typical total number of pupils tested per school
S=number of schools in ANOVA
A=state average proportion of eligible pupils who did not provide scores
R=mean (across forms) of the coefficient alpha values for that grade and content area

Model II (Schools fixed, forms random, pupils sampled from an infinite population):

$SE^2(PAC) = \underline{\quad}_f/F+\underline{\quad}_{sf}/F+\underline{\quad}_w\ /n_T$

$\underline{\quad}_f=\{MS(f)\_MS(w)\}/S\_n*$

$\underline{\quad}_{sf}=\{MS(sf)-MS(w)\}/n*$

Other terms remain unchanged from Model I.

Model III (Schools random, forms random, pupils sampled from an infinite population):

$\underline{\quad}_f=\{MS(f)-MS(sf)\}/S\cdot n*$

Other terms remain unchanged from Model II.

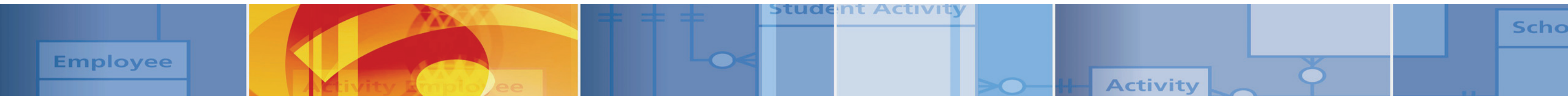
(Yen, 1997, page 13)

References

More complete references and documents related to No Child Left Behind can be found at: www.espsolutionsgroup.com

- *2000 Disaggregated Achievement Report Guide Sheet.* (2000). Florida Department of Education, Curriculum, Instruction & Assessment, Evaluation & Reporting.
- Alker, H. R., Jr. (1975). Polimetrics: Its Descriptive Foundations. In F. Greenstein, and Polsby, N. (Ed.), *Handbook of Political Science. Reading: Addison-Wesley.*
- American Institutes for Research (2002). National School-Level State Assessment Score Database. (Compact Disc).
- Chou, F. (2002). Louisiana Department of Education, personal communication.
- Clements, B. S. (1998). *Protecting the Confidentiality of Education Records in State Databases*: Evaluation Software Publishing, Inc., Study for the Massachusetts Department of Education.
- Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement,* 57, 373-399.
- Evans, T., Zayatz, L., & Slanta, J. (1996). *Using Noise for Disclosure Limitation of Establishment Tabular Data:* U. S. Bureau of the Census.
- Fienberg, S. E., Steele, R. J., & Makov, U. (1997). *Statistical Notions of Data Disclosure Avoidance and Their Relationship to Traditional Statistical Methodology: Data Swapping and Loglinear Models:* Carnegie Mellon University, Haifa University.
- Gravetter, F. J., & Wallnau, L. B. (2000). *Statistics for the Behavioral Sciences* (Fifth ed.). Belmont, CA: Wadsworth/Thomson Learning.
- Gurr, T. R. (1972). *Politimetrics: An Introduction to Quantitative Macropolitics.* Englewood Cliffs, NJ: Prentice-Hall.
- Hambleton, R. K., & Slater, S. C. (1997). Reliability of Credentialing Examinations and the Impact of Scoring Models and Standard-Setting Policies. *Applied Measurement in Education*, 10(1), 19-38.
- Hays, W. L. (1994). *Statistics*. Austin: Harcourt Brace College Publishers.
- Hilton, G. (1976). *Intermediate Politometrics.* New York: Columbia University Press.
- Hoffman, R. G., & Wise, L. L. (2000). *School Classification Accuracy Final Analysis Plan for the Commonwealth Accountability and Testing System.* Alexandria, VA: Human Resources Research Organization (HumRRO).
- Jaeger, R. M., & Tucker, C. G. (1998). *Analyzing, Disaggregating, Reporting, and Interpreting Students' Achievement Test Results: A Guide to Practice for Title I and Beyond.* Washington, D.C.: Council of Chief State School Officers.
- Kane, M. (1996). The Precision of Measurements. *Applied Measurement in Education,* 9(4), 355-379.

- Kim, J. J., & Winkler, W. E. (1997). *Masking Microdata Files*. Paper presented at the American Statistical Association.
- King, G. (1991). *On Political Methodology.* Paper presented at the American Political Science Association, Atlanta, Georgia.
- Ligon, G. D. (1998). *Small Cells and Their Cons (Confidentiality Issues):* NCES Summer Data Conference.
- Ligon, G. D., Clements, B. S., & Paredes, V. (2000). *Why a Small n is Surrounded by Confidentiality: Ensuring Confidentiality and Reliability in Microdatabases and Summary Tables*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans.
- Ligon, G.D., Jennings, Judy, & Clements, B.S. (2002). *Confidentiality, Reliability, and Calculation Alternatives for No Child Left Behind.* Unpublished paper for CCSSO CAS/SCASS.
- Linn, R. L., & Haug, C. (2002). Stability of School-Building Accountability Scores and Gains. *Educational Evaluation and Policy Analysis*, 24(1), 29-36.
- Moore, R. A. (1997). *Controlled Data-Swapping Techniques for Masking Public Use Microdata Sets*. Unpublished manuscript, Washington, D C.
- *North Carolina Student Accountability Standards*. North Carolina Department of Public Instruction. Retrieved, from the World Wide Web: www.dpi.state.nc.us/student_promotion/sas_guide/standard_error.html
- Opperdoes, F. (1997). *Bootstrapping*. Retrieved, from the World Wide Web: http://www.icp.ucl.ac.be/~opperd/private/bootstrap.html
- Rai, K. B., & Blydenburth, J. C. (1973). *Political Science Statistics*. Boston: Holbrook Press.
- Rogosa, D. (1999). *Statistical Properties of Proportion at or above Cut-off (PAC) Constructed from Instruments with Continuous Scoring.* UCLA: National Center for Research on Evaluation, Standards, and Student Testing.
- *Statistical Policy Working Paper 2--Report on Statistical Disclosure and Disclosure-Avoidance Techniques*. (1978). Federal Committee on Statistical Methodology.
- Winkler, W. E. (1997). *Views on the Production and Use of Confidential Microdata:* Bureau of the Census.
- Yen, W. M. (1997). The Technical Quality of Performance Assessments: Standard Errors of Percents of Pupils Reaching Standards. *Educational Measurement*: Issues and Practices (Fall), 5-15.
- Zayatz, L., Moore, R., & Evans, B. T. (undated). *New Directions in Disclosure Limitation at the Census Bureau*. Washington, DC: Bureau of the Census.



About ESP Solutions Group

ESP Solutions Group provides its clients with *Extraordinary Insight*™ into PK-12 education data systems and psychometrics. Our team is comprised of industry experts who pioneered the concept of "data driven decision making" and now help optimize the management of our clients' state and local education agencies.

ESP personnel have advised school districts, all 52 state education agencies, and the U.S. Department of Education on the practice of K-12 school data management. We are regarded as leading experts in understanding the data and technology implications of the **No Child Left Behind Act (NCLB)**, **Education Data Exchange Network (EDEN)**, and the **Schools Interoperability Framework (SIF).**

Dozens of education agencies have hired ESP to design and build their student record collection systems, federal reporting systems, student identifier systems, data dictionaries, evaluation/assessment programs, and data management/analysis systems.

To learn how ESP can give your agency *Extraordinary Insight* into your PK-12 education data, email info@espsg.com.

This document is part of *The Optimal Reference Guide* Series, designed to help education data decision makers analyze, manage, and share data in the 21st Century.

ESP Solutions Group

(512) 879-5300
www.espsolutionsgroup.com