

Best Practices for Determining Subgroup Size in Accountability Systems While Protecting Personally Identifiable Student Information

INSTITUTE OF EDUCATION SCIENCES CONGRESSIONALLY MANDATED REPORT



Best Practices for Determining Subgroup Size in Accountability Systems While Protecting Personally Identifiable Student Information

INSTITUTE OF EDUCATION SCIENCES

U.S. Department of Education

John B. King, Jr.
Secretary

Institute of Education Sciences

Ruth Neild
Deputy Director for Policy and Research Delegated Duties of the Director

National Center for Education Statistics

Peggy G. Carr
Acting Commissioner

The Institute of Education Sciences (IES) is the statistics, research, and evaluation arm of the U.S. Department of Education. We are independent and non-partisan. Our mission is to provide scientific evidence on which to ground education practice and policy and to share this information in formats that are useful and accessible to educators, parents, policymakers, researchers, and the public.

If you have any comments or suggestions about this or any other IES product or report, we would like to hear from you. Please direct your comments to

IES, U.S. Department of Education
Potomac Center Plaza
550 12th Street SW
Washington, DC 20202

January 2017

The IES home page address is <http://ies.ed.gov>.

The IES Publications and Products page address is <http://ies.ed.gov/pubsearch>.

This publication is only available online. To download, view, and print the report as a PDF file, go to the IES Publications and Products address shown above.

This report was prepared with assistance from the Quality Information Partners under Contract No. ED-CFO-16-A-0126. Mention of trade names, commercial products, or organizations does not imply endorsement by the U.S. Government.

Suggested Citation

Seastrom, Marilyn (2017). Best Practices for Determining Subgroup Size in Accountability Systems While Protecting Personally Identifiable Student Information. (IES 2017-147). U.S. Department of Education, Institute of Education Sciences. Washington, DC. Retrieved [date] from <http://ies.ed.gov/pubsearch>.

Technical Contact

Marilyn Seastrom
National Center for Education Statistics (NCES)
Chief Statistician
(202) 245-7766
Marilyn.Seastrom@ed.gov

Every Student Succeeds Act of 2015

SECTION 9209. REPORT ON SUBGROUP SAMPLE SIZE.

- (a) *REPORT.*— . . . the Director of the Institute of Education Sciences shall publish a report on—
- (1) best practices for determining valid, reliable, and statistically significant minimum numbers of students for each of the subgroups of students, as defined in section 1111(c)(2) of the Elementary and Secondary Education Act of 1965 (20 U.S.C. 6311(c)(2)), as amended by this Act, for the purposes of inclusion as subgroups of students in an accountability system described in section 1111(c) of such Act (20 U.S.C. 6311(c)), as amended by this Act; and
 - (2) how such minimum number that is determined will not reveal personally identifiable information about students
- (b) *PUBLIC DISSEMINATION.*—The Director of the Institute of Education Sciences shall work with the Department of Education’s technical assistance providers and dissemination networks to ensure that such report is widely disseminated—
- (1) to the public, State educational agencies, local educational agencies, and schools; and
 - (2) through electronic transfer and other means, such as posting the report on the website of the Institute of Education Sciences or in another relevant place.
- (c) *PROHIBITION AGAINST RECOMMENDATION.*—In carrying out this section, the Director of the Institute of Education Sciences shall not recommend any specific minimum number of students for each of the subgroups of students, as defined in section 1111(c)(2) of the Elementary and Secondary Education Act of 1965 (20 U.S.C. 6311(c)(2)), as amended by this Act.

Acknowledgements

I would like to thank Thomas Szuba for his editorial contributions. Any errors are my responsibility.

Executive Summary

The Every Student Succeeds Act (ESSA) of 2015 (Public Law 114-95) requires each state to create a plan for its statewide accountability system. In particular, ESSA calls for state plans that include strategies for reporting education outcomes by grade for all students and for economically disadvantaged students, students from major racial and ethnic groups, students with disabilities, and English learners. In their plans, states must specify a single value for the minimum number of students needed to provide statistically sound data for all students and for each subgroup, while protecting personally identifiable information (PII) of individual students. This value is often referred to as the “minimum n-size.”

Choosing a minimum n-size is complex and involves important and difficult trade-offs. For example, the selection of smaller minimum n-sizes will ensure that more students’ outcomes are included in a state’s accountability system, but smaller n-sizes can also increase the likelihood of the inadvertent disclosure of PII. Similarly, smaller minimum n-sizes enable more complete data to be reported, but they may also affect the reliability and statistical validity of the data.

To inform this complex decision, Congress required the Institute of Education Sciences (IES) of the U.S. Department of Education to produce and widely disseminate a report on “best practices for determining valid, reliable, and statistically significant minimum numbers of students for each of the subgroups of students” (Every Student Succeeds Act of 2015 (ESSA 2015), Public Law 114-95). Congress also directed that the report describe how such a minimum number “will not reveal personally identifiable information about students.” ESSA prohibits IES from recommending any specific minimum number of students in a subgroup (Section 9209).

IES produced this report to assist states as they develop accountability systems that

- (1) comply with ESSA;
- (2) incorporate sound statistical practices and protections; and
- (3) meet the information needs of state accountability reporting, while still protecting the privacy of individual students.

State education agencies (SEAs) will likely rely on the expertise of statistical and research professionals to make critical choices about the design, development, operation, and use of state accountability systems. When tasked with establishing a minimum n-size, these statistical experts can look to this report to identify several fundamental questions to consider while evaluating the statistical, data, and privacy implications of specific minimum n-size decisions. To that end, this report presents a thorough review of the statistical and privacy considerations that are most relevant to state efforts to establish a minimum n-size.

As presented in this report, the minimum n-size refers to the lowest statistically defensible subgroup size that can be reported in a state accountability system. Before getting started, it is important to understand that the minimum n-size a state establishes and the privacy protections it implements will directly determine how much data will be publicly reported in the system.

Chapter 1 includes a listing of “Key Steps in Establishing a Minimum Number of Students and Protecting Personally Identifiable Student Data in a State Accountability System” for analysts and policymakers to consider. Each of the steps is explored in the subsequent chapters.

First, states should establish a team of policymakers who are supported by technical staff with sufficient statistical and data expertise to lead the effort to establish a minimum n-size for their state accountability system.

Second, states should verify that the results will be statistically valid. A result is valid if it accurately measures what it is intended to measure; if the result can be generalized to other places, people, and times; and if the statistical conclusions drawn from the result are reasonable (i.e., credible or believable). The analysts and the policymakers should ask themselves whether the results are reasonable and believable and whether the results observed in a specific school, district, or state will support comparisons over time, between subgroups within schools, districts, or states, or between schools, districts, or states.

Third, states should confirm that the results will be statistically reliable. A result is reliable if it is consistent, stable, and reproducible from one use to the next, and relatively error free. The analysts and the policymakers should evaluate the amount and nature of errors in the results for each subgroup in a population that is intended to be reported separately.

Fourth, states should ensure that the results will be statistically sound. This step involves reaching a decision as to whether the results in a state accountability system will be treated as coming from a population or a sample. The arguments for and against both perspectives, and the implications for accountability systems that flow from adopting one approach or the other, are presented. Simply put, the population perspective draws on decades of experience using descriptive statistics to study population trends and patterns that are observed in universe or census data for all members of a defined population. The sampling perspective sees state accountability systems as focused on a *school's* performance (i.e., the school's ability to serve its students) on a set of outcome measures, as opposed to the performance of a *particular set of students at one point in time*; therefore, each year's set of students is viewed as a sample from a larger population of similarly defined groups of students over time. Measuring meaningful or significant differences and the analytic trade-offs under each of these perspectives is discussed, with the caution that before a state settles on an approach for its accountability system, the state's analysts and policymakers should use data to examine the impact on the validity, reliability, and credibility or statistical soundness (i.e., statistical conclusion validity) of the results when different scenarios for assumptions are explored.

Fifth, the statistical rigor that informed the selection of the minimum n-size should be documented and how this minimum number is statistically sound should be described. This step is important to the requirement for a full justification of the decisions made in establishing a state's accountability system, and for communicating and collaborating with various stakeholders.

Sixth, the state should identify recommended privacy controls to be used to ensure that the state accountability system does not inadvertently disclose personally identifiable information. The third chapter of the report starts by acknowledging that a minimum of 301 students would be needed (with results reported as integers) to avoid reporting results that risk disclosing information about one or two students. However, since data in a state's accountability system will most likely be reported for smaller subgroups of students, additional privacy controls known as disclosure avoidance techniques are presented. The techniques presented include primary and complementary suppression, ranges, top and bottom coding, and rounding.

The seventh step requires confirming that the minimum n-size, in combination with the techniques selected in step six, is sufficient to not reveal any personally identifiable information.

Finally, in the eighth step, the state should describe how it collaborated with teachers, principals, other school leaders, parents, and other stakeholders when determining the minimum number.

This report also includes several features intended to make the information more accessible to a broader audience, although the nature of the topic requires that the primary audience of this *Report* possess an understanding of basic statistics. Important points are highlighted in text boxes on the right-hand side of a page and technical notes are displayed in full-width boxes. While short examples are included throughout the text, extended examples are offset in the main body. These extended examples illustrate how statistical advisors might effectively evaluate different alternatives prior to establishing any specific minimum number of students for subgroups in their accountability systems.

Table of Contents

Every Student Succeeds Act of 2015.....	ii
Acknowledgements	iii
Executive Summary	iv
Chapter 1. Introduction	1
Why minimum n-size matters	1
Background.....	1
About this report.....	2
Chapter 2. Best Practices for Establishing a Valid, Reliable, and Statistically Sound Minimum Number of Students for State Accountability Systems.....	5
Will the results be valid?	5
Will the results be reliable?	7
Will the estimates be statistically sound?.....	8
Adopting a population perspective or a sampling perspective.....	8
Comparing population and sampling perspectives and the selection of meaningful or significant differences	19
Further implications of selecting a population or sampling perspective.....	22
Chapter 3. Protecting Personally Identifiable Information	24
Data Protection Techniques to Minimize Inadvertent Disclosures in Public Reporting	25
Primary and Complementary Cell Suppression.....	25
The Use of Ranges and Top and Bottom Coding to Replace Specific Data Values.....	29
Rounding.....	32
Caveat: The Hierarchical Structure of Education Data.....	33
Additional Resources.....	34
Appendix A: Sampling.....	35
Appendix B: ED DRB Data Protection Schema	37
References	38

Chapter 1. Introduction

The Every Student Succeeds Act (ESSA) of 2015 (Public Law 114-95) requires that each state education agency create a plan for its accountability system. State plans must include strategies for reporting education outcomes by grade for all students and for economically disadvantaged students, students from major racial and ethnic groups, students with disabilities, and English learners. In their plans, states must specify a single value for the minimum number of students needed to provide statistically sound data, for all students and for each subgroup, while protecting personally identifiable information of individual students. This value is often referred to as the state's *minimum n-size*.

Choosing a minimum n-size is a complex decision that involves important and difficult statistical and policy trade-offs. To inform states' decisions, this guide presents best practices for determining a valid, reliable, and statistically sound minimum n-size for the student groups that are publicly reported in a state accountability system. The guide also describes practical privacy protections that can help states ensure that the minimum n-sizes they establish and the data they report will not reveal personally identifiable information about students.

This report responds to ESSA's requirement that the U.S. Department of Education's Institute of Education Sciences (IES) produce and widely disseminate a report on best practices for establishing a minimum n-size. The best practices described in this report are derived from a thorough review of the statistical and privacy considerations that are most relevant to state efforts. Readers will note that this report does not recommend any specific minimum n-size values. ESSA specifically proscribes IES from making any such recommendations in this report.

Why minimum n-size matters

From a technical perspective, minimum n-size refers to the lowest statistically defensible subgroup size that can be reported with protections for personally identifiable information in a state accountability system. In addition to being the law (ESSA 2015, Public Law 114-95), there are many reasons for states to apply the robust statistical procedures and privacy protections identified in this report when planning their state accountability systems. This is because the minimum n-size a state establishes and the privacy protections it implements will directly determine how much data will be publicly reported in the system. A state's statistical experts can use the methods and recommendations presented in this report to strike what they conclude to be a responsible and appropriate balance between the trade-offs associated with data release and data protection.

What Is a Minimum N-Size?

In the context of this report, **minimum n-size** refers to the lowest statistically defensible subgroup size that can be reported in a state accountability system with protections for personally identifiable information.

Background

How best to establish a minimum n-size for accountability reporting has been a topic of policy discussions for more than a decade. Some policymakers and researchers are concerned that setting the minimum subgroup size too high results in the exclusion of data for a substantial number of subgroups.

Others are concerned that setting the minimum subgroup size too low can result in misinterpretation, as when a change for a small number of students results in a substantial change in the percentage of students in one outcome category versus another. Further, a too-low n-size can produce a large margin of error around the estimate for an outcome measure that could limit the utility of the outcome measure. Finally, a too low n-size can reveal personally identifiable information about individual students in small subgroups.

Discussions about the most appropriate minimum n-size in accountability systems intensified with the 2001 reauthorization of the Elementary and Secondary Education Act (Public Law PL 107-110), known as No Child Left Behind (NCLB). NCLB required the reporting of achievement data and other results for elementary and secondary public school students, including separate reports for economically disadvantaged students, students in major racial and ethnic groups, students with disabilities, and English learners. Unless the number of students in a group was insufficient to yield statistically reliable information, or the results would reveal personally identifiable information about a student, NCLB required states to include data for all students and for each specified subgroup. The U.S. Department of Education reiterated these requirements in the 2002 regulations that were issued to support NCLB (Title I – Improving the Academic Achievement of the Disadvantaged; Final Regulations on Title I, July 5, 2002). The Department also reinforced these requirements in non-regulatory guidance (Report Cards-Title I Part A – Non-Regulatory Guidance, September 12, 2003).

In October of 2008, the Department issued a Final Rule with amended regulations requiring states to modify their Consolidated State Application Accountability Workbook to include a description of how they determined the minimum number of students sufficient to yield statistically reliable information for each purpose for which disaggregated data were used. Further, the Department proposed that states be required to ensure, to the maximum extent practicable, that all student subgroups were included, particularly at the school level, for purposes of making accountability decisions (34 CFR Part 200, Title I– Improving the Academic Achievement of the Disadvantaged; Final Rule; Sec. 200.7 Disaggregation of data, Federal Register, 2008, Volume 73, Number 210). With the reauthorization of the ESEA as the Every Student Succeeds Act of 2015, the requirement for states to determine and explain their minimum subgroup size became federal law.¹

About this report

This report is intended primarily for the technical personnel who support education policymakers as they make critical choices about the design, development, operation, and use of state accountability systems. The content of this report reflects the statistical and data complexities inherent to establishing a minimum n-size and ensuring effective data privacy provisions in state accountability systems. The report assumes that the reader has an understanding of basic statistics, and many common statistical terms are used without definition.

However, recognizing that policymakers may also want to consult this report, we have sought to make concepts more accessible to these audiences by highlighting important points and placing technical

¹ Proposed regulations to support ESSA were under consideration during the development of this report, but they had not yet been finalized and, therefore, are not discussed here.

notes in text boxes. We provide extended examples in boxes offset from the main text. These extended examples illustrate how statistical advisors might evaluate different alternatives prior to establishing any specific minimum n-size for their accountability system.

The steps that must be addressed by each state are summarized in Box 1 and provide the organization for Chapter 2 and Chapter 3.

BOX 1: Key Steps for Establishing a Minimum Number of Students and Protecting Personally Identifiable Student Data in a State Accountability System

Establishing a minimum n-size has significant implications for data quality and privacy protections in a state accountability system. State leaders may want to undertake the following steps to ensure that the process of establishing a minimum n-size adheres to ESSA requirements and supports the overarching goals of the state accountability system., as described in greater detail throughout this report.

- Step 1. Establish a team with sufficient statistical, and data expertise to lead the effort to establish a minimum n-size for your state accountability system.
- Step 2. Verify that the resulting estimates will be statistically valid.
 - Evaluate external validity and statistical conclusion validity, which are both relevant to establishing a minimum n-size.
- Step 3. Confirm that the resulting estimates will be statistically reliable.
 - Evaluate the reliability of outcome measures for each subgroup in a population that is intended to be reported separately.
- Step 4. Ensure that the resulting estimates will be statistically sound.
 - Determine whether the outcome measures will be treated as a population or sample.
 - Establish criteria for triggering a “meaningful difference” (i.e., the smallest change in value in a reporting group that constitutes a significant difference).
- Step 5. Document the statistical rigor that informed the selection of the minimum n-size and describe how this minimum number is statistically sound.
- Step 6. Identify recommended privacy controls to be used (such as primary and complementary suppression, ranges, top and bottom coding, and rounding) to ensure that the state accountability system does not inadvertently disclose personally identifiable information.
- Step 7. Confirm that the specified minimum number, in combination with the privacy controls selected in step 6, is sufficient to not reveal any personally identifiable information.
- Step 8. Describe how the state collaborated with teachers, principals, other school leaders, parents, and other stakeholders when determining the minimum number.

The remainder of the report is presented in the following chapters and appendices A and B.

Chapter 2. Best Practices for Establishing a Valid, Reliable, and Statistically Sound Minimum Number of Students for State Accountability Systems provides an in-depth discussion of the statistical concepts and methods necessary to consider when determining a statistically defensible minimum n-size for a state accountability system.

Chapter 3. Protecting Personally Identifiable Information reviews best practices for minimizing the likelihood of inadvertently disclosing individual student information in state accountability reporting.

Appendix A. Sampling provides a brief discussion of statistical concepts underlying sampling for non-technical readers.

Appendix B. Data Protection Schema from the U.S. Department of Education’s Disclosure Review Board (ED DRB) provides additional information related to the privacy protections and rationale established by the ED DRB.

References lists the citations for all sources used to substantiate the recommendations offered in the report.

Chapter 2. Best Practices for Establishing a Valid, Reliable, and Statistically Sound Minimum Number of Students for State Accountability Systems

Assembling a team with appropriate expertise is the first step that a state should take to establish a valid, reliable, and statistically sound minimum n-size. The team should include the policymakers responsible for decisions about the state accountability system, and, given the technical nature of the analysis needed to guide the decision, the team should also include staff members with sufficient statistical and data expertise to conduct the analysis. ESSA requires that the team consult with teachers, principals, other school leaders, parents, and other stakeholders when considering which minimum n-size to choose (ESSA 2015, Public Law 114-95, Section 1111, (c)(3)(a)(ii)).

The state team must address several crucial questions in selecting a specific value for a minimum n-size in accountability system reporting. These questions are:

- Will the results be valid?
- Will the results be reliable?
- Will the results be statistically sound?
 - Should the results be treated as population parameters from a universe data collection?
 - Should the results be treated as estimates from a sample survey data collection?
- What steps need to be taken to establish a minimum size for reporting subgroups?
- Will personally identifiable student information be protected by the minimum subgroup size (Chapter 3)?
 - What additional steps can be taken to protect personally identifiable student information?

Will the results be valid?

In common language, validity refers to the state of being logically or factually sound. From a statistical perspective, validity is the degree of correspondence between a measurement and the process or product being studied (National Forum on Education Statistics, 2005). A measurement instrument or test is valid if it accurately measures what it is intended to measure (Vogt, 2005). Validity refers to both the procedures and the conclusions, with a focus on methodological soundness or appropriateness (Graziano and Raulin, 1989). These concepts apply to each of the measures in a state's accountability system, from measures of school quality or student success to graduation rates and academic achievement.

Validity

Validity is the degree of correspondence between a measurement and the process or product being studied.

A measure is valid if it accurately measures what it is intended to measure.

Although there are many types of validity, *external validity* and *statistical conclusion validity* are particularly relevant to establishing a minimum n-size. External validity refers to the extent to which the results of a study can be generalized beyond the immediate study to other settings (ecological validity), other people (population validity), and over time (historical validity) (McLeod, 2007). For example, ecological validity permits student accountability results to be compared across students in the same subgroup between schools; population validity enables comparisons to be drawn in one school between

students in different subgroups at one point in time; and historical validity allows results to be compared within one subgroup at different points in time.

Statistical conclusion validity is concerned with whether the statistical conclusions drawn from the results of a study are reasonable (i.e., credible or believable) (Graziano and Raulin, 1989; Trochim, 2006). For example, in 2014, 35 percent (6 of 17) of the students who were English learners at ABC Elementary School scored at the proficient level or above on the state assessment, but when the results for 2015 came in, only 24 percent (4 of 17) of the English learners scored at the proficient or above level. The state elected to treat their accountability data as universe data and identified values over 10 percent as significant changes. The school principal was upset when he saw this 11 percentage point drop. He looked into the underlying data and realized that this significant decline was the result of a change in the performance of only two students—and he questioned whether it was valid for the performance of two students to trigger a significant difference in the accountability system.

Therefore, the analysts and the policymakers should ask themselves whether the results observed are reasonable and believable and whether the results observed in a specific school, district, or state will support comparisons over time, between subgroups within schools, districts, or states, or across schools, districts, or states. Although the analysts' and policymakers' decisions concerning what is reasonable and believable are subjective and involve judgement, these decisions should be clearly explained so that they are available for public scrutiny.

Technical Note 1: More on Statistical Validity

Statistical conclusion validity is concerned with whether the statistical conclusions drawn from the results of a study are reasonable. Consider the fact that in reaching a conclusion about whether there is a relationship or a difference between two variables, there are three possible outcomes:

- 1) The researcher's conclusion may be correct.
- 2) The researcher may conclude that a relationship or difference exists when, in fact, it does not exist (Type I error or false positive).
- 3) The researcher may conclude that there is no relationship or difference when in fact it does exist (but the researcher's data and analysis failed to detect it) (Type II error or false negative).

Incorrect conclusions may result from not collecting enough data to detect the relationship or difference (i.e., an insufficient population or sample size); from violations of the assumptions that underlie statistical tests for the basis of a comparison; or from the use of measures that are unreliable.

Any measure consists of some unmeasurable true value plus some amount of measurement error that can reduce the validity of the measure (Trochim, 2006; Biemer, et al., 1991; Viswanathan, 2005). In any measurement, there can be two types of errors: random and systematic. Validity can be reduced by the presence of either random or systematic measurement errors. Random errors arise from factors that randomly affect measurement of the variable across the sample. If these errors are truly random they add variability but do not affect the mean of a measure. Systematic errors arise from factors that affect measurement across the sample in a uniform (i.e., systematic) manner. Because each source of systematic error tends to be directional (i.e., either positive or negative), they are more likely to result in bias in measurement.

Will the results be reliable?

Reliability refers to the degree of consistency, stability, or reproducibility of a measure, test, or observation from one use to the next (Vogt, 2005; Boudett, City, and Murnane, 2008; Levy and Lemeshow, 2008). In short, reliability involves the quality of measurement, and any measure consists of some unmeasurable true value plus some amount of measurement error (Trochim, 2006; Biemer, et al., 1991; Viswanathan, 2005). Thus, the reliability of a measure can be jeopardized by errors that arise from factors that affect the measurement of the variable across the sample. If these errors are truly random, they add variability but do not affect the mean.

Reliability

Reliability is the degree of consistency, stability, or reproducibility of a measure, test, or observation from one use to the next.

Reliability involves the quality of measurement.

Therefore, the analysts and the policymakers should evaluate the amount and nature of errors in the results. Considering first errors that are systematic (i.e., nonrandom); when this occurs, the result is likely to be biased. When bias exists and is constant over time or across settings, the result is consistent, stable, and reproducible and thus reliable. However, the result is not valid due to the lack of accuracy resulting from the bias. However, if the factors contributing to the bias change over time or across different settings, the results are not likely to be consistent, stable or reproducible. As a result, the quality required for reliability and the accuracy required for validity are not achieved, and the results in question are neither valid nor reliable.

Considering next errors that are truly random, in this case, the value for the estimate will not be affected by the error. If the amount of random error is relatively small in each of two results that are being compared, the results are of sufficient quality to be considered reliable and sufficient accuracy to be considered valid. However, if the amount of random error is large, substantial amounts of variability across the results from individual students may make it difficult to reach any statistical conclusions based on a comparison of results across time or across different settings. When this happens, the results are neither valid nor reliable. In this case, if the analyst and policymaker conclude that there is no difference between two results, they are at risk of making a Type II error (i.e., identifying a false negative).

Given the need to report results separately for subgroups in a state accountability system, analysts should measure reliability separately for each subgroup within the set of results for each outcome measure. This is because results based on a small number of cases are less reliable (i.e., a small error in the counts for a small subgroup has a larger relative effect than would be the case with the same size error in the count for a larger subgroup). Thus, results from state level data may be more reliable than the results at the district and school levels. In sum, reliability must be studied by analysts and evaluated by policymakers for the set of results for each outcome measure for each subgroup at the school, district, and state levels.

Technical Note 2: More on Measurement Error

To maximize the reliability of the measures in a state's accountability system, a technical expert should pilot test the assessment or survey data collection instrument, seek information from respondents on the difficulty of the questions and any impacts of the interview environment (e.g.,

cognitive interviews), train test administrators and interviewers to maintain consistency and avoid the introduction of test administrator or interviewer error, monitor data processing controls, use statistical procedures to evaluate the amount of error in the resulting data, and use multiple related measures that do not share the same systematic errors to triangulate across multiple measures (Marczyk, DeMatteo, and Festinger, 2005; Trochim, 2006).

Although there are several sources of error that are of potential concern in a state accountability system, steps can be taken to guard against them. For example, the ESSA requirement for a 95 percent participation rate addresses concerns over error from nonresponse. Undercoverage due to a failure to identify all units to be measured is another potential problem. In the case of a state accountability system, error from undercoverage exists only if there are public schools that are not in the state reporting system (e.g., newly formed charter schools). In any data collection, the possibility exists for measurement error from coding, and processing quality checks (e.g., range edits, relative change from prior years) should be used to minimize this source of error. Similarly, unclear data collection instruments and instructions can also contribute to measurement error. Therefore, with the introduction of new metrics in a state's accountability system, special attention should be paid to the clarity of the data collection instruments and instructions.

Will the estimates be statistically sound?

To meet the ESSA requirement for establishing a statistically sound statewide value for the minimum number of students, policymakers must address three key, interrelated questions:

- Will our accountability system adopt a population perspective or a sampling perspective?
- What will our state consider to be a meaningful or significant difference in the results in our accountability system?
- How many students must have a change in status for our state's accountability system to recognize the change as a meaningful or significant difference?

Adopting a population perspective or a sampling perspective

State accountability systems are designed to include data for all students, or put differently, for an entire *population* or *universe* of students. For this reason, whether to treat the data for these students as a *population* or as a *sample from a population* may seem like a straightforward decision. However, that is not the case. There are arguments for and against both perspectives, and there are implications for accountability systems that flow from adopting one approach or the other.

Arguments for the population perspective

The United States has several longstanding ongoing census or universe data collections that are regularly used by demographers, economists, urban planners, and epidemiologists, among others to describe and understand patterns and trends in American society (see Technical Note 3). These population data are taken at face value and allow analysts and researchers to describe the data to facilitate policymakers' and the general public's understanding of the patterns and trends in the size and demographic, social, and economic characteristics of the population. A population perspective assumes that there is no sampling error and the data are more accurate than sample data (Myers, 1992). Arguably, population data are less subject to misinterpretation because they are not subject to

the limitations imposed on the interpretation of the data when errors attributed to sampling come into play (Petersen, 1969).

From a practical perspective, within a state accountability system a population perspective permits simpler summaries of outcomes and comparison of differences between groups and over time than are possible with a sampling perspective. In a state accountability system, the population of interest can be defined at the subgroup level within a school, district, or state and subgroup comparisons can be drawn within or across schools, districts, or states. An advantage of this simplicity is the greater ease with which a public audience can understand and engage with the accountability system. Descriptive statistics can be computed directly from the observed cases, and subgroup performance can be summarized using one or more measures of central tendency, such as the mean, median, or mode. The amount of variation can be captured by examining the range of data, the mean deviation, the variance, or the standard deviation. Alternatively, the data may be arrayed into a frequency distribution, with counts of the number of population members with each value in the distribution. Such counts can then be used with the size of the population to compute a relative distribution showing the proportion of the population within each value in the distribution.

Technical Note 3: Examples of Census or Universe Data

The United States has conducted a census of the population every ten years since 1790 (Petersen, 1969). The resulting data have been used to provide data on the basic demographics of the population (age, sex, race, marital status), on household characteristics (family size, relationship of household members to the head of household), and on social and economic characteristics of individuals (occupation, employment status, school enrollment status, educational attainment, veteran status, retirement status).

United States vital statistics data on all births and deaths occurring in a year have been collected since 1933 (Linder and Grove, 1947). These data allow researchers and policymakers to monitor population growth by basic demographic characteristics and patterns of death by age and cause.

The federal government started collecting biennial counts of the number of public schools and institutions of higher education in 1870; by the mid-1940s there were annual data on higher education enrollment and the number of earned degrees; and by the early- to mid-1950s there were annual data on the number of kindergarten through grade 12 public schools, students, and teachers. Education universe data are currently published through the Common Core of Data (CCD) and the Civil Rights Data Collection (CRDC) at the elementary and secondary levels and through the Integrated Postsecondary Education Data System (IPEDS) at the postsecondary level. These data are used to monitor participation in education, the completion of various education credentials and degrees, the size of the educational labor force, and educational expenditures.

Arguments for the sampling perspective

When a sample is drawn from a population and data are collected from a subset of the population in lieu of a census, inferential statistics are used to infer information about the population based on the sample (Blair and Blair, 2015; Lodico, Spaulding, and Voegtle, 2006; Barnett, 2002). Although data for an accountability system are typically collected from all students, the focus is on the overall level of performance in an identified subgroup over some period of time. In this instance, the results for a particular group of students at one point in time are viewed as a sample from the universe of similarly defined groups over time. In an accountability system, the outcome measures are used to gauge a *school's* performance (i.e., school effectiveness), as opposed to the results of a *particular set of students*

(Hill and DePascale, 2003). Although data for an accountability system are typically collected from all students, the focus is on a *school's* performance on the results for a set of outcome measures, as opposed to the results of a *particular set of students*. For this reason, in a 2014 report on the uses of assessments for education accountability, the Joint Committee on Educational and Psychological Testing argued that it is appropriate to treat annual measures of student performance as a sample rather than a population – even if data from all students are used.² The logic is that each year of results for a specific outcome measure can be regarded as data for a sample from a larger population.³

This position is consistent with a 1997 article in which Cronbach et al. discussed using standard errors to provide a measure of uncertainty in scores to avoid over interpretation of results when generalizing assessment findings (Cronbach, Linn, Brennan, and Haertel, 1997). Similarly, a 2002 Council of Chief State School Officers report discussed using standard errors and associated confidence intervals “to infer how well the observed score represented the “true” percent proficient for that school given a sample of all possible students who could attend that school” (Marion, White, Carlson, Erpenbach, Rabinowitz, and Sheinker, 2002, page 66).

Following this line of reasoning, a sampling approach to interpreting universe data can be applied to each of the measures in a state’s accountability system.⁴ Once a sampling perspective has been adopted, the information from the sample can be used to infer the characteristics of the entire population. These inferences can be used to estimate or predict the value of a population characteristic, such as when an estimate is expressed as a value plus or minus a specified margin of error. Inferential statistics can be used to test hypotheses about the value of a measure relative to a fixed value, a different measure of the same characteristic within a sample, or the same measure from a different sample.⁵ Inferential statistics require the calculation of a mean, the variance around the mean, the standard deviation, and the standard error.⁶

² This Joint Committee was formed by three professional associations with a strong professional interest in educational and psychological testing—the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME)—to revise the existing 1999 *Standards*.

The Joint Committee on Educational and Psychological Testing’s formal operational standard states:

When average test scores for groups are the focus of the proposed interpretation of the test results, the groups tested should generally be regarded as a sample from a larger population, even if all examinees available at the time of measurement are tested. In such cases, the standard error of the mean should be reported, because it reflects variability due to sampling examinees as well as variability due to individual measurement error (Standard 2.17).

³ The Joint Committee further posited that considering data over time, each year’s results reflect the experience of a sample of students in a longitudinal sense (i.e., given static conditions, comparable groups from the same population will recur over time).

⁴ The Joint Committee extended this sample interpretation to other results, such as the proportion of Hispanic students who are proficient on an assessment or the proportion of Black students who graduate from high school on time.

⁵ In addition to sampling error, estimates inferred from samples are subject to measurement error and to sampling bias that may occur as a result of coverage bias, selection bias, or nonresponse bias.

⁶ The mean is the average of the individual measures of the members of the sample, computed as the sum of the individual estimates divided by the number of estimates. The variance is a measure of the dispersion or spread of the members of the sample, computed by subtracting each observed estimate from the mean, squaring each difference, summing over the differences, and dividing by the sample size minus 1. The standard deviation is a measure of the average amount of the

Importantly, when taking a sampling perspective, observed differences between two groups or between points in time may not be *statistically significant*, or statistically distinguishable from zero. States may need to take special care in explaining to the public what statistical significance does and does not mean. For non-technical readers of this report, a brief discussion of statistical concepts underlying sampling is provided in Appendix A.

The population perspective: Identifying a meaningful difference in outcomes

Recall from the discussion about a population perspective that it is assumed that there is no sampling error and the data are more accurate (i.e., both reliable and valid) than sample data (Myers, 1992). If a state chooses a population perspective for its accountability system, it does not need to take sampling error into account to assess whether differences between groups or across time are statistically significant (i.e., significantly different from zero). However, if a state adopts a population perspective, the state will need to define how large a difference between two values must be to qualify as *meaningful*.

If the accountability measures will be reported as the percentage of students with a specific outcome, differences over time or between subgroups can be described in terms of percentage point differences. A state's predefined *meaningful difference* expressed in percentage points can be used to

- compare differences between subgroups (e.g., the percentage of English learners who are at or above proficient versus the percentage of students fluent in English who are at or above proficient);
- compare differences within a subgroup over time (e.g., the percentage of students with disabilities at or above proficient in year one compared to year two); or
- compare differences within a subgroup across schools (e.g., the percentage of Hispanic students who are at or above proficient in one school compared to the percent of Hispanic students who are at or above proficient in other schools in the same district or state).

What is a "Meaningful" Difference?

Each state must define how large a difference must be to qualify as meaningful.

Researchers caution against misinterpreting small differences in student outcome measures, especially in the case of differences between groups or differences within groups over time. For example, a small change in the scores of 2 out of 20 students in a class can produce a relatively large percentage change in what is reported to the public.

The potential impact of such a change in a small number of students should be considered when states determine the size of a "meaningful" difference in accountability reporting.

Determining a *meaningful difference* is, in part, a policy decision. Is the state willing to consider a difference of three percentage points to be meaningful? Or does it want a higher bar of five points, or ten points? Accountability experts have argued that state policymakers should be cautious about misinterpreting small differences in the results from student outcome measures, especially in the case

dispersion or spread of the members of the sample, calculated as the square root of the variance. The standard error refers to error in the estimates due to random fluctuations in different samples; it measures the average of the amount of difference (i.e., standard deviation) expected from different samples of the same size; the standard error of the mean is calculated by dividing the estimate of the standard deviation of sample means by the square root of the sample size.

of subgroup differences or changes over time (Linn, Baker, and Herman, 2002). Boudet, City, and Murnane (2008) advocate giving greater weight to differences that are sizable or that persist over time and recommend a more cautious “wait and see” approach to small differences and one-time differences.

Establishing a meaningful difference also requires an awareness of the sizes of all student subgroups for which data are to be reported. Caution should be used when interpreting student results from small subgroups, because a change in the status of a few group members (for example, 2 or 3 students out of 20 or 25 students in a class or a population subgroup) can produce a relatively large change in the group’s percentage (Joint Committee on Standards for Educational and Psychological Testing, 2014).

Therefore, when an accountability system takes a population perspective, the impact of a change in a small number of students should be considered in establishing the size of a difference or change that is considered meaningful and in determining the minimum number of students needed to provide valid, reliable, statistically sound data for all students and each subgroup. The goal is to strike a balance between the number of students required to trigger a meaningful difference and the size of the smallest population that will yield such a meaningful difference.

Table 1 provides the data needed to determine how small a population can be while still ensuring that a change in the status of two, three, four, or five students in the group does not produce a meaningful change at the level that the state has specified. Using Table 1, the minimum size of the population required to yield a specified meaningful difference for a given small number of students can be determined by following these steps:

1. Identify the percentage change required for difference to be considered meaningful (e.g., more than 5 percent or more than 10 percent).
2. Select the small number of students that should not trigger a meaningful difference (i.e., 2, 3, 4, or 5).
3. Read down the selected Small Number of Students column until you see the percentage value for the defined meaningful difference identified in step 1.
4. Read across the row to the left to identify the necessary minimum population size to support these choices.

Table 1. Percent of population represented by a change of 2, 3, 4, or 5 students at different population sizes.

Population Size	Small Number of Students				Population Size	Small Number of Students			
	2	3	4	5		2	3	4	5
5	40	60	80	100	60	3	5	7	8
6	33	50	67	83	61	3	5	7	8
7	29	43	57	71	62	3	5	6	8
8	25	38	50	63	63	3	5	6	8
9	22	33	44	56	64	3	5	6	8
10	20	30	40	50	65	3	5	6	8
11	18	27	36	45	66	3	5	6	8
12	17	25	33	42	67	3	4	6	7
13	15	23	31	38	68	3	4	6	7
14	14	21	29	36	69	3	4	6	7
15	13	20	27	33	70	3	4	6	7
16	13	19	25	31	71	3	4	6	7
17	12	18	24	29	72	3	4	6	7
18	11	17	22	28	73	3	4	5	7
19	11	16	21	26	74	3	4	5	7
20	10	15	20	25	75	3	4	5	7
21	10	14	19	24	76	3	4	5	7
22	9	14	18	23	77	3	4	5	6
23	9	13	17	22	78	3	4	5	6
24	8	13	17	21	79	3	4	5	6
25	8	12	16	20	80	3	4	5	6
26	8	12	15	19	81	2	4	5	6
27	7	11	15	19	82	2	4	5	6
28	7	11	14	18	83	2	4	5	6
29	7	10	14	17	84	2	4	5	6
30	7	10	13	17	85	2	4	5	6
31	6	10	13	16	86	2	3	5	6
32	6	9	13	16	87	2	3	5	6
33	6	9	12	15	88	2	3	5	6
34	6	9	12	15	89	2	3	4	6
35	6	9	11	14	90	2	3	4	6
36	6	8	11	14	91	2	3	4	5
37	5	8	11	14	92	2	3	4	5
38	5	8	11	13	93	2	3	4	5
39	5	8	10	13	94	2	3	4	5
40	5	8	10	13	95	2	3	4	5
41	5	7	10	12	96	2	3	4	5
42	5	7	10	12	97	2	3	4	5
43	5	7	9	12	98	2	3	4	5
44	5	7	9	11	99	2	3	4	5
45	4	7	9	11	100	2	3	4	5
46	4	7	9	11	101	2	3	4	5
47	4	6	9	11	102	2	3	4	5
48	4	6	8	10	103	2	3	4	5
49	4	6	8	10	104	2	3	4	5
50	4	6	8	10	105	2	3	4	5
51	4	6	8	10	106	2	3	4	5
52	4	6	8	10	107	2	3	4	5
53	4	6	8	9	108	2	3	4	5
54	4	6	7	9	109	2	3	4	5
55	4	5	7	9	110	2	3	4	5
56	4	5	7	9	111	2	3	4	5
57	4	5	7	9	112	2	3	4	4
58	3	5	7	9	113	2	3	4	4
59	3	5	7	8	114	2	3	4	4
					115	2	3	3	4

Example 1: Determining population size for a specified meaningful change

Assume a state's policymakers decide to follow the recommendation to focus on differences that are sizable (Boudet, City, and Murnane, 2008). As a result, the policymakers decide to treat differences as meaningful only if they are greater than 10 percentage points (Step 1). The policymakers also want to avoid a scenario in which a change of one or two students in any category of an outcome measure would trigger a meaningful difference (Step 2).

To examine this scenario, the state's analyst reads down the *Small Number of Students* column heading for "2" in Table 1 until the last entry of 11 percent appears (with light grey highlighting) (Step 3). Reading across the row to the population size column on the left, the analyst sees that a change of 11 percentage points would occur if 2 students moved from one category of an outcome measure to another in a population or subgroup of interest of 19 (i.e., $[(2/19)*100 = 11 \text{ percent}]$). Since the policymakers do not want a change in the outcome of two students to result in a meaningful difference, the analyst looks to the next row in the table to see that 2 out of 20 students would produce a 10 percentage point change (i.e., $[(2/20)*100 = 10 \text{ percent}]$) (Step 4). Thus, as long as there are at least 20 students in the population or subgroup of interest, a change in the status of 2 students would not produce a meaningful difference that crosses the policymaker's threshold of greater than 10 percentage points.

After seeing the results from the first scenario, the state's policymakers decide to consider what would happen if a more conservative approach were applied, such as increasing from 2 to 3 the minimum number of students that could produce a meaningful change. To examine this new scenario, the state's analyst looks down the *Small Number of Students* column heading for "3" to identify the last value of 11 percent (light grey highlighting), reads across the row to the population size column on the left, and sees that a change of 11 percentage points would occur if 3 students moved from one category of an outcome measure to another in a population or subgroup of interest of 28 students (i.e., $[(3/28)*100 = 11 \text{ percent}]$). Since the policymakers do not want a change in the outcome of 3 students to result in a meaningful difference, the analyst looks to the next row in the table to see that 3 out of 29 students would produce a 10 percentage point change (i.e., $[(3/29)*100 = 10 \text{ percent}]$). Thus, as long as there are at least 29 students in the population or subgroup of interest, a change in the status of 3 students would not produce a meaningful difference that crosses the policymakers' threshold of greater than 10 percentage points.

Some of the policymakers' colleagues in an adjacent state suggest that more than a 10 percentage point requirement for a meaningful difference is too stringent, so the state's policymakers decide to examine a third scenario that reduces the meaningful difference value to more than 5 percentage points. The data in the *Small Number of Students* column heading for "2" show that 37 students (dark grey highlighting) are needed for the population or subgroup of interest (i.e., $2/37$

= 5 percent). For the value of “3” in the Small Number of Students column heading, the data show that at least 55 students (dark grey highlighting) are required (i.e., $3/55 = 5$ percent).

Note that the shadings in columns 4 and 5 show the minimum number of students required if the number of students required to result in a meaningful difference of greater than 5 percent (dark grey highlighting) or greater than 10 percent (light grey highlighting) is raised to 4 or 5

The sampling perspective: Identifying a significant difference in outcomes

Taking a sampling perspective requires that the state consider how well its accountability system will be able to detect statistically significant differences for subgroups at the school, district, and state levels. The sampling perspective assumes that the population parameter is not known and must be estimated using established methods for statistical inference.

Established facts about statistical sampling provide information about margins of error. Put simply, at a given level of confidence, subgroups with smaller numbers of students will have larger margins of error. State policymakers must use these statistical facts to answer questions such as:

- *Does this margin of error meet our criteria for statistical conclusion validity?*
- *Does it provide useful information about student progress?*
- *Do estimates meet the reproducibility, consistency, and stability criteria of reliability?*

To illustrate how subgroup size relates to margin of error, Example 2 provides three scenarios using samples of different sizes.

Example 2: Determining the statistical conclusion validity of estimates at the school, district, and state levels.

Consider ABC Elementary School, which has 30 students with disabilities in the third grade. Because the focus of accountability data is on the proportion (percentage) of students with a particular characteristic, such as students with disabilities who are at or above the proficient level, the standard error formula for proportions rather than means is relevant for analytical purposes.

Standard error of the proportion $[se(p)] = \text{square root } [(p(1-p))/n]$

If the level of confidence is set at the customary 95 percent and the proportion proficient, p , is 0.50, the standard error of the proportion, $se(p)$, is calculated by multiplying the proportion by 1 minus the proportion, and then dividing that product by the sample size, n , and taking the square root of the resulting number.

The state’s analyst applies this formula to data for ABC Elementary School and conducts the analysis for the 30 third-graders with disabilities, the resulting calculation is $p(1-p)$

equals $0.5 * 0.5 = 0.25$, which is divided by the sample size of 30, resulting in a value of 0.008333. The square root of this value yields a standard error for p of 0.091287 (or 9.1287 percent). The margin of error is determined by multiplying the 1.96 associated with the 95 percent confidence interval⁷ by this standard error, 9.1287, resulting in a margin of error of 17.89. The confidence interval is then computed as the estimate plus or minus the margin of error.

Thus, when taking a sampling perspective, if 50 percent of the 30 disabled third-grade students scored at or above the proficient level, with a margin of error of 17.89 percentage points, the analyst can be 95 percent confident that the population percentage (the true value) falls between 32.11 and 67.89. This wide confidence interval indicates, with 95% confidence, that the true population value for the percentage of third-grade students with disabilities scoring at or above the proficient level in ABC elementary is somewhere in this 36 percentage point range.

The state's analyst shares his findings with the policymakers who are charged with setting the minimum number of students for a subgroup that will be used in their state's accountability system. Together, the analyst and the policymakers must decide whether this margin of error meets their criteria for statistical conclusion validity. That is, they must ask themselves and their constituencies whether treating all values between 32.11 percent and 67.89 percent as not significantly different from 50 percent will provide useful information about student progress. Similarly, they must consider whether, for their purposes, estimates that range across 35.78 percentage points from one use to the next meet the reproducibility, consistency, and stability criteria of reliability.

In contrast, consider instead that ABC Elementary School is one of six elementary schools in their school district. Across the six schools, the district has a total of 190 third-graders with disabilities. The state's analyst conducts the analysis for the school district using the assumptions applied at the school level (i.e., the level of confidence is set at 95 percent and the proportion at or above the proficient level, p , is 0.50), the square root of the result of 0.25 divided by the sample size of 190 yields a standard error for p of 0.036474 (or 3.6474 expressed in percentage points) and the resulting confidence interval is 1.96 times this standard error, resulting in a value of 7.11. Thus, if 50 percent of the 190 disabled third-grade students scored at or above the proficient level, with a margin of error of plus or minus 7.11 percentage points, the analyst can be 95 percent confident that the percent of the disabled third-graders scoring at or above the proficient level in the district (the true value) falls between 42.89 percent and 57.11 percent. This is a narrower confidence interval, indicating with 95% confidence that the true population value is somewhere in this 14 percentage point range. The analyst and the policymakers must answer the same set of questions. Does this margin of error meet their criteria for statistical conclusion validity? Is it reasonable to treat all values between 42.89 percent and 57.11 percent as not significantly different from 50 percent? Do estimates that range across 14 percentage points from one use to the next meet the reproducibility, consistency, and stability criteria of reliability?

⁷ 1.96 is the exact value associated with two standard errors from the mean.

At the state level, there are 20 districts with a total of 3,500 third-graders with disabilities. The state’s analyst conducts the state-level analysis for this subgroup of students, using the assumptions already applied at the school and district levels (i.e. the level of confidence is set at 95 percent and the proportion at or above the proficient level, p , is 0.50), the square root of the result of 0.25 divided by the sample size of 3,500 yields a standard error for p of 0.0084515 (or 0.8452 expressed in percentage points) and the resulting margin of error is 1.96 times this standard error, resulting in a value of 1.66. Thus, if 50 percent of the 3,500 disabled third-grade students in the state scored at or above the proficient level, with a confidence interval of plus or minus 1.66 percentage points, the analyst can be 95 percent confident that the percent of the disabled third-graders in the state who scored at or above the proficient level (the true value) falls between 48.34 percent and 51.66 percent. This is a narrow confidence interval, indicating that the true population value is somewhere in this 3 percentage point range. The analyst and the policymakers must ask themselves the same set of questions. Does this margin of error meet their criteria for statistical conclusion validity? Is it reasonable to treat all values between 48.34 percent and 51.66 percent as not significantly different from 50 percent? Do estimates that range across 3 percentage points from one use to the next meet the reproducibility, consistency, and stability criteria of reliability?

When evaluating expectations for data in a state accountability system that uses the sampling perspective, states should recognize that the assumption of a proportion of 0.5 will yield a larger standard error than would be the case with larger or smaller proportions. For example, with a sample size of 30, the proportions of 0.6, 0.7, 0.8, and 0.9 would result in standard errors of 17.5, 16.4, 14.3, and 10.7 percent, respectively.⁸ When comparable data are available for a prior year’s data, the actual percentage of students reaching the desired outcome should be used to study possible effects on standard errors if a sampling perspective is applied rather than the 0.5 assumption used in Example 2.⁹

Example 2 uses a 95 percent confidence level. With a 95 percent confidence interval, the analysts are accepting a 5 percent possibility of concluding that a difference exists when it does not (i.e., Type I error, or accepting a false positive).¹⁰ While 95 percent is the level typically used for statistical testing, state

⁸ Proportions of 0.4, 0.3, 0.2, and 0.1 would also result in standard errors of 17.5, 16.4, 14.3, and 10.7 percent, respectively.

⁹ Although the standard error of the proportion formula $[se(p)] = [\text{square root } [(p(1-p))/n]]$ is applied in the examples, there are a number of free online tools that can be used to calculate standard errors and evaluate sample sizes for proportions. An Internet search for the term “sample size calculators for proportions” will generate links to a range of tools. Note that these tools typically ask for the size of the population from which the sample is drawn—given the rationale for treating these results as sample data, the default for a large population should be used. Since some of the online tools use proportions and others use percentages, users should pay close attention to whether a tool is using the decimal or percentage version of a proportion.

¹⁰ At the 95 percent level, an analyst understands that 95 percent of the values in the distribution of sample values will fall in the range calculated from the confidence interval. Expressed in other terms, there is a 5 percent probability of concluding that a relationship or difference exists when, in fact, it does not.

analysts and policymakers may want to consider whether a 90 percent confidence level (or some lower level) is sufficient for the intended use.¹¹

In making a decision to alter the confidence level from 95 percent to 90 percent, or some lower level, the analyst must understand that the probability of concluding that there is no difference when one does exist decreases (i.e., Type II error) (Ferguson and Takane, 1989). The probability that a statistical test will correctly identify a real relationship or difference is known as statistical power (see technical note 4).¹² Information regarding the power of a statistical analysis and the probabilities of making Type I and Type II errors can be helpful to an analyst wanting to make a recommendation on the minimum n-size and the corresponding confidence interval or significance level to use when testing whether a relationship or difference exists within state accountability system data.

Technical Note 4: Type I and II Errors and Power

Recall from technical note 1 that a researcher may conclude that a relationship or difference exists when, in fact, it does not exist (Type I error or false positive); or that there is no relationship or difference when in fact there is (but the researcher's data and analysis failed to detect it) (Type II error or false negative). Statistical power is the probability that a statistical test will correctly identify a real relationship or difference. It is calculated as the inverse of the probability of concluding that there is no difference when there is a difference (Type II error).

Conventional wisdom, absent a justification for how to handle a specific comparison, calls for the use of a power level of 80 percent and a statistical significance level of 5 percent (corresponding to a 95 percent confidence interval) (Cohen, 1988; Ellis, 2010). Note that a test based on a large sample has more statistical power, and is less likely to produce a Type II error, than the same test based on a smaller sample. In fact, if a sample is too small, the risk of overlooking meaningful effects is increased. Such a test is said to be underpowered. Furthermore, if the relationship or size of the difference that the analyst is measuring is small, a small sample will generate more variability, making it more difficult to detect a relationship or difference (Ellis, 2010).

To see the impact of changing the significance level, consider the example where an analyst wants to know if a sample of 30 students has enough power to determine whether an observed proportion of 0.5 is significantly different from a reference value of 0.3 with the probability of making a Type I error set at 5 percent (i.e., 95 percent significance level). The analyst uses a power calculator for a 1-sample 2-sided equality, and learns that the chance of making a Type II error is 40.76 percent. Expressed in terms of power, this translates into a 59.24 percent chance that a statistical test will correctly identify a real relationship or difference. In this case, the analyst must advise the state's policymakers as to whether correctly describing the relationship or difference 59 percent of the time is acceptable.

Unsatisfied with the low power of the test, the analyst might decide to explore the likely ramifications of

¹¹ NCES requires significance testing with sample surveys using a 95 percent level. In contrast, the Census Bureau uses a 90 percent level.

¹² With a sample size of 30, the convention of applying a 5 percent significance level and an 80 percent power level results in a 20 percent chance of making a Type II error, thus implying that a Type I error is 4 times more important than a Type II error (i.e., $20/5 = 4$). Holding the sample size at 30, if the significance level is increased to 10 percent, the power level is 70 percent, resulting in a 30 percent chance of making a Type II error. In this case, there is an underlying assumption that a Type I error is 3 times more important than a Type II error (i.e., $30/10 = 3$).

increasing the risk of making a Type I error to 10 percent. In this case, the analyst considers whether the trade-off from increasing the chance of identifying false positives to 10 percent is worth it to increase the power of the analysis. With a minimum n-size of 30, this change in the significance level from 5 percent to 10 percent translates into an increase in the power of the analysis from 59.24 percent to 70.80 percent (i.e., the Type II error rate is 29.20 percent). For the analyst advising the policymakers who need to decide on a minimum n-size for a state accountability system, the percentage of the schools that are incorrectly identified as making progress or slipping backwards increases from 5 percent to 10 percent—in other words, there is a chance that 1 in every 10 schools identified with a change did not really change. At the same time, the increase in statistical power increases the chance of correctly identifying a difference from just under 6 out of 10 schools to just over 7 out of 10 schools.

Comparing population and sampling perspectives and the selection of meaningful or significant differences

Population data support a smaller minimum number of students for signifying a meaningful difference than would be possible from a sampling interpretation of the data for a significant difference of the same size. For example, if a state decided that more than 10 percentage points is its standard for a meaningful or significant difference:

- From a *population perspective*, a subgroup would need to show a percentage point difference of more than 10 to be identified with a meaningful difference from one time point to the next or between two subgroups at one point in time. In Example 1, a change of 3 students results in a difference greater than 10 percentage points when there are 28 students in a subgroup.¹³ Thus, for a subgroup change or difference of more than 10 percentage points that is not triggered by 3 or fewer students, **the minimum n-size is 29.**
- From a *sampling perspective*, the size of a significant difference decreases as the sample size increases. As a result, there will be instances with small subgroup sizes where the only differences that can be identified as statistically significant are larger than 10 percentage points and instances with large subgroup sizes where differences of less than 10 percentage points are identified as statistically significant. However if a state were to choose a sampling perspective and set its standard for a meaningful or significant difference at 10 percentage points, a sampled subgroup would need to have a margin of error of 10 percentage points, for a difference of more than 10 percentage points from one time point to the next or between two subgroups at one point in time to yield a statistically significant difference.¹⁴ Importantly, when using a sampling perspective, the percentage of the subgroup that is in the category of interest (e.g., the percent of Hispanic fourth-graders who performed at or above proficient on a state assessment) must be taken into account since the standard error and, hence, the margin of

¹³ In subgroups with 29, 30, or 31 students, a change or difference of 3 students would be 10 percentage points, a value below the established meaningful or significant difference.

¹⁴ These conclusions could be drawn by comparing the confidence intervals on the two estimates to ensure that they are not overlapping or by conducting a test for statistical significance that accounts for joint standard error.

error, is tied to that percentage. Specifically, the standard error decreases as the percentage increases.¹⁵ As noted in Example 2, if there are 30 students in a subgroup, an estimate of 50 percent would have a margin of error of plus or minus 17.9 percentage points, with estimates of 60, 70, 80, and 90 percent yielding margins of error of 17.5, 16.4, 14.3, and 10.7 percent, respectively. With a sampling approach, a sample size of 96 is needed to reach the point where the margin of error of 10 percent occurs for a subgroup with 50 percent of the students in one category. Thus, when using a sampling approach to test for a difference of 10 percentage points with 50 percent of the students in a subgroup in one category, **the minimum n-size is 96.**

Example 3 illustrates the minimum n-sizes required by the population and sampling perspectives for a school, district, and state.

Example 3: Comparing the results at the school, district, and state levels from a population versus sampling perspective

Using the school, district, and state from Example 2, the team responsible for setting the minimum n-size for the state must decide between a population perspective and a sampling perspective. From a population perspective, the minimum number of students required for defining a change or difference is tied to the percentage established by the state team as a meaningful difference. From a sampling perspective, the minimum number of students required for defining a change or difference is tied to the margin of error derived from the distributional statistics of the sample. The team is considering a minimum n-size of 30 students for their state's accountability system. The team identifies the subgroup of third-grade students with disabilities as one subgroup that has school-level data with the minimum n-size of 30, so they work through some scenarios to evaluate the impact of the two perspectives on this subgroup of students.

Population: *First, the team decides to explore the ramifications of treating their data as a population. They set the size of a meaningful difference at more than 10 percent of the subgroup and agree that they do not want a change or a difference of only 3 students to produce a meaningful difference. They see that this year, 50 percent, or 15, of the 30 third-graders with disabilities enrolled in ABC Elementary School performed at or above the proficient level. They confirm that a change in the performance of 3 students next year (i.e., 10 percent of 30) would not exceed their pre-established value of more than 10 percentage points for a meaningful difference; as a result they determine that a change in the status of 4 students is needed to produce a meaningful change of more than 10 percentage points (because a change in the status of 4 out of 30 is a change of 13.3 percent). To show a meaningful increase in the next year, at least **19 (15+4) or 63.3 percent of the school's 30 third-graders with disabilities will need to perform at or above the proficient level.***

¹⁵ With a subgroup size of 30, all of the percentage values below 90 percent would have a margin of error greater than 10 percentage points, and each margin of error would be based on a change or difference of 4 or 5 students.

At the school district level, the district has a total of 190 third-graders with disabilities. A meaningful difference is more than 10 percent of the subgroup so, at the district level, 20 students would need to change their status regardless of the percentage of third-graders who are proficient or above in the first year (i.e., 10 percent of 190 is 19, so 20 students need to change to produce a meaningful difference). For the sake of comparison, the team assumes the percentage of third-graders with disabilities in the district at or above the proficient level this year is also 50 percent, or 95 students. To show a meaningful increase in the next year, at least **115 (i.e., 95+20) or 60.5 percent of the district's 190 third-graders with disabilities would need to perform at or above the proficient level.**

When the data across the 20 districts in the state are rolled up to the state level, there are 3,500 third-graders with disabilities. Since 10 percent of this group is 350, at least 351 students in this subgroup would have to move into the proficient level (or above) to register a meaningful difference at the state level. Again, staying with an estimate of 50 percent at or above proficient this year (1,750 students), to show a meaningful increase in the next year, at least **2,101 (i.e., 1750+351) or 60.02 percent of the state's 3,500 third-graders with disabilities would need to perform at or above the proficient level.**

Sample: The team also evaluates the implications of treating their data as a sample. From a sampling perspective, the size of a significant difference is driven by the size of the standard error, and thus the margin of error—with the size of the margin of error decreasing as the sample size increases. A school with 50 percent of the 30 third-grade students with disabilities scoring at or above the proficient level in one year has a margin of error of 17.9 for the 50 percent estimate. As a result, it is not possible to detect differences that are less than 18 percent. Since 17.9 percent of 30 is 5.4, in order to show a significant increase in the next year, the school would need to have at least 6 more third-graders with disabilities score at or above the proficient level for a total of **21 of the 30 third-graders with disabilities scoring at or above the proficient level.** Thus to demonstrate a significant improvement at the school level, 70 percent of the third-graders with disabilities would have to score at or above proficient. Thus, since the percentage of students needed to show a significant improvement is greater than the 10 percentage points set with the population approach, the district can only exhibit a significant change with a larger percentage of third-graders with disabilities scoring at or above proficient than is the case with a population perspective (i.e., 70 percent sample perspective versus 63.3 percent population perspective).

The six schools with third-graders with disabilities account for a total of 190 third-graders with disabilities in the district. For the sake of comparison, assume that 50 percent, or 95, of the 190 third-graders with disabilities in the school district also scored at or above the proficient level this year. The margin of error on this larger number of students is 7.1 percentage points. Since 7.1 percent of 190 is 13.5, in order to show a significant increase in the next year, the district would need to have at least 14 more third-graders with disabilities score at or above the proficient level, for a total of **109 third-graders (95+14) with disabilities performing at the proficient level.** Thus to demonstrate a significant improvement at the district level, 57 percent of the third-graders with disabilities would have to score at or above proficient. Thus, since the percentage of students needed to show a significant improvement is less than the 10 percentage points set with the population approach, the district can exhibit a significant change with a smaller percentage of third-graders with disabilities scoring at or above

proficient than is the case with a population perspective (i.e., 57 percent sample perspective versus 61 percent population perspective).

*Again, for the sake of comparison, assume that 50 percent, or 1,750, of the 3,500 third graders with disabilities in the state also scored at or above the proficient level this year; the margin of error on this larger number of students is 1.7 percentage points. Since 1.7 percent of 3,500 is 59.5 or 60 students, in order to show a significant increase in the next year, the district would need to have **at least 1811 (1750+61) third-graders with disabilities score at or above the proficient level.** Thus to demonstrate a significant improvement at the state level, 52 percent of the third-graders with disabilities would have to score at or above proficient. Thus, since the percentage of students needed to show a significant improvement is less than the 10 percentage points set with the population approach, the district can exhibit a significant change with a smaller percentage of third-graders with disabilities scoring at or above proficient than is the case with a population perspective (i.e., 52 percent sample perspective versus 60 percent population perspective).*

Further implications of selecting a population or sampling perspective

A comparison of the results in example 3 demonstrates a clear trade-off between the population and sampling perspectives. If the margin of error for the number of students in a subgroup is larger than the pre-established meaningful or significant percentage point difference, that subgroup will demonstrate progress using a population perspective but not using a sampling approach. Thus, when there are small subgroups at the school level, more subgroups will demonstrate progress using a population perspective than would be the case using a sampling approach. But, when the margin of error values associated with larger numbers of students in subgroups at the district level are less than the pre-established meaningful or significant percentage point difference, more subgroups at the district level will demonstrate progress using a sampling perspective than using a population approach. Similarly, for subgroups at the state level, when the margin of error values are less than the pre-established percentage point difference, more subgroups at the state level will demonstrate progress using a sampling perspective than using a population approach.

It is important to note that because the margin of error at the state level will be smaller than the margin of error at the district level, subgroup differences at the state level are more likely to reach statistical significance than subgroup differences at the district level. Similarly, the sampling perspective may be perceived as giving an advantage to larger districts and larger states, since the smaller margin of errors associated with larger numbers of students are more likely to result in statistically significant differences. One solution to level the playing field is to combine the concept of a meaningful difference with a sampling perspective, defining differences as those differences that are statistically significant and meet the pre-established percentage point difference.

Example 3 explores only some of the options. Before a state settles on an approach for its accountability system, the state’s analysts and policymakers should use their existing data to examine the possible impact on the validity, reliability, and credibility or statistical soundness (i.e., statistical conclusion validity) of the results when different scenarios for assumptions are explored. For example, from a population perspective states should compare a minimum difference of greater than 10 percentage points versus a minimum difference of greater than 5 percentage points (or any other percentage point value that the state wants to consider). A state’s analysts and policymakers should also explore different values required to produce a change. Example 3 uses the assumption that 3 students should not be able to support a meaningful change; states may want to examine the impact of using 2, 4, or 5 in place of 3. Note that each of these scenarios will point to different minimum n-sizes. In a similar vein, states may want to use existing data from the state to analyze the impact of changing from a 95 percent significance level to a 90 percent confidence level, or of moving the minimum n-size below or above 30 students—paying particular attention to the validity, reliability, and credibility or statistical soundness (i.e., statistical conclusion validity) of the results at the school, district, and state levels.

Before a state settles on an approach for its accountability system, the state’s analysts and policymakers should use existing data to examine the possible impact on the validity, reliability, and credibility or statistical soundness (i.e., statistical conclusion validity) of the results when different scenarios for assumptions are explored.

Finally, researchers have observed that the number of subgroups for which data are not reported increases as the minimum number of students required for inclusion in a state accountability system increases. This observation raises questions of whether results are generalizable (external validity), credible (statistical conclusion validity), and stable (reliability) when large portions of important subgroups cannot be reported in an accountability system for statistical or privacy reasons. Put another way, these researchers are asking whether an accountability system that does not report data that do not meet minimum reporting thresholds can be viewed as a valid accountability system (Simpson, Gong, and Marion, 2006; Linn, Baker, and Herman, 2002; Harr-Robins, Song, Hurlburt, Pruce, Danielson, and Garet, 2013). As a result, for each scenario a state considers, there should be an analysis of the number and percent of schools, districts, and students that will not have their data reported for each reporting subgroup. This information should be taken into account when a state studies the results from the different scenarios and makes a final decision on a minimum n-size for the state’s accountability system.

Chapter 3. Protecting Personally Identifiable Information

ESSA requires this report to describe how the minimum number that is determined as part of a state's accountability system will not reveal personally identifiable information (PII) about students. Building on research initiated by the National Center for Education Statistics (Seastrom, 2010b), the Department of Education's Disclosure Review Board (ED DRB) concluded that in order to fully protect against reporting estimates that reflect the experiences of one or two students and risk disclosing information about those students, a minimum of 301 students would be needed to report estimates in integers, and 3,001 students would be needed to report estimates to one decimal place (Appendix B: ED DRB Data Protection Schema).

However, as a practical matter, education data need to be reported for groups much smaller than 301 students. Therefore, the ED DRB's efforts have focused on data protection techniques that are applied in the analysis and reporting of data. In the interest of encouraging the use and reporting of as much data as possible while still protecting individual student's PII, this report follows the approaches taken by the ED DRB and emphasizes data protection techniques that support the analysis and reporting of data.

To begin this discussion, it is useful to review the definition of *personally identifiable information* in the Family Educational Rights and Privacy Act (FERPA) (20 U.S.C. Section 1232g; 34CFR Part 99).¹⁶

BOX 1: Family Educational Rights and Privacy Act Definition of Personally Identifiable Information

FERPA defines **personally identifiable information** as information that includes:

- the name and address of a student's family;
- a personal identifier, such as the student's Social Security number, student number, or biometric record;
- other indirect information, such as the student's date, place of birth, and mother's maiden name;
- other information that, alone or in combination, is linked or linkable to a specific student that would allow a reasonable person in the school community, who does not have personal knowledge of relevant circumstances, to identify a student with reasonable certainty; and
- information based on a targeted request (Seastrom, 2010a).*

* A "targeted request" refers to a request for information in which the person asking for the data has an expectation that it will relate to a specific student. For example, if there was a rumor published in the local paper that a public official was disciplined for cheating during his senior year in high school, a request to the high school for the disciplinary records of students who were caught cheating during the year the public official was a senior would be considered a targeted request.

Because each state's accountability system is comprised of a set of outcome measures that aggregate the results of individual students, the most relevant aspects of the FERPA definition include "other indirect information, such as the student's date and place of birth and mother's maiden name" and

¹⁶ For more information about FERPA, visit the U.S. Department of Education's Family Policy Compliance Office website at <http://www2.ed.gov/policy/gen/guid/fpco/ferpa/index.html>. For other privacy-related resources, including methods for protecting sensitive and individually identifiable student information, visit the U.S. Department of Education's Privacy Technical Assistance Center (PTAC) at <http://ptac.ed.gov/>.

“other information that, alone or in combination, is linked or linkable to a specific student.” The challenge is for state accountability systems to display aggregate data without enabling a viewer to learn personally identifiable information about an individual student. Given the requirement to report the results for each outcome measure by each subgroup (i.e., economically disadvantaged students, students from racial and ethnic groups, children with disabilities, and English language learners), the permutations and volume of reported data raise the possibility of a viewer combining this information to identify an individual student’s data with reasonable certainty.¹⁷

Data Protection Techniques to Minimize Inadvertent Disclosures in Public Reporting

A range of data protection techniques are available to help guard against the unintentional release of PII in public reporting. Policymakers and analysts preparing data for their state’s accountability system should consider these protection options in light of the anticipated use of accountability system data. Note, however, that once a minimum number of students needed for valid, reliable, and statistically sound estimates have been established, each additional action taken to protect data for public release has a potentially negative impact on the remaining amount and quality of information available for reporting (Federal Committee on Statistical Methodology, 2005).

One approach to balancing transparency in public reporting and data protection concerns is to use the data for analytical and evaluative purposes prior to implementing data protections, and then implement suitable data protection techniques when preparing the data for public release. If this approach is implemented, the protected (reported) data must be presented in a way that is consistent with analytical conclusions and policy decisions driven by the unprotected data without inadvertently disclosing PII. With such questions in mind, the following discussion reviews several data protection best practice methods currently in use with aggregated education data.

The Federal Committee on Statistical Methodology acknowledges that “publicly available data may not be adequate for certain statistical studies” (Federal Committee on Statistical Methodology, 2005, page 8).

For some uses, decisions may need to be made before data protections are implemented.

Primary and Complementary Cell Suppression

Establishing the minimum number of students in a population or a subgroup is the first step in protecting PII. However, as demonstrated in example 4, displaying results for a small category within a subgroup can inadvertently lead to the identification of an individual student.

The **threshold rule** specifies a minimum reporting size for breakout categories in public reporting.

Primary suppression refers to the process of withholding data values in public reporting data that do not meet the threshold rule—in other words, removing data to protect the identity of individual students.

¹⁷ Examples of the many and varied ways in which an individual student might be identifiable through the release of aggregated results for student subgroups are provided in a 2010 NCES report *Statistical Methods for Protecting Personally Identifiable Information in Aggregate Reporting* (Seastrom 2010b), which can be downloaded from the National Center for Education Statistics (NCES) at <http://nces.ed.gov/pubs2011/2011603.pdf>.

To guard against the unintentional identification of an individual student in publicly reported data (example 4), it is a good practice to specify a minimum reporting size for categories within an accountability measure (e.g., achievement levels on a student assessment). This minimum, which is sometimes referred to as the threshold rule, identifies the categories in a subgroup that are sensitive because the number of students in the category is less than the specified threshold rule. Some data collection agencies set this number at 5, while others set it at 3 (Federal Committee on Statistical Methodology, 2005).

Example 4: Suppression of small subgroups—but not small reporting categories—can lead to inadvertent disclosure of PII in public reporting.

Consider the following data table in a public report. Are there any disclosures of PII that could reasonably be foreseen and avoided through the application of data protection techniques for public reporting?

Table 2. School-level grade 4 mathematics assessment results in a state with a minimum reporting group size of 10

		Percent		Below			
		Assessed	Tested	Basic	Basic	Proficient	Advanced
Total	%	100	100	12.5	31.3	34.4	21.9
	N	†	32	4	10	11	7
White	%	100	100	0	22.7	45.5	31.8
	N	†	22	0	5	10	7
Hispanic	%	100	100	40	50	10	0
	N	†	10	4	5	1	0

† Denotes not applicable.

NOTE: Details may not sum to totals because of rounding.

If a minimum reporting size of 10 is applied to these data, the number of students tested meets that threshold for the Total Count as well as the White and Hispanic subcategories. However, when the assessment results of the 10 Hispanic students are reported across the four achievement levels (see highlighted cells), the number of students at each achievement level falls below the established minimum reporting size of 10: there are 4 students in the Below Basic achievement group, 5 students in the Basic achievement group, 1 student in the Proficient achievement group, and 0 in the Advanced group. Because the minimum size rule was applied for the total number of students tested in the subgroup, these achievement level data are reported in the example and inadvertently disclose personally identifiable information. By reporting that only one Hispanic child scored at the Proficient level (and 0 above), anyone who is able to identify that Hispanic child, such as that student's parents who know he or she scored at the Proficient level, also knows that all of the other Hispanic children in the fourth grade failed to reach the proficient achievement level on the mathematics assessment.

To protect from the inadvertent disclosure of private information, the data for categories that fall below the threshold are not displayed (i.e., they are suppressed, which is known as primary suppression). In the case of example 4, any assessment results (Below Basic, Basic, Proficient, and Advanced) with fewer than 5 students would be suppressed. In this case, the results for category 1 (Below Basic), category 3 (Proficient), and category 4 (Advanced) are suppressed. However, if only one category of data is suppressed, and data for the subgroup total and the remaining categories are displayed, the suppressed data can be easily reconstructed (example 5).

Example 5: The necessity of complementary suppression

Consider the example of reporting counts and the percentage distribution for 39 students on a 4-category performance metric. In table 3a the data for the 2 students in performance category 1 (Below Basic) are suppressed and there are 15 students in performance category 2 (Basic), an additional 17 in performance category 3 (Proficient), and 5 students in performance category 4 (Advanced). To help protect against the identification of one or both of the 2 students in category 1, an analyst performs primary suppression on the data for these 2 students (table 3a).

Table 3a. School-level grade 4 reading assessment results in a state with a minimum reporting group size of 10 and the application of primary suppression for a subgroup minimum threshold of 5.

	Percent		Below				
	Assessed	Tested	Basic	Basic	Proficient	Advanced	
Black	%	100	†	38	44	13	
	N	†	†	15	17	5	

† Denotes not applicable.

‡ Reporting standards not met.

NOTE: Details may not sum to totals because of rounding.

But with only one value suppressed, the missing count and percentage can be easily reconstructed by summing the counts in the three unsuppressed categories and then subtracting that value from the total [$39 - (15 + 17 + 5) = 2$]. Similarly, the percentages reported for the three unsuppressed categories can be summed and then subtracted from 100 percent to ascertain the value for the suppressed category [$100 - (38 + 44 + 13) = 5$]. To avoid relatively simple recovery of suppressed data, a second category should be suppressed. This is referred to as complementary suppression (Federal Committee on Statistical Methodology, 2005) (table 3b).

Table 3b. School-level grade 4 reading assessment results in a state with a minimum reporting group size of 10 and the application of complementary suppression for a subgroup minimum threshold of 5.

		Percent		Below			
		Assessed	Tested	Basic	Basic	Proficient	Advanced
Black	%	100	100		38	44	
	N	†					

† Denotes not applicable.

‡ Reporting standards not met.

NOTE: Details may not sum to totals because of rounding.

The category selected for complementary suppression (typically the next smallest category) would be category 4 (Advanced) with 5 students. This leaves the data user knowing that 32 of the 39 students (82 percent) are in performance categories 2 (Basic) and 3 (Proficient), and the remaining 7 students (18 percent) are in the lowest and/or highest performance category, but provides no information about how the students are distributed across those groups—are they all in the lowest performance category, all in the highest performance category, or split between the two either evenly or disproportionately? Thus, while the use of complementary suppression decreases the likelihood of the inadvertent disclosure of PII, it can also decrease the utility of the available data. In this case, the missing data have potentially important information regarding whether the students represented by suppressed data values are in need of targeted assistance or are already achieving at an advanced level.

Insofar as the ESSA calls for the use of at least three outcome categories for each measure used in the state accountability systems, it is possible to envision a case in which three categories are intended to be reported, but one of the three requires suppression because it does not meet the reporting size threshold and, therefore, necessitates the subsequent application of complementary suppression in a second category. Thus, the net effect of these critical privacy techniques is that two of the three reporting categories end up being suppressed in public reporting.

While primary and complementary suppression can be powerful tools for protecting sensitive data in one row or one column of data, in the case of two-way or multi-level tables, such as the percentage of students performing at different achievement levels disaggregated by race/ethnicity, interactions between the suppressed cells, other values in the same rows and columns as the suppressed data, the row totals, column totals, and the overall total can be used to recover the suppressed data. Such complexity warranted the following caution from the 2005 Federal Committee on Statistical Methodology Working Paper: “While it is possible to select cells for complementary suppression manually, in all but the simplest of cases, it is difficult to guarantee that the result provides adequate protection” (page 17). One solution to this problem is to combine cell suppression with other data protection techniques.

Complementary suppression

refers to excluding data from publication as necessary to avoid the recovery of data that have undergone primary suppression. In many cases, data that have been withheld due to complementary suppression do not disclose PII on their own, but can be used to calculate data that have been suppressed to protect PII.

The Use of Ranges and Top and Bottom Coding to Replace Specific Data Values

Protecting the ends of the distribution is especially important when safeguarding individual student data is a concern. In cases in which the values within individual cells approach 0 percent or 100 percent, bottom and top coding is often employed to minimize the risk of identifying individual students (Federal Committee of Statistical Methodology, 2005, page 25). Such recoding is typically accomplished by substituting “greater than 95 percent” (“> 95”) for all percentage values that are above 95 percent and “less than 5 percent” (“<5”) for all percentage values that are below 5 percent. This method is undertaken to avoid reporting the fact that all, or nearly all, of the students in a population subgroup share the same achievement level or the same outcome, or that very few or none of the students share a particular outcome. Doing so reflects similar logic as the threshold rule: recoding the ends of the distribution ensures that there is a sufficient number of students in the category to protect the identity of individual students.

Recoding Data Values

Reporting a range of values can be used to avoid reporting that all or nearly all (or none or nearly none) of the students in a population or subgroup share the same achievement level or the same outcome.

For example, data values of 96, 97, 98, 99, and 100 percent can be reported as “>95 percent.”

Similarly, values that are “less than 5 percent” (i.e., 4, 3, 2, 1, and 0 percent) are coded as “<5 percent.”

Top and bottom coding involves reporting a group of outcomes as a range of possible values that fall above or below a specified cut point. Just as the top- and bottom-coded ranges protect small numbers of students from being identified (as having or not having a specific educational outcome), other parts of a distribution can be recoded into ranges to reduce the amount of data loss that occurs with small cell suppression (Federal Committee of Statistical Methodology, 2005, pages 18 and 26). The extent of recoding required to protect small categories is related to the size of the subgroup, with a larger recoded range required for smaller subgroups. At a minimum, results should not be published for outcomes based on the experiences of one student. In Seastrom 2010b, the recommendation was to not include the group or subgroup totals or the exact counts that support the percentage distributions in resulting accountability tabulations. This protection was coupled with the introduction of reported ranges across the distribution to ensure that each category represented at least two students (see Recommendations in Seastrom 2010b).¹⁸

In subsequent work, the ED DRB concluded that it often is not practical to suppress population and subgroup totals in public reporting because the loss of potentially important information decreases the utility of the data. As a result, the methodology underlying the Seastrom 2010b report was adapted to a

¹⁸ The recommended ranges were constructed so that each percent value in the reported range represented at least 2 students, with some of the percentages in the range representing 3 students. In such a scenario, if the total number of students in the subgroup is excluded from the tabulation, there is no way of calculating whether the actual number of students in the category is 2 or 3. In contrast, if the number of students in the subgroup is displayed or otherwise known, a viewer trying to undo data protections could conceivably uncover the fact that the actual number of students is 2.

schema that includes reporting the group and subgroup totals and using a threshold of 3, rather than 2, for the reported ranges (i.e., each percentage in a displayed range could represent at least 3 students).

The ED DRB schema uses primary suppression for the percentages for all subgroups with only 0 to 5 students.¹⁹ It also advises that reported categories be displayed as a range of possible values to prevent suppressed data from being recovered (through calculations using the totals and unsuppressed values in other categories). These ranges also insure that each displayed range includes at least 3 students, thus ensuring that results are not reported based on 1 or 2 students. The size of the reported ranges is determined by the size of the group whose data are being displayed. As the number of students in the group or subgroup increases, the size of the range decreases until there are more than 300 students in the group or subgroup (table 4).

Table 4. Reporting ranges for percentages, by reporting group or subgroup size.

Population Size	Reporting Ranges
0 - 5	Suppressed
6 - 15	<50%, ≥50%
16 - 30	≤20%, 21-39%, 40-59%, 60-79%, ≥80%
31 - 60	≤10%, 11-19%, 20-29%, 30-39%, 40-49%, 50-59%, 60-69%, 70-79%, 80-89%, ≥90%
61 - 300	≤5%, 5-9%, 10-14%, 15-19%, 20-24%, 25-29%, 30-34%, 35-39%, 40-44%, 45-49%, 50-54%, 55-59%, 60-64%, 65-69%, 70-74%, 75-79%, 80-84%, 85-89%, 90-95%, ≥95%
301 - 3,000	≤1%, whole number percentages, ≥99%
More than 3,000	≤0.1%, percentages to one decimal place, ≥99.9%

On the low end of the distribution, data for 6 to 15 students are recoded into ranges of less than 50 percent and greater than or equal to 50 percent; at the upper end of the distribution, data for 301 to 3,000 students are recoded as less than or equal to 1 percent, greater than or equal to 99 percent, and as whole numbers between 2 and 98 percent.²⁰ (See Appendix B for a discussion of how these numbers of students and the related ranges were determined.) The use of these ranges maximizes the amount of information that can be released, especially at the school or district level, while reasonably protecting personally identifiable student information. The ED DRB has used this approach to recoding for percentage distributions of assessment performance level results as well as for adjusted cohort graduation rates reported at the school level.

¹⁹ In instances in which there is a compelling interest to preserve reported zeroes, the 0 is reported and values of 1-5 are suppressed.

²⁰ Data for groups of 3,001 or more students can be recoded as less than or equal to 0.1 percent or greater than or equal to 99.9 percent, with numbers with one decimal place of precision between the two tails of the distribution.

In instances in which the rates or percentage distribution of an outcome measure are displayed at the state level, the ED DRB has approved less restrictive recoding—the logic being that aggregation to the state level already minimizes the risk of disclosure of individual data. In these cases, the outcome percentage is suppressed for those populations and subgroups with 1 to 5 students and, if the population total is displayed, the data for the next smallest subgroup are also suppressed. For populations and subgroups ranging from 6 to 3,000 in size, categories that are close to 0 or 100 percent are bottom or top coded using the values for the tails of the distributions for the specified population sizes shown in table 4, and the rest of the data are displayed as whole number percentages. For populations and subgroups with more than 3,000 students, categories that are close to 0 or 100 percent are bottom or top coded, and the rest of the data can be displayed as tenths of a percent.

There are other instances where the data displayed are limited to counts of students with a shared characteristic (e.g., students with disabilities, migrant students, students in limited English proficiency programs). In some cases, the ED DRB has concluded that counts aggregated to the state or local education agency level do not need further protections. For example, in the IDEA 618 Part B Dispute Resolution (state-level data), the fact that one student could be included in multiple disputes and that a single dispute could involve multiple students prevents the counts in this dataset from being used to identify individual students.²¹

For data releases in which protections are considered necessary, the ED DRB has approved several scenarios that rely on suppression or recoding, including top and bottom coding.

ED DRB Scenario 1: Counts of students are protected by primary and complementary suppression of small categories. In this scenario, state-level counts of students with disabilities are suppressed for each subgroup that includes categories with 1 or 2 students, and complementary suppression is then used for the next smallest nonzero subgroup within each state. If national totals are included, each category suppressed within a single state requires complementary suppression in at least one other state. School district counts of migrant students provide another example where the ED DRB approved the use of data suppression to protect count data; in this case the counts were suppressed for districts with less than 30 migrant students.

ED DRB Scenario 2: Primary and complementary suppression of small categories are used in addition to top coding of categories with counts that are close to the total. In this scenario, school-level counts of limited English proficiency (LEP) students in LEP programs are suppressed for all categories, subgroups, and totals with counts of less than 5 students, and categories that are within 5 students of the total count are top coded to avoid releasing data that would reveal the fact that all or nearly all of the students in that group or subgroup are LEP students enrolled in LEP programs. To further protect the released data, complementary suppression is used to protect any category that is suppressed or recoded.

²¹ As another example, under IDEA Part B, Coordinated Early Intervening Services (CEIS) reporting is highly unlikely to enable the identification of special education students through counts of students receiving CEIS because (1) the data are not disaggregated by any demographic characteristics; (2) the data are cumulative over a 2-year period; (3) not all students receiving CEIS are identified as needing special education or related services; and (4) not all special education students receive CEIS.

ED DRB Scenario 3: Counts of categories are bottom coded to less than or equal to 3 and row and column totals are not reported. In this scenario, state-level counts of “students with disabilities who are proficient or above” that are reported by testing conditions for the state assessment are bottom coded in categories with 3 or fewer students (i.e., ≤ 3). But rather than using complementary suppression, row and column totals and the overall total are not reported for this data file, meaning that more data values are displayed (because there was not complementary suppression) and those values that did receive primary suppression cannot be calculated through simple subtraction from a total count that is not displayed.

Rounding

Another method of data perturbation, or adding noise to the data, is the use of a systematic rounding routine. Such an approach has been applied to data from the Civil Rights Data Collection (CRDC).²² Like most of the examples in this discussion of alternate methods of protecting against the identification of individual students, the CRDC data do not include individual-level data; instead they are aggregated at the school and district levels. However, CRDC data include disaggregations for groups of students with shared characteristics and, as a result, pose a risk of student identification when subgroups are small or when all, or nearly all, of the subgroup members share a common response on one data element. As such, there is an increased possibility that a reasonable person in the school community could determine with some certainty that a specific individual student is included in a reported value. To guard against this possibility, two slightly different systematic rounding techniques are applied to data at the school and district levels.

In order to preserve meaningful zeroes, actual 0 values are reported for a subset of CRDC data elements, but the remaining values are rounded to the middle value in consecutive groupings of 3 values (e.g., values of 1, 2, and 3 are reported as 2; values of 4, 5, and 6 are reported as 5; and so on). For the remaining data elements that require protection, the zeroes are not preserved. As a result, the rounding schema includes reporting values of 0, 1, and 2 as ≤ 2 ; values of 3, 4, and 5 are rounded to 4; values of 6, 7, and 8 are rounded to 7; and so on. To maintain consistency, all row and

For the purpose of this discussion, **data perturbation** is the intentional introduction of an element of uncertainty (e.g., top and bottom coding and rounding) into data before they are published in order to minimize the likelihood of the identification of an individual student.

Rounding

Rounding refers to altering a number to another approximately similar value for the purpose of convenience or, in this context, to introduce an acceptable level of uncertainty that protects data values without substantially changing their meaning.

For example, in a subset of the CRDC,

- values of 1, 2, and 3 are reported as 2;
- values of 4, 5, and 6 are reported as 5;
- etc.

To maintain consistency, all row, column, and overall totals are calculated as the sum of the rounded data. Percentage distributions across subgroup categories are also computed using the perturbed data.

²² Visit <http://ocrdata.ed.gov/> for more information about the U.S. Department of Education’s Civil Rights Data Collection.

column totals and the overall totals are calculated as the sum of the rounded data. Percentage distributions across subgroup categories are also computed using these perturbed data. These two rounding approaches are, in effect, the equivalent of grouping the observed values into ranges of 3 and then using the center value rather than reporting the data as a range. Although each of these ranges has only three possible solutions, the fact that the reported totals for groups and subgroups are the summation of the rounded values prevents precise values from being reconstructed and introduces enough uncertainty to any single value to minimize the likelihood of the identification of a specific individual student.

Caveat: The Hierarchical Structure of Education Data

Education data for elementary and secondary education have a hierarchical structure, with individual student data being aggregated to classrooms, grade levels, schools, school districts, states, and the nation. Because the data at each level are an aggregate of the data from lower levels, this hierarchy leads to interdependencies in the data that must be taken into account when assessing the risk of identifying a specific individual through released data.

Given the hierarchical nature of education data, policymakers and analysts should understand that protection decisions made for one level of data (such as data intended to be used in schools) may limit the amount of detail that can be reported at another level (such as the district, state, or nationally).

Minimally, there is an interest in ensuring that protections implemented at one level in the hierarchy do not undo protections used at a lower or higher level (Federal Committee on Statistical Methodology, 2005). This is certainly the case in a state accountability system that includes data aggregated from students, grade levels, schools, and school districts to the state level. For example, in the case of the CRDC data, which is collected and reported at multiple levels of the education system, a decision was made to apply rounding at the lowest level of data (e.g., the school) and to maintain consistency by calculating all row totals, column totals, and the overall totals as the sum of the rounded data. As a result, data that are submitted at the school level undergo rounding at the school level. These numbers are then combined to produce district data, which are in turn combined with data from other districts to produce state-level data. The state data are subsequently combined to produce national data—the results of which may or may not be closely related to the results that would come from the aggregation of unrounded data originating at the school or district level.

As discussed previously, the ED DRB has approved the use of small cell suppression coupled with complementary suppression. However, when considering the hierarchical nature of education data (which is typically apparent in the geographical or organizational nesting of the education system), this assumes that the suppression schema adequately accounts for the hierarchical nesting of data when the same data are going to be reported at multiple levels. For example, when data are suppressed for a specific subgroup in one school within a district, the results for that subgroup must be suppressed for either another school in that district or for the entire district. Otherwise, simple subtraction allows the suppressed values to be calculated from the totals. The same logic applies to suppressions implemented at the district level when the data are aggregated to the state level and to state-level data when they are aggregated to national totals (Seastrom, 2010b).

The data protection methods implemented at one level in the education hierarchy have strong implications for subsequent data protection decisions at higher and lower levels in the hierarchy.

Policymakers and analysts must consider the varied uses of the data when deciding how and where to implement data protection schema. For example, if the school is the primary unit of analysis, data protections should be applied at the school level with a goal of maximizing the amount of information available at the school level. The policymaker and analyst should understand that protection decisions made for school-level data may limit the amount of detail that can be reported at the district or state level. Conversely, if a state outcome measure is the focus of interest, then maximizing the level of detail available at the state level may require protections that limit the amount of information available at the district and school levels.

Additional Resources

Under the auspices of the congressionally mandated National Cooperative Education Statistics System, the National Center for Education Statistics founded the National Forum on Education Statistics (the Forum) to help support the production, maintenance, and use of comparable and uniform elementary and secondary education statistics across states and school districts. Through the Forum, representatives of state, local, and federal agencies and other organizations with an interest in education data collaborate to develop resources and produce reports and best practice guides intended to provide helpful information that will further the development and maintenance of a robust system of education statistics. Of direct relevance to this *Report* are the three Forum reports published on the topic of privacy in student data. The most recent report in this series is the *2016 Forum Guide to Education Data Privacy*. This guide includes information on federal and state privacy laws; the interrelationships among data governance, data security, and data privacy; and includes a set of case studies that highlight the management of 11 common privacy issues that arise in using student data.

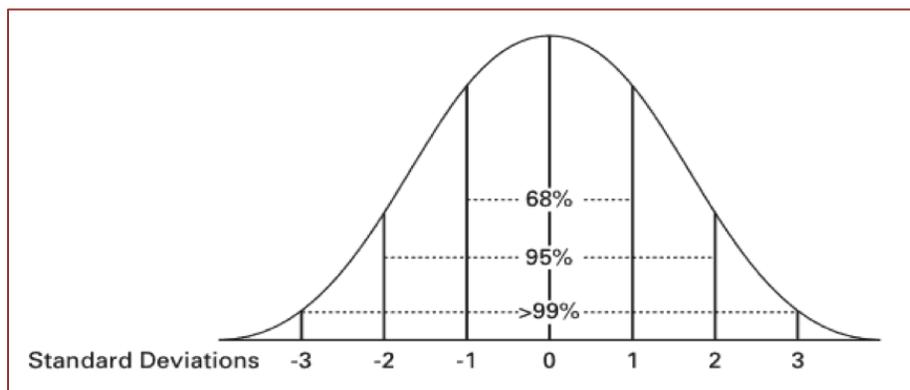
The Department of Education's Privacy Office supports a Privacy Technical Assistance Center (PTAC) that serves as a resource for education stakeholders wanting to learn about best practices for protecting personally identifiable student information and for promoting compliance with the Family Educational Rights and Privacy Act (FERPA). PTAC has developed a series of guidance documents on protecting student privacy when using student-level data systems for education decision-making and reporting, and will be publishing additional guidance on protecting privacy in public reporting over the coming year. Information about the PTAC and its resources is available at <http://ptac.ed.gov/>. SEA and LEA officials who need technical assistance or have questions on FERPA may contact PTAC's Student Privacy Help Desk by emailing PrivacyTA@ed.gov.

Appendix A: Sampling

The normal distribution is fundamental to how the estimate of a population characteristic and the related measures of dispersion are used to draw inferences about the population.

A normal distribution is a theoretical continuous probability distribution that is represented graphically by a plot with all of the possible values of a variable on the horizontal (x) axis and the probability of those values occurring on the vertical (y) axis. In a normal distribution, the estimates of a variable are clustered around the mean in a symmetrical pattern, meaning that the upper and lower halves of the distribution are mirror images of one another. Thus, the distribution is highest in the middle, creating a unimodal or bell-shaped distribution in which the mean, median, and mode are all the same.

For all normal curves, the area under the curve is equal to 1 and the probability that a normally distributed random variable falls within a specific interval is the area under the curve within the applicable interval on the x axis. In a normal distribution, 68 percent of the values of a variable fall within plus or minus 1 standard deviation from the sampling mean (i.e., 1 standard error) and 95 percent fall within a confidence interval that is defined as the range that is plus or minus 2 standard deviations from the sampling mean (i.e., 2 standard errors).



SOURCE: Retrieved from http://grants.hhp.coe.uh.edu/doconnor/PEP6305/Ncurve_SDs.gif

The Central Limit Theorem is critical to understanding why the normal distribution is important for determining a sound value for a minimum number of students. Simply put, the Central Limit Theorem states that even for populations that are far from normal, the sampling distribution of the mean will be normally distributed as long as the sample size is large enough.²³ Notably, the sampling distribution becomes nearly normal for a wide range of population distributions as long as the sample size is greater

²³ Central Limit Theorem: For any population that has a mean μ and a finite variance σ^2 , the distribution of sample means (each based on N independent observations) will approach a normal distribution with mean μ and variance σ^2 divided by N , as N approaches infinity (Barry H. Cohen, 2001).

than or equal to 30 (Cohen, 2001; Cohen and Lea, 2004; Mendenhall and Ott, 1980; Urdan, 2001; Vogt, 2005).²⁴

The amount of variability in sampling means decreases as the sample size increases.²⁵ Thus while a sample size of 30 is sufficient under the Central Limit Theorem, the margin of error for a sample of 30 is larger than it would be with a larger sample. The analyst or policymaker setting the minimum number of students in a population or subgroup required for use in the accountability system must evaluate whether the margin of error associated with the minimum number of students meets their criteria for statistical conclusion validity and for reliability.

²⁴ An important caveat is that the sample size needed to approximate a convergence of the sampling distribution to normality is related to how close the population distribution is to normal, with larger sample sizes required the further the population distribution is from normal.

²⁵ The standard error (i.e., the standard deviation of the sample means), is computed by dividing the standard deviation of the sample by the square root of the sample size. As the sample size increases, the denominator increases and the resulting estimate of the standard error decreases; in other words, the amount of error around the measured mean decreases.

Appendix B: ED DRB Data Protection Schema

The U.S. Department of Education's Disclosure Review Board (ED DRB) concluded that it often is not practical to rely solely on suppression as a protection schema in public reporting because the loss of potentially important information decreases the utility of the data. As a result, the ED DRB uses both suppression and range data protection schema for all subgroups with 5 or fewer (0 to 5) students (table 4 in the body of the report). To prevent suppressed data from being recovered and to further protect against the identification of individual students in subgroup categories that include fewer than 3 students, the reported categories are displayed as a range of possible values.

This process is initiated by recoding the low end of a distribution, which involves determining the number of students that would need to be included in the recoded value to ensure that the product of that number and the range of the recode yields 3 students. In order to include percentages based on 6 students, while still ensuring that at least 3 students are in each displayed category, data for very low counts of students must be recoded as less than 50 percent or greater than or equal to 50 percent (i.e., <50%, ≥50%).

For the next category in the distribution, the decision was made to identify the number of students needed to support a range of 20 percentage points. Because 20 percent of 16 is 3.2, a count of 15 was identified as the upper end of the number of students to include in the 50 percent recode and 16 was selected as the low end of the number of students for the next group of recodes. However, since the sum of the remaining reported categories subtracted from 100 percent yields the exact percentage that was recoded, simply protecting the percentages at one end of a percentage distribution is not sufficient to protect the original contents of the recoded category. Rather than relying on complementary suppression, the ED DRB approach collapses the percentage distributions of the remaining categories into ranges that are the same width as that identified for the low end of the distribution. Thus, the width of the ranges for populations or subgroups that include 16 students is set at 20 percentage points and the results are recoded in 20 percentage point intervals (i.e., ≤20%, 21-39%, 40-59%, 60-79%, ≥80%).

Ten percentage points was identified as the range for the next set of recodes. Because 10 percent of 31 is 3.1, 30 was identified as the upper end of the number of students to include in the 20 percent recode and 31 was selected as the low end of the number of students for the recodes with a 10 percentage point width (i.e., ≤10%, 11-19%, 20-29%, 30-39%, 40-49%, 50-59%, 60-69%, 70-79%, 80-89%, ≥90%).

For the next set of recodes, the range was dropped to 5 percent. Because 5 percent of 61 is 3.05, 60 was identified as the upper end of the number of students to include in the 10 percent recode and 61 was selected as the low end of the number of students for recodes with a 5 percentage point width (i.e., ≤5%, 5-9%, 10-14%, 15-19%, 20-24%, 25-29%, 30-34%, 35-39%, 40-44%, 45-49%, 50-54%, 55-59%, 60-64%, 65-69%, 70-74%, 75-79%, 80-84%, 85-89%, 90-95%, ≥95%).

This process continued with the next set of recodes limited to a range of 1 percentage point. Because 1 percent of 301 is 3.01, the upper end of the number of students to include in the 5 percent recode was set at 300, and 301 was selected as the low end of the number of students for recodes with a 1 percentage point width. Thus the results for groups of students that include at least 301 students can be reported as whole number percentages (i.e., ≤1%, whole number percentages, ≥99%).

Continuing this logic one more step, in those cases in which there is an interest in also reporting data to one decimal point, 3,001 students are needed (i.e., ≤0.1%, percentages to one decimal place, ≥99.9%).

References

- Barnett, V., 2002. *Sample Surveys: Principles and Methods*. New York, NY: Oxford University Press.
- Biemer, P.P., Groves, R.M., Lyberg, L.E., Mathiowetz, N.A., and Sudman, S., 1991. *Measurement Errors in Surveys*. New York, NY: John Wiley & Sons.
- Blair, E. and Blair, J., 2015. *Applied Survey Sampling*. Thousand Oaks, CA: Sage Publications Inc.
- Boudett, K.P., City, E., and Murnane, R.J., 2008. *Data Wise: A Step-by-Step Guide to Using Assessment Results to Improve Teaching and Learning*. Cambridge, MA: Harvard University Press.
- Code of Federal Regulations, Title 14—Education, Part 99. *Family Educational and Privacy Rights*, (34CFR99). Washington, DC: GPO Access e-CFR, Retrieved from http://ecfr.gopaccess.gov/cgi/t/text-idx?c=ecfr&sid=44d350c26fb9cba4a156bf805f297c9e&tpl=/efcbrowse/Title34/34cfr99_main_02.tpl.
- Code of Federal Regulations, Title 14—Education, Part 200. Notice of Public Rulemaking (NPRM) *Title 1—Improving the Academic Achievement of the Disadvantaged*. (34CFR200). Washington, DC: GPO Access e-CFR. Retrieved from <https://www2.ed.gov/legislation/FedRegister/proprule/2002-3/080602a.html>.
- Code of Federal Regulations, Title 14—Education, Part 200. Final Rule. *Title 1—Improving the Academic Achievement of the Disadvantaged*. (34CFR200). (34CFR99). Washington, DC: GPO Access e-CFR. Retrieved from <https://www2.ed.gov/legislation/FedRegister/finrule/2003-1/010803a.htm>.
- Cohen, B.H., and Lea, R.B., 2004. *Essentials of Statistics for the Social and Behavioral Analysis*. Hoboken, NJ: John Wiley & Sons.
- Cohen, J., 1988. *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Erlbaum.
- Cronbach, L. J., Linn, R.L., Brennan, R.L., and Haertel, E.H., 1997. *Generalizability Analysis for Performance Assessments of Student Achievement or School Effectiveness*. Educational and Psychological Measurement, Volume 57, Number 3, June 1997, 373-399.
- Deming, W.E., 1990. *Sample Design in Business Research*. New York, NY: John Wiley & Sons.
- Ellis, P.D., 2010. *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results*. Cambridge, UK: Cambridge University Press.
- Everitt, B.S., 1999. *The Cambridge Dictionary of Statistics*. Cambridge, UK: Cambridge University Press.
- Federal Committee on Statistical Methodology, 2005. *Report on Statistical Disclosure Limitation Methodology, Statistical Policy Working Paper 22 (Second version)*. Washington, DC: Statistical and Science Policy, Office of Information and Regulatory Affairs, Office of Management and Budget. Retrieved from <http://fcsm.sites.usa.gov/reports/policy-wp/>.

- Ferguson, G.A. and Takane, Y., 1989, Sixth Edition. *Statistical Analysis in Psychology and Education*. New York, NY: McGraw-Hill Book Company.
- Graziano, A.M. and Raulin, M.L., 1989. *Research Methods: A Process of Inquiry*. New York, NY: Harper & Row, Publishers, Inc.
- Harr-Robins, J., Song, M., Hurlburt, S., Pruce, C., Danielson, L., and Garet, M., 2013. *The Inclusion of Students with Disabilities in School Accountability Systems: An Update*. National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences. U.S. Department of Education, Washington, DC.
- Hill, R.K. and DePascale, C.A., 2003. *Reliability of No Child Left Behind Accountability Designs*. The National Center for the improvement of Educational Assessment, Inc.
- Joint Committee on Educational and Psychological Testing, 2014. *Standards for Educational and Psychological Testing*. Washington, DC: AERA.
- Joint Committee on Standards for Program Evaluation, 1999. Second Edition. *The Program Evaluation Standards* Thousand Oaks, CA: SAGE Publications.
- Levy, P.S. and Lemeshow, S., 2008. *Sampling of Populations: Methods and Applications*. Hoboken, NJ: John Wiley & Sons.
- Linder, F.E. and Grove, R.D., 1947. *Vital Statistics Rates in the United States, 1900-1940*. National Office of Vital Statistics, U.S. Public Health Service: Washington, DC.
- Linn, R.L., Baker, E.L., and Herman, J.L., 2002. "Minimum Group Size for Measuring Adequate Yearly Progress." *The CRESST Line*, Newsletter of the National Center for Research on Evaluation, Standards, and Student Testing, UCLA. Los Angeles, CA.
- Lodico, M.G., Spaulding, D.T., and Voegtle, K.H., 2006. *Methods in Educational Research: from Theory to Practice*. San Francisco, CA: Jossey-Bass.
- Marczyk, G., DeMatteo, D., and Festinger, D., 2005. *Essentials of Research Design and Methodology*. Hoboken, NJ: John Wiley & Sons.
- Marion, S., White, C., Carlson, D., Erpenbach, W .J., Rabinowitz, S., and Sheinker, J. , 2002. *Making Valid and Reliable Decisions in Determining Adequate Yearly Progress*. Council of Chief State School Officers, Washington, DC.
- McLeod, S. A., 2007. *What is Validity?* Retrieved August 12, 2016 from www.simplypsychology.org/validity.html.
- Mendenhall, W. and Ott, L., 1980. *Understanding Statistics*. North Scituate, MA: Duxbury Press.
- Myers, D., 1992. *Analysis of Local Census Data: Portraits of Change*. San Diego, CA: Academic Press Inc.

- National Forum on Education Statistics, 2005. *Forum Guide to Education Indicators* (NFES 2005–802). National Center for Education Statistics, U.S. Department of Education. Washington, DC.
- National Forum on Education Statistics, 2016. *Forum Guide to Education Data Privacy* (NFES 2016–096). National Center for Education Statistics, U.S. Department of Education. Washington, DC.
- Petersen, W., 1969. *Population*. London, England: The MacMillan Company.
- Seastrom, M., 2010a. *Basic Concepts and Definitions for Privacy and Confidentiality in Student Education Records* (NCES 2011-601). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education, Washington, DC.
- Seastrom, M., 2010b. *Statistical Methods for Protecting Personally Identifiable Information in Aggregate Reporting* (NCES 2011-603). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education, Washington, DC.
- Simpson, M.A., Gong, B., and Marion, S., 2006. *Effect of Minimum Cell Sizes and Confidence Interval Sizes for Special Education Subgroups on School Level AYP Determinations*. (Synthesis Report 61). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Trochim, W.M. K., 2006. Second Edition, *Research Methods Knowledge Base*. Retrieved August 12, 2016 from <http://www.socialresearchmethods.net/kb/>.
- Upton, G. and Cook, I., 2006. *Oxford Dictionary of Statistics*. New York, NY: Oxford University Press.
- U.S. Code, Title 20—Education, Chapter 31—General Provisions Concerning Education, Subchapter III—General Requirements and Conditions Concerning Operation and Administration of Education Programs: General Authority of the Secretary, Part 4—Records, Privacy, Limitation on Withholding Federal Funds, Section 1232g. *Family Educational and Privacy Rights*, (20USG1232g). Washington, DC: GPO Access. Retrieved from <http://frwebgate4.access.gpo.gov/cgi-bin/TEXTgate.cgi?WAISdocID=79948617532+0+1+0&WAIAction=retrieve>.
- Urdan, T.C., 2001. *Statistics in Plain English*. Mahwah, NJ: Lawrence Erlbaum Association Publishers.
- Viswanathan, M., 2005. *Measurement Error and Research Design*. Thousand Oaks, CA: SAGE Publications.