Institute of Education Statistics
National Center for Education Statistics

NATIONAL INSTITUTE OF STATISTICAL SCIENCES
WHITE PAPER EXPERT PANEL

SIGNIFICANCE AND NON–SIGNIFICANCE IN NCES REPORTS

TABLE OF CONTENTS

NATIONAL INSTITUTE OF STATISTICAL SCIENCES

SIGNIFICANCE AND NON-SIGNIFICANCE IN NCES REPORTS

EXECUTIVE SUMMARY

For decades the National Center for Education Statistics (NCES) has collected data on the state of education, nationally and internationally, via validated assessments, surveys, and collections of administrative data. Many NCES reports of these data focus on “significant” findings. The prime challenges facing NCES are: What to report as significant, how to report it, and how to explain it.

NCES charged the National Institute of Statistical Sciences (NISS) with convening a panel of technical experts to focus on how significance of findings from data is reported in NCES publications, presentations of data summaries in a variety of forms on the website, and other citations of significance of NCES statistical summaries of NCES data that are produced by or for the Center.

The broad charge to the panel was to examine the representation of significance in recent NCES publications, and to deliberate the conceptual issues of defining significance prior to making recommendations to NCES. In addition, the panel was asked specifically to consider possible definitions of significance including the dichotomy (significant or not, $p < 0.05$) in current use for NCES reports. A second specific request was for the panel to consider possible publication practices and whether to restrict publication to significant findings (i.e., meeting the threshold definition, $p<0.05$). Because reports on multiple variables pose special problems, the panel was asked to review practices for handling multiple tests and to make recommendations for ensuring that quoted probabilities (p-values) are correct. A final request was for the panel to provide advice on effective communication of the meaning of a “significant finding” to a broad readership. The panel met in person in September 2018. This white paper is based on the panel’s report.

The panel’s discussion covered four areas: concepts of significance and importance, statistical issues, standards, and publication practices.

The overarching goal is to reduce the gap in information and in understanding between statisticians and policy makers and the lay public. Therefore, the panel encourages NCES to ensure that reports accurately reflect in full all the important complexities in the data. Recommendations follow, grouped by area.

* White Paper September 2019

NATIONAL INSTITUTE OF STATISTICAL SCIENCES
EXPERT PANEL WHITE PAPER

SIGNIFICANCE AND NON-SIGNIFICANCE IN NCES REPORTS

Primary Recommendations: Significance and Importance

- Lead with magnitude of effect; follow with significance.
- Communicate importance in terms of magnitude and associated variance, probability (e.g., p-value, interval or other) and strength of evidence or sensitivity.
- Replace dichotomization and eliminate nebulous expressions (e.g., "substantially")

Statistics and Methodology

- Expand the collection of analytic methods employed to meet the needs for analysis and interpretation. In particular univariate methods used alone can be seriously misleading because unidimensional analyses cannot reflect interactions, clustering or differences in responses among subsets of the population.

  NOTE: Multivariate analysis is often necessary for accurate interpretation of the data, but such an analysis does not imply causality.
- For multiple tests (or probability statements or intervals) indicate the required adjustments to calculated probabilities.

Planned and Exploratory Analyses

- Require an analytic plan at the outset that specifies analysis to be done and commits to full reporting of all planned analyses.
- Anticipate and allow exploratory analyses that are discretionary, but when reported are separated and clearly identified in the text, noting that probability calculated cannot be correct without adjustment for conditional decisions and multiplicity.
- Report analysis details and process to provide technical support for interpretations as supplemental material.

Standards and Guidelines

- Review and revise (as needed) Standards and Guidelines every 3 to 5 years with attention to relevant advances in statistical and technological methodology. Start with an immediate comprehensive review.
- Add one or more new Standards (and accompanying Guidelines) in each of the following

areas. Seek external consultants with specific expertise where appropriate.

- Statistical graphics and data visualization
- Measuring and reporting model fit for survey and administrative data

- Require that submission of reports for review include specific response to each Standard or Guideline indicating "consider. . ."

Publication Practices

- Write clearly but accurately so that information as interpreted by a broad readership will be consistent with deeper analyses of the data that support the reported results.
- Ensure complete publication of results for all statistical analyses and include statistical methods employed (especially tests!).
- Disseminate reports at two levels by providing details of analyses including analytic process and supporting statistical information. For example, expand Data Point to supply deeper data analysis results by appending or linking to detail required by a more sophisticated reader or policymaker to validate methodology, results and conclusions or to make decisions.
- Indicate precision (and/or probability measure of significance) wherever data is presented
  - text, table, graph, other data visualization.
- Use technology wisely to link elaborations and detailed explanations, additional graphics or data visualizations, and important definitions to simple statements in online reports.

Note on Implementation

The expert panel recognizes that transitioning away from a threshold-based, single-variable-at-a-time conception of significance will require effort, expertise and time to accomplish. Attainable change will be a balance of feasibility in terms of resources (staff time, funding, etc.) with best practices; however, this does not change the urgency for moving forward.

PREFACE

The National Center for Education Statistics (NCES) charged the National Institute of Statistical Sciences (NISS) with convening a panel of technical experts to focus on how significance of findings from data is reported in NCES publications, presentations of data summaries in a variety of forms on the website, and other citations of significance of NCES statistical summaries of NCES data that are produced by or for the Center.

On 13-14 September 2018 the panel of technical experts met in person. This white paper is based on their full report of their deliberations and recommendations.

NATIONAL INSTITUTE OF STATISTICAL SCIENCES
EXPERT PANEL WHITE PAPER

SIGNIFICANCE AND NON-SIGNIFICANCE IN NCES REPORTS

I. BACKGROUND

For decades NCES has collected data on the state of education, nationally and internationally, via validated assessments, surveys, and collections of administrative data. The information these provide is extensive and the quality of these data is exceptional. NCES releases aggregate data and summaries as a public data file. For sample surveys, public use files also include microdata at the person or the institutional level with integrated disclosure avoidance protections. Data at a finer granularity are available for research purposes upon application for a license from NCES.

The overarching goal for NCES reports is to reduce the gaps in information and understanding between statisticians and policy makers at all levels and the lay public.

NCES publishes and also posts online summaries, brief non-technical reports and occasional longer reports. These reports are available to the general public, the education community, education researchers, and policymakers. Currently all these reports focus on “significant” findings.

The prime challenges facing NCES are: What to report as significant, how to report it, and how to explain it. What does “significant” mean in technically accurate but non-technical terms for a lay audience? In technical terms, what justifies the designation “significant”? NCES put this broad question, encompassing what constitutes “significance” of information, of evidence, of findings or of statistics (depending on the data user’s vernacular) and how “significance” is measured, before the expert panel.

To address these questions, the panel first considered the relationships among statistical significance, the prevalent usage of that term, and importance of a finding or result. After framing the issues and identifying the key concepts, the panel turned its attention to the NCES context and relevance of these concepts to NCES reports.

PART ONE

II. STATISTICAL SIGNIFICANCE

The term, “statistical significance,” is usually used in referring to a probability calculated for a particular context, i.e., under the assumption that a specified default case is true. The (statistical) probability, expressed as a p-value, is a measure of rarity in that specified context. Thus the measure of significance depends upon default case being true and upon the magnitude of the departure from the default, the strength of evidence provided by the data and the inherent uncertainty. When multiple statements are made, probability calculations further depend on properly accounting for any dependence among the statements and the contexts and the variables (responses and factors).

Probability and Rarity

“Statistical significance” presented as a p-value specifically refers to the tail of the default distribution, i.e., the total probability of values at least as extreme as the value calculated from the observed data when the *specified default is true*. Often the default case is a zero between-group difference or zero influence of a tested factor. So the probability of deviating from the default by at least as much as is presented by the data, i.e., the “p-value,” reflects: first, the *magnitude* of the difference seen from the data; and second, the *strength* of that evidence, usually quoted instead in terms of its uncertainty (variance, standard error, coefficient of variation or interval). The smaller the p-value, the less likely for the default case to produce data like those actually observed, and the more likely these data would be coming from some alternative to the default. Thus, the p-value, as a numerical representation, can be treated as an indicator of rarity. When the chosen default is not credible or even is known to be untrue, the calculated p-value is not wrong; rather it is silly because the chosen default was already considered untenable.

A widespread practice for at least a century[1], has been to designate the probability of one in twenty (5%) as “rare” in order to assign findings into two classes: “statistically significant” and “statistically non-significant.” Alternatively, “Statistical significance” of information may be reported as an actual p-value rather than a binary classification.

Either way, the strength of evidence influences the probability statement, leading to the questions: How much precision is possible? and How much precision is needed? On the one hand, a large p-value raises the question: Is the p-value large and the classification “non-significant” simply because the evidence was not sufficient (e.g., the sample size was too small) to detect an important difference with precision? On the other hand, the question is raised by a small p-value: Is the p-value so small only because the evidence is so extensive (e.g., the sample size was enormous) or observations so precise that otherwise meaningless differences are “significant”? Answer to the first question is a statistical measure of sensitivity that can be calculated from the default distribution. Answer to the second question must be made in substantive terms, determined by judgment about what constitutes a meaningful magnitude.

[1] Karl Pearson is credited with introducing the notion of a p-value, then referred to as “P” in 1900 (*Philosophical Magazine, Series 5*). In *Statistical Methods for Research Workers* published in 1925, Sir Ronald Fisher proposed the use of 0.05 as the threshold and that has continued in common use to the present day.

Discussions of "significance" are most often formulated – perhaps because most easily articulated – in terms of a single univariate statement. Even this simplest case illuminates the prime question: Is probability due to random chance the "right measure" of significance?

Calculated Probabilities and Independence

In any case, the accuracy of the calculated probability depends on the accuracy of the specifications used in the calculation, i.e., the assumptions about the default probability distribution. In the simplest case – one observed outcome on one homogeneous group to yield one probability statement – meeting the critical assumptions is not usually a problem. Often, however, two or more probability statements are made about different aspects of the same data, either comparisons of different subsets of the data or assessment of different elements of the outcome. In these cases, the univariate distribution for each outcome variable observed on one specified group is no longer the correct and relevant distribution for calculating probabilities.

When there are multiple outcome variables, the relevant distribution is the joint (bivariate or multivariate) distribution. For example, Figure 1 in the *Student Victimization in U.S. Schools*[2], makes it clear that the three specific questions about location of bullying (in schools, outside on school grounds, on a school bus) cannot be independent because the groups of students for these three questions overlap. Since the numbers of students bullied at these locations total more than the number who report bullying, some students must have been targets for bullying in two or more of the three locations. Thus for this example, when these statements about bullying in three locations are taken together, the probabilities will not be correct if calculated independently.

When each of several survey questions or factors is used separately to partition the data; this simply reassembles the data in different ways. In such a case the overlaps from one set of partitions to another usually create dependence that must be properly taken into account in the probability calculation. As an example, for three groups of students in the *Student Victimization in U.S. Schools* (Any Victimization (A), Theft Victimization (T), Violent Victimization (V)), the first group contains all students in the other two; and those two also overlap because some students experienced both violent and theft victimizations. Because of this interdependence, p-values for any response variable for all or any pair of the comparisons A vs T, A vs V and T vs V cannot be treated as separate (simplest case) univariate comparisons. This is in addition to the fact that for three groups (A, B, C), the three tests (A vs B, B vs C and C vs A) would not be independent even if the groups were non-overlapping. Interdependence in this form is a common occurrence. Statistical methods that adjust for multiple comparisons have been developed for precisely these situations and should be put to use.

Relative Importance

When data are multivariate and relationships among variables are complex, a univariate concept of significance is inadequate. Hierarchical linear models and generalized linear models are good examples. In such cases, a good index to relative importance is needed for determining which components are essential for a statistically accurate description of the multi-variable data. Viewed as a single dimensional calculation, the probability, or area of the tail of the distribution, for one selected variable changes

[2] ***Student Victimization in U.S. Schools***: Results from the *2015 School Crime Supplement to the National Crime Victimization Survey*, STATS IN BRIEF, NCES, Draft June 2017.

depending on the ignored dimensions. So the importance of a factor could be measured conditionally on a particular context defined by the other dimensions (factors or variables). Alternatively the relative contribution of the selected variable in the presence of all factors could be measured. The correct choice for means of measurement depends on the inference to be drawn.

Non-literal Use of p-Values

Data exploration is an important research activity but is quite distinct from the data representation and analysis discussed above. For exploration, probabilities in the form of p-values do not have validity as numeric values but rather serve as a tool for examining potentially relevant factors or potentially related outcomes. Exploratory modeling or analysis is data-driven and spontaneous and sequential in nature. Multiple responses may be examined, multiple factors may be included/excluded and multiple ways of defining subgroups can be freely considered. Used in this context the p-value loses its meaning as an accurate probability. Instead it becomes a scaled indicator of the relative contributions from multiple sources of variation, thereby suggesting directions for future inquiry. Even as an indicator it still is influenced by the strength of evidence, which is likely to vary across response variables and across different partitioning's of respondents into subsets. So interpretation of this indicator still must take into account those dependencies when subgroups overlap or when factors or multiple responses are not independent.

III. SIGNIFICANCE AND IMPORTANCE

The panel re-examined the problem of defining significance by considering the larger question of importance.

Significance as Dichotomy

As noted earlier, the standard threshold for statistical significance has been maintained at ($p < .05$), and NCES has up to now adopted that fixed threshold approach to determining and reporting significance. However, even for a single variable considered in isolation (rarely the case in NCES reports) this creates problems for several reasons:

i) all report readers must accept this arbitrary threshold,
ii) all nuance of relative likelihood of divergence from the default is lost,
iii) the fixed threshold results in varying degrees of sensitivity to differences depending on strength of evidence (e.g., sample size) and inherent uncertainty,
iv) zero or essentially-zero difference from the default can never be "significant;" it can only be "default," and
v) asymmetry means the broad range of other "non-significant" values cannot be distinguished from zero.

Adopting a threshold approach can simplify writing – only a single statement at the beginning of a report is required – but it can also complicate understanding. For example, a "significant" effect observed for all students may be "non-significant" for all boys and also "non-significant" for all girls. If "non-significant" is taken as "accept the default." Logically, then if neither of the two groups differs from the default, how can the default be rejected, i.e. "a statistically significant effect" be found, when the groups are combined? Of course, this could be an example of Simpson's Paradox, which would require a deeper explanation. But most often this complication to statistical inference simply reflects the smaller sample sizes for the subpopulations of boys and of girls than for the total of the two.

The alternative of reporting a p-value provides a sliding scale indicator that addresses both shortcomings i) and ii) above. However, when standing alone, it still is predicated on the asymmetric notion of a default and the potential for a "significantly different" alternative to be established based on the data. And it is still subject to the tangled issues of magnitude of difference and strength of evidence.

As an illustration, findings may be differences between groups or differences from a default value (often but not necessarily zero). The importance of the information can then be the *absence* of a difference as well as the *existence* of a difference. For example, a report can include both the information that: "students who reported being the victim of any crime at school also reported being bullied at school at a higher rate than students who reported not being victims of crime," and also the information that: "no significant differences were found in the percentages of male students and female students who reported being the victim of any crime.[3]" But while the first statement demonstrates significant differences, the second statement includes a range of "non-significant differences" as well as zero, which the statement implies; and the breadth of that range depends on the strength of evidence and the uncertainty.

Measures of Magnitude

Magnitude of the effect of a variable or of the difference for a comparison has significance both with respect to "statistical significance" and with respect to importance as determined in the substantive context. On the one hand, an observed difference may be too small to be of interest, in which case the question of its reproducibility or "statistical significance" is moot. On the other hand, for a calculated difference of magnitude judged to be of substantive interest, the question of "statistical significance" turns on the precision of that difference estimate and on the strength of evidence.

Taking magnitude of effect as the starting point makes sense. Also, starting with a measure of magnitude/size of effect allows a symmetric approach that eliminates an artificial hypothesized default value in favor of an estimate of magnitude of effect. This resolves shortcoming iv).

When there is a prespecified hypothesis leading to a decision, the combination of magnitude plus p-value provides the two dimensions of information that are needed for discriminating between meaningful/negligible differences and at the same time comparing with the default (null hypothesis) likelihood of the observed value. When the inference is an estimate rather than a test (i.e., no decision to be made), magnitude is still required but a measure of its precision (typically its variance) replaces the p-value. By providing both estimated magnitude and precision, whatever the estimated value, statistical inference via estimation resolves shortcoming v).

NOTE: In this discussion the term "effect size" is avoided because in the social science and education literatures this is commonly applied to a normalized measure of magnitude. Also, for NCES reports, magnitudes in the original units would be preferable for most audiences.

[3] ***Student Victimization in U.S. Schools***: Results from the *2015 School Crime Supplement to the National Crime Victimization Survey*, STATS IN BRIEF, NCES, Draft June 2017.

Strength of Evidence

The latent component of significance is the strength of evidence, or its equivalent that connects sensitivity of available information to magnitude of effect. Given the available evidence, what is the attainable precision or the attainable sensitivity? This is at the heart of shortcoming iii).

In the case of a hypothesis to be tested, strength of evidence is often framed as the relationship of the necessary sample size to the magnitude of effect to be detected at some particular p-value. It can be deduced from a series of power curves for different sample sizes. It also can be shown (in a graph, for example) for different sample sizes as the relationships between the attainable p-value and the observed difference from the default.

When the observed magnitude of a difference is of interest in the substantive context, then the strength of evidence responds to the question of sufficiency of the data. It also provides a quantitative response to the question: How much more data would be required for an adequate strength of response?

In the case of estimation, the strength of evidence is more easily inferred from the confidence interval or from the empirical or fitted distribution of the data or of the summary statistic.

Completeness - Multiple Variables and Multiple Tests

The univariate case referred to above is simpler than the multiple variable case presented in NCES reports that give results for several variables as factors potentially influencing responses. Usually results are reported for multiple responses as well.

One aspect of richness in NCES data is that the data are national in scope and therefore reflect the heterogeneity of the population that leads to complexity of these data. Properly representing this complexity requires looking beyond single variables as influential to consideration their interactions.

With multiple variables come multiple tests or multiple inferences. In the case of research or statistical analysis reports, it is common in statistical practice to lay out the plan for analysis in the form of a specified set of hypotheses and the appropriate statistical tests. Essentially this analysis plan, structured to examine substantive theoretical constructs and conjectures, acts as a demonstrable commitment to report the results for all of the conjectures, whether significant or not.

When the analysis is not that simple (and it rarely is), either the pre-planned tests exceed the number of possible independent tests or multiple tests are planned using the same population or subset. In such cases measures of significance need to account for the dependencies. A Bonferroni adjustment is one method for preserving *overall* the desired level of statistical significance (i.e., $p < 0.05$ for the ensemble of all tests). As one alternative, the False Discovery Rate (FDR) is now commonly used in other contexts. Avoiding hypothesis testing altogether and reporting estimates with associated uncertainties instead of binary test results is another alternative among several.

Follow-on hypotheses, or simply further exploratory analysis, are distinguished in function from initially planned analyses also in being conditional on observing results from planned analyses. Consequently standard calculations for p-values for subsequent analyses will not be correct. These follow-on data analyses may already be anticipated but without a specific plan or may be fundamentally exploratory and spontaneous. Preparing for a two-stage analysis this way allows for first-stage testing of a "legitimate"[4] set of prespecified hypotheses with probability measures of significance at face value. For the second-stage analysis that deepens data exploration, the significance measure only takes the role of useful indicator of relative importance without a probability attached. What is important in a report is making a clear distinction between the interpretation of the significance measure as an actual probability measure or as a relative indicator.

Multiple Variable Fallacies

Two common fallacies – not endorsed by NCES Statistical Standards or Guidelines – are: i) to assume that factors influencing responses are independent of each other and ii) to assume that a set of hypothesis tests are independent provided the number of tests does not exceed the number of admissible (independent) tests possible.

Language from the past century used the term "dependent variable" to denote a response variable and the term "independent variable" to denote a factor potentially influencing response. This has been widely misconstrued to mean mutual independence among the "independent variables," which patently need not be, and oftentimes is not, the case. With interdependence, quoting "statistical significance" as measured for each of these variables independently is specifically not correct because tests for these interdependent variables cannot be independent. For example, socioeconomic status, father's education level and mother's education level are not independent, hence tests of the influence of these factors on a student's score cannot be independent either.

Often the influence of a set of factors is tested after partitioning the population according to yet another factor. Extending the hypothetical example above, subsets might be based on ethnicity or on degree of urbanicity. Clear understanding of the data may now, in fact, center on the interactions. Hence the effects averaged over the heterogenous population may not accurately depict the results for any one of the subsets. In this case, it is impossible for significance statements for any of the factors to be accurate without taking into account the interactions. Even for a lay report reader, the insertion of a two- three- or multi-way table into a purely descriptive report may convey the importance of the interaction, if not its full implications. A model-based analysis can accommodate this complexity and may be relatively easy to explain.

The restriction of the statistical approach to hypothesis testing also leads to the second fallacy. In some NCES reports, examination of changes in a series of observations is done by fixing on one observation as an index case and then comparing it to every other observation, one at a time. Examples include grade by grade comparison of results to 7th grade results or comparing every previous year to the current one. While there are N grades or N years and N-1 tests, these tests are not statistically independent when all rely on the same index case. The separately calculated probability measures of significance must be adjusted. What is additionally unfortunate is that the results are reduced to a series of individual questions intrinsically of less interest (e.g., 7v8, 7v9, 7v10, 7v11, 7v12, or 2017v2016, 2017v2015, etc.) rather than examination of logical breakdowns of the series (e.g., middle grades v high school) or of overall trends (e.g.,

[4] "legitimate" is used here to mean tests that are mutually independent with respect to the subsets of the population and that do not exceed the number of degrees of freedom available.

trend over past decade up to current year). (The panel noted that some instances of this kind of reporting are dictated by international agreement for international surveys.)

IV. IMPORTANCE RETHOUGHT

Significance vs Importance

When breaking down the term "statistical significance," "statistical" signifies probability and numbers, and "significance" is better characterized as attached to Information. The role of statistics and "statistical significance" is to illuminate the *importance* of the information.

The components of magnitude of effect, probabilistic interpretation (whether as threshold or p-value or precision measure or distribution) and strength of evidence (or sensitivity) are all needed to define significance. Significance is a three-dimensional concept: magnitude, associated credibility (including both likelihood and strength of evidence) and completeness.

The *starting point is magnitude of effect*, not relative rarity under a default condition. Inference about magnitude also requires a measure of precision or variance (which typically also figures in calculation of a p-value and/or strength of evidence). Completeness means representing the roles of correlations, interactions, dependence and complex relationships among variables – either factors or responses – and using the statistical methods that document these most clearly.

A binary classification does not serve either an unsophisticated public audience nor an audience that seeks deeper understanding. In addition, a dichotomy suggests an unjustified sharp distinction. Thus, significance is better represented as a full-spectrum concept, i.e., preferring p-values to a threshold. Indication of the strength of evidence is required as well in order to view the observed significance level against what is attainable with the available sample size and precision. A full-spectrum concept also implies using significance measures that are universally applicable, i.e., across all magnitudes including zero and near-zero values.

Intrinsic Importance

The potential value of information is ultimately – and originally – determined within a substantive context: Which information carries the greatest relevance and potential for insight or for decision making? For example, in the examining observations in a series, are the essential questions for policy makers: Responses for 7th graders v 11th graders?[5] OR Results for 2017 v results for 2011? OR are the more important questions those about the trend from 2007 – 2017 OR about comparison of middle grades results v high school results? OR some other question that could be answered better with a different statistical approach than a series of t-tests?

For NCES there is a perennial conundrum in reporting to an unsophisticated public audience while simultaneously providing sufficiently detailed information to policymakers at all levels from local to federal, private to public. Bypassing interactions among variables to capture the main factor effects might be

[5] ***Student Victimization in U.S. Schools***: Results from the *2015 School Crime Supplement to the National Crime Victimization Survey*, STATS IN BRIEF, NCES, Draft June 2017, p.2. "To assist policymakers, researchers, and practitioners in making informed decisions concerning crime in schools, the National Center for Education Statistics (NCES) collects data on student criminal victimization. . ."

effective in drawing attention of the general public to interesting features of the data. By itself, ignoring important relationships misses out on conveying valuable information that often could be communicated to a general audience without a lot of technical detail.

Making wise policy typically requires a more careful examination of the information than just quotation of main effects or marginal frequencies. Clear understanding depends on accounting for important interactions among explanatory factors, dependence among outcome variables, and often more complex relationships between factors and outcomes. Limiting statistical methodology to simple descriptive statistics fails to meet policy making needs. (The panel acknowledges that NCES makes public data very accessible to policy makers to take on the task of data analysis. However, the panel also recognizes that not all policy makers have the skill; even fewer have the time for such an analysis.)

A two-level report could mitigate the problem by providing a “broad strokes” report for the general public with a linked more comprehensive report including more detailed statistical analyses. Other solutions would also be possible using the technology available for web-based reports.

PART TWO

V. NCES CONTEXT

Publication and Audiences

Upon release, NCES public data files and accompanying reports are consulted by a diverse audience for different purposes and with attention focused on different parts of the data.

This means that NCES is presented with the usual issues intrinsic to reporting statistical analyses and findings: expressing uncertainty, definition of statistical significance, selectivity in reporting, reporting for multiple variables or tests. In addition, NCES faces two challenges because of the varied (statistical) sophistication of the audience. These are complexity of content and articulation of significance (as a statistical concept) in non-technical terms.

NCES reports, with some important exceptions, differ from research reports written for professional journals or other publications with readership from a specific discipline. Research publications typically cite specific hypotheses to be tested by data. In these cases, probability statements about significance can be formulated for a pre-determined set of inferences; and ancillary observations or data explorations can be clearly labelled to acknowledge that no meaningful probability can be assigned to indicate significance. NCES research reports, although rare in recent years, clearly fall into the category of reports written for professional publications.

In contrast, most reports currently produced by NCES serve to inform the public about available data or to highlight interesting observations about the data. These reports include *Data Points, First Looks,* and *Stats in Brief*. Since NCES data users are able to use the public data however they choose, no one or several factors or items in the list of response variables can be identified as necessarily primary or as only ancillary. Moreover, presenting every feasible partitioning of respondents into subsets is an impossibly-granulated task.

As described earlier, NCES has adopted as its all-purpose definition of “significance” as a classification: “significant”/ ”not significant” that is based on a single-variable test of zero effect or comparison of zero difference set at α=0.05, i.e., p-value less than 0.05. With some exceptions (notably international reports), no allowance is made for multiple tests, nor are any dependencies among variables noted. “Non-significant” results may not be reported at all. Many reports focus on “significant” effects and comparisons. For example, from the ECLS- K:2011 First Look: “All differences reported are statistically significant at the p < .05 level and are at least one-fifth of a standard deviation in size. Adjustments were not made for multiple comparisons.”[6]

One unfortunate consequence is the introduction of approximate language (e.g., “not measurably different”) to indicate results that are not well explained in terms of the threshold value. Often this arises when the sample size for a subset of the population is small. While the magnitude is sufficient to be of interest, the strength of evidence is not sufficient for meeting the threshold value. The use of vague terms that have no real definition in either a technical or non-technical sense should be replaced by clear statements of magnitude, precision and strength of evidence.

Publication Standards

At the present time the majority of reports are written by contractors without NCES staff listed as either author or co-author. To enable contractors to produce these NCES reports successfully and consistently across all contractors, NCES developed a set of standards that embody statistical principles and practices, accompanied by extensive guidelines for meeting each of these standards.

Standards were first published in 1992. In 2002 a comprehensive revision including substantial expansion of guidelines was released with a more recent updating in 2012. Chapter 5 specifically addresses the purpose of data analysis and reporting: “To ensure that statistical analyses, comparisons, and inferences included in NCES products are based on appropriate statistical procedures.” These standards and guidelines are available to all NCES staff, are provided to all contractors engaged in writing reports for NCES and are also published on the NCES website.

NCES originated a comprehensive set of clearly articulated statistical standards (Chapter 5 of *2012 NCES Statistical Standards*) for reporting survey and assessment data on education that has since served as the basis for some other federal data reporting standards. The detailed Guidelines that accompany these standards have proved to be a useful resource in writing NCES reports.

As contemporary best practices of statistics are changing, especially for reporting via electronic media, a revision of the 2012 is now due. Many of the Standards express statistical principles that are unchanged, but a few now need reconsideration as does a greater portion of the Guidelines. In Section 5.1, the first two points (5.11 and 5.12), outlining the need for an *a priori* analytic plan and the correct use of survey weights in computing summary statistics, are as important today as when they were first written. On the other hand, the two points (5.13 and 5.15) that deal with definition and representation of significance in terms of a fixed threshold (p < 0.05) need to be rethought and revised.

Addition of a new section should be considered to address standards for reporting statistical models and the proper treatment of significance in the presence of complex relationships. Within this section or

[6] ***ECLS-K:2011 First Look***: Findings from the *Third-Grade Round of the Early Childhood Longitudinal Study, Kindergarten Class of 2010-11.*

separately, standard and guidelines are needed for the measurement and reporting of model fit, including modeling with survey data, administrative data and data from heterogeneous sources.

The section on visualization (5.4: Tabular and Graphic Presentations) was written before web-based documents predominated. Hence, while correct, it focuses on the formatting of tables and simple graphs rather than the (statistical) information that modern visualizations can communicate most effectively.

New standards for visualization that are of the same high quality with guidelines of the same usefulness as in the rest of the Standards document are needed.

Because the Guidelines provide practical advice on proper representation of statistics in reports, these are the most subject to need for revision – especially for web documents. NCES has used “consider . . .” to reference statistical techniques that are called for in specific situations (e.g., multiple comparisons) but are not otherwise required. The flexibility of “consider . . .” also would allow for distinguishing levels of sophistication in a report where requirements for accessibility by a broad public differ from more technical requirements for decision-makers or researchers. On the other hand, this flexibility in application has also allowed these “considerations” to be bypassed without adequate rationale as manuscripts were being drafted.

Publication Practices

Within NCES the review process addresses content plus a technical review for compliance with the Statistical Standards. For “descriptive” reports that essentially quote from data tables, the dual focus is on accuracy of information and on potential for disclosure of identifying information. More comprehensive research reports or statistical analysis reports, may receive additional review that is more akin to a professional journal’s review – with commensurate delay.

For sample surveys, the foundation to much of NCES work, study units are randomly selected using a well-defined study design. This permits valid inference to the sampled population. However, even with a complex sample design, any given study cannot incorporate into the design (or apply at random) all the factors commonly considered to be of interest. Thus, for those factors associations but *not* cause-and-effect relationships can be determined using standard statistical analyses. Going beyond identification and characterization of associations to drawing cause-and-effect conclusions would require specific causal inference methods. (Note: As with any study, important factors that are not measured still have the potential to obscure relationships, but their impact may be missed when they are unobserved or are not investigated in the analysis.)

On occasion there has been confusion about what constitutes descriptive or summary analysis, what constitutes a comprehensive statistical analysis and what constitutes analysis of cause and effect. One consequence of this confusion is difficulty in obtaining approval for publication when manuscripts include a wider variety of statistical methodologies that are widely regarded as best practices and are commonly used at other federal statistical agencies. IF the author writes carefully when documenting patterns in the data, a reader or a reviewer should not infer causation. The breadth of statistical technologies applicable to NCES data continues to expand. Standard statistical analyses, generally accepted by federal statistical agencies as statistical best practices should not be called into question. However, when new statistical conceptualization or methodology lies outside a reviewer’s personal expertise, consultation should be obtained from a statistician who is expert in that area.

VI. REPORTING AND COMMUNICATING

Information – Brevity and True Picture

The tension between reporting in brief terms and giving a complete and clear picture of the significant information presents a serious challenge. For briefer reports, a more journalistic style works better than the format and style for research publications. By working from an initially proposed agenda, the primary issues can be presented at the outset together with a summary of the principal inferences. For the broad population, this assures that the "correct message" is taken even if the full report is not read with the details following this summary. For policy makers, an even deeper understanding may require either an extended publication (available on demand or via link) to further explain the significance of more complex relationships among variables and factors. Technology can facilitate the deliberate addressing of needs for both quick summaries and a more comprehensive description of relationships.

The transition from reporting a hypothesis test to more fully descriptive statistics can start by linking the observed size of effect to the range of conditions for which the observed data are consistent. To deter a naïve reader from concluding that non-significant means zero, one alternative is to present a range of "true" values, i.e., a set of values that could have produced the observed data with reasonable frequency. By way of example, the following explanation might be given in reporting on gender differences:

> "Previously, the score differences based on gender have been {reported or observed} to be as high as 4 points out of a possible score of 20. Based on the current study, the observed mean score difference of 2.2 would be consistent (i.e., at least a 1 in 20 chance of occurring) for any true mean difference between 1.6 and 2.8."

Report Structure and Language

Drafting an analytic plan for each report provides a basic structure for the later reporting of results. At the same time, this plan ensures that key findings are fully reported whatever their significance or p- values turn out to be. Even for a relatively brief report, some needs for information on more complex relationships among factors are predictable. For these relationships, variables can be identified in advance and used to pre-plan for a more comprehensive analysis. Starting from (rather than finally arriving at) this more complete understanding, also enables selection of the best and most accurate presentation of these results to the broad public, whether in simpler language or in graphics. While direct, lay language is important for addressing the broad public, accurate standard technical language is important for precise communication of this same information to the research community.

Just as different levels of sophistication are required to meet general and policy-decision needs, accessible language and even the best representation of the data may differ by audience needs. Well-considered [two- and multi-way] tables and graphics can often communicate results and relationships far more accessibly than text. Graphics presented in the exemplar reports given to the panel contribute little to overall understanding of the data and are not successful in communicating the importance of the information presented. Clearly this is an area where advances in technology and statistical graphics have a great deal to offer and have not been tapped.

Since NCES reports are now primarily disseminated via electronic media, it makes sense to take full advantage of the technology to link brief reports to more extensive ones, to connect important words (e.g., "p-value," "interaction") to definitions, possibly at several levels from general to technical, and to connect graphics to numerical values and inferences. This kind of linking can also simplify the language required in text by requiring only a single reference value. This obviates the need for cumbersome conventions requiring inclusion of long parenthetical asides for each cited statistic: test name, statistic value and p=value (e.g., t-test, two-tailed, t=1.875, df=100, p= 0.0637). However, this information would appear as part of the linked technical material.

VII. PRACTICAL CONSIDERATIONS

Processes

NCES Reports are primarily written by contractors. The NCES process for producing these reports draws on three strengths: first, NCES staff use their expertise to set topics for the reports; second, NCES makes extensive training available to contractors (via workshops and web-based tutorials); and third, the comprehensive statistical standards for reports that NCES has drafted with carefully written guidelines for their implementation are openly available and are required for reports by NCES contractors. For many reports, there are already templates and specifications or examples from previous studies. With a revised view of significance, the necessary updating and/or revision of these templates will not be a quick task and will require technical expertise to provide both language and examples.

When the oversight role of NCES staff only comes after the submission of a complete draft manuscript, the opportunity for substantial change is limited because additional funding would be needed for reanalyzing or redrafting. Requiring NCES approval of an analytic plan prior to the actual drafting of a manuscript could alleviate this. For example, it could ensure that the significance of both primary factors and potential interactions would be reported correctly. It would also allow NCES to ensure that all guidelines were being met and that decisions about implementing suggestions ("consider . . .") in the Guidelines were reviewed and documented prior to drafting the report. During the transition to a different representation of significance early staff input to the report development process would be particularly valuable.

Solutions and Transition

The conceptual transition is key: to start with magnitude in determining importance, and to use uncertainty and strength of information in gaging it. Then the dual transitions away from a threshold-based definition and beyond a single-variable approach will require committing effort, expertise and time to accomplish.

Procedurally, when important data features involve relationships or interactions among multiple variables and/or multiple factors, the details of the analysis methodology may warrant a more technical (statistical or substantive) review than usual. NCES could follow the common practice at other statistical agencies of soliciting these from external reviewers with appropriate expertise.

Use of technology may be successful in addressing some of the more difficult challenges. NCES has been able to take advantage of technology to pursue a number of ideas for NAEP. Implementation for other NCES data collections and other NCES reports will take more planning and more time to implement.

Conveying concepts, definitions and inferences to the multiple audiences for NCES reports might take advantages of linkages for definitions, layered from the most general to the technical. Similarly, summary analyses for the broad public could be linked to deeper and more detailed presentations for policy makers with perhaps a third layer giving model specifications that would communicate most efficiently with researchers.

Expanding the graphics using modern visualization methods for statistical information may communicate some concepts more accurately and more accessibly than text, even without interactive capabilities. This is one of the areas for growth that could benefit greatly from external technical consultants in the area of statistical graphics.

PART THREE

VIII. SUMMARY OF FINDINGS

Overview

The richness of the NCES data, its relevance for education policy decisions and its high-quality call for thoughtful and technically sound analyses so that the general public and policy makers as well as education researchers can make reliable and accurate inferences even when data are complex.

Survey designs for NCES data collections are carefully constructed bearing in mind the kinds of data summaries that will need to be generated and represent best practices for large-scale surveys and assessments. Scrutiny of collected data similarly follows best practices for ensuring the quality of the data files.

By comparison, statistical reports based on NCES data are currently limited in scope and also limited to only a subset of best statistical practices. The use of a threshold definition to define statistical significance is one impediment. The restriction to a very limited set of descriptive methods is another that is even greater. Both make communication with audiences at all levels of technical sophistication difficult.

Goals for Reporting

To meet the overarching goal of reducing the gap in information and in understanding between statisticians and policy makers and the lay public, NCES is encouraged to ensure that reports accurately reflect important complexities in the data.

Primary Recommendations: Significance and Importance

- Lead with magnitude of effect; follow with significance.
- Communicate importance in terms of magnitude and associated variance, probability (e.g., p- value, interval or other) and strength of evidence or sensitivity.
- Replace dichotomization and eliminate nebulous expressions (e.g., “substantially”). Dichotomization enforces an arbitrary single threshold on all data users and for all purposes. In addition, the threshold value may often be inaccurate, especially in the case of multiple tests or multiple comparisons.

Statistics and Methodology

- Expand the collection of analytic methods employed to meet the needs for analysis and interpretation. In particular univariate methods used alone can be seriously misleading because unidimensional analyses cannot reflect interactions, clustering or differences in responses among subsets of the population.

  NOTE: Multivariate analysis is often necessary for accurate interpretation of the data, but such an analysis does not imply causality.

- For multiple tests (or probability statements or intervals) indicate the required adjustments to calculated probabilities.

Planned and Exploratory Analyses

- Require an analytic plan at the outset that specifies analysis to be done and commits to full reporting of all planned analyses.
- Anticipate and allow exploratory analyses that are discretionary, but when reported are separated and clearly identified in the text, noting that any probability calculated cannot be correct without adjustment for conditional decisions and multiplicity.
- Report analysis details and process as supplemental material to provide technical support for interpretations.

Standards and Guidelines

- Review and revise (as needed) Standards and Guidelines every 3 to 5 years with attention to relevant advances in statistical and technological methodology. Start with an immediate comprehensive review.
- Add one or more new Standards (and accompanying Guidelines) in each of the following areas. Seek external consultants with specific expertise where appropriate.
  - Statistical graphics and data visualization
  - Measuring and reporting model fit for survey and administrative data
- Require that submission of reports for review include specific response to each Standard or Guideline indicating “consider. . .” with the decision to include or the rationale for not including the listed analysis as “planned.”

Publication Practices

- Write clearly but accurately so that information as interpreted by a broad readership will be consistent with deeper analyses of the data that support the reported results.
  - Analyses may be complicated to account for the survey design, for covariates, for clustering or for disparate subpopulations, but the explanations should be simply stated.
- Ensure complete publication of results for all planned statistical analyses and include statistical methods employed (especially tests!).
- Disseminate reports at two levels by providing details of analyses including analytic process and supporting statistical information. A more sophisticated reader or policymaker should be able to review the data analysis in deeper and more technical detail in order to validate methodology, results and conclusions, and to make decisions.

  - For example, consider expanding Data Point to supply deeper data analysis results by appending or linking to outline the data analysis process and to give supporting technical information.
- Indicate precision (and/or probability measure of significance) wherever data is presented – text, table, graph, other data visualization.
- Use technology wisely to link elaborations and detailed explanations, additional graphics or data visualizations, and important definitions to simple statements in online reports.

Note on Implementation

The expert panel recognizes that transitioning away from a threshold-based, single-variable-at-a-time concept of significance will require effort, expertise and time to accomplish. Attainable change will be a balance of feasibility in terms of resources (staff time, funding, etc.) with best practices. However, this does not change the urgency for moving forward.

APPENDICES

Appendix A: Charge to the Panel

Appendix B: Expert Panel Members’ Biosketches

Appendix A: Charge to the Panel

The charge to this panel of technical experts is to examine the representation of significance in NCES publications, deliberate the issues, make recommendations to NCES and, where appropriate, provide advice on the implementation of those recommendations. With pressure from some professional journals to impose more rigid practices, the findings and recommendations from this expert panel hold great importance to NCES in ensuring that all NCES publications (in-house or contracted) meet a high standard. Three particular concerns for the panel to address are:

- The black/white approach to defining "significant finding" (α-level test) vs descriptive measure of significance (p-value) vs confidence intervals vs other alternatives;
- The importance of reporting "non-significance" in order to recognize the lack of relationship especially in case of anticipated or suspected dependence or to indicate a continuing open question – implicit in this concern is the measurement and reporting of magnitude of effect;
- The dual problems of multiple tests and misstatements of significance when either there is direct or indirect dependence among tests that impacts significance statements or when multiple tests are done as part of data exploration but are considered to be and are reported as confirmatory.

The charge to the panel extends to two specific additional tasks. First, following the panel's recommendations, the panel is also asked to provide advice on how to effectively communicate the meaning of "significant" findings to the broad readership of NCES reports. Second, a further goal is to articulate some principles for use by NCES staff in preparing (or reviewing) reports, and to examine the current standards with recommendations for any changes in implementation.

Appendix B: Expert Panel Members' Biosketches

Michael L. Cohen, PhD

Title: Senior Program Officer for the Committee on National Statistics at the National Academies of Sciences, Engineering, and Medicine

Michael Cohen is currently serving as study director for the Standing Committee for Improving Motor Carrier Safety Measurement and for the Workshop on Transparency and Reproducibility in Federal Statistics. He is also assisting on the study on Reproducibility and Replicability in Science. Previously, he was a mathematical statistician at the Energy Information Administration, an assistant professor at the School of Public Affairs at the University of Maryland, and a visiting lecturer in statistics at Princeton University. His general area of interest is the use of statistics in public policy, with particular focus in census undercount, model validation, and robust estimation. He is a fellow of the American Statistical Association and an elected member of the International Statistical Institute. He received a B.S. in mathematics from the University of Michigan and an M.S. and Ph.D. in statistics from Stanford University. Finally, he has served as Associate Editor of the International Statistical Review and he is Editor of Statistics and Public Policy.

Jee-Seon Kim, PhD

Title: Professor in the Department of Educational Psychology at the University of Wisconsin-Madison

Dr. Kim received her BS and MS in Statistics and Ph.D. in Quantitative Psychology. Her research focuses on the development and application of statistical methods for addressing empirical questions in the social and behavioral sciences. Dr. Kim is particularly interested in experimental and quasi-experimental designs, multiple imputation for missing data, item response theory models, multilevel models and latent variable models, including methods for modeling change, learning, individual differences, and human development using longitudinal data. She has explored various advances and applications of these methods, including aspects of model development and testing. Dr. Kim has participated in numerous research projects funded by different agencies including National Institutes of Health, Institute of Education Sciences, National Science Foundation, and the Agency for Healthcare Research and Quality, and has extensive experience in study design, data management, analysis, and dissemination. Dr. Kim currently serves as an associate editor for Psychological Methods and Psychometrika.

Finbarr "Barry" Sloane*,* PhD

Title: Program Director in the Knowledge Building Cluster (EHR/DRL), Building Community and Capacity in Data Intensive Research in Education (BCC-EHR), Division of Research on Learning in Formal and Informal Settings (EHR/DRL) at the National Science Foundation

Dr. Sloane, a native of Ireland, received his PhD in Measurement, Evaluation, and Statistical Analysis from the University of Chicago with specialization in Mathematics Education and Multilevel Modeling. His research has appeared in Educational Researcher, Reading Research Quarterly, and Theory into Practice; he serves on the editorial boards of a number of journals including: Irish Educational Studies, Mathematical Thinking and Learning, and Reading Research Quarterly. Prior to accepting the appointment as Program Director at NSF, he was on the faculty at Arizona State University, College of Education.

Linda J. Young, PhD

Title: Chief Mathematical Statistician & Director of Research and Development, USDA's National Agricultural Statistics Service

Linda J. Young is Chief Mathematical Statistician and Director of Research and Development of USDA's National Agricultural Statistics Service. She oversees efforts to continually improve the methodology underpinning the Agency's collection and dissemination of data on every facet of U.S. agriculture. Prior to joining NASS, Dr. Young served on the faculties of three land grant universities: Oklahoma State University, University of Nebraska, and the University of Florida. She has three books and more than 100 publications in over 50 different journals, constituting a mixture of statistics and subject-matter journals. A major component of her work has been collaborative with researchers in the agricultural, ecological, and environmental sciences. She has been the editor of the Journal of Agricultural, Biological and Environmental Statistics. Dr. Young has served in a broad range of offices within the professional statistical societies, including President of the Eastern North American Region of the International Biometric Society, Vice-President of the American Statistical Association, Chair of the Committee of Presidents of Statistical Societies, and member of the National Institute of Statistical Science's Board of Directors. Dr. Young is a fellow of the American Statistical Association (ASA), a fellow of the American Association for the Advancement of Science (AAAS), and an elected member of the International Statistical Institute (ISI).

Panel convened by National Institute of Statistical Sciences

Nell Sedransk*, PhD*

Title: Director, National Institute of Statistical Sciences-DC

Dr. Nell Sedransk is the Director of the National Institute of Statistical Sciences. She is an Elected Member of the International Statistical Institute, also Elected Fellow of the American Statistical Association. She is coauthor of three technical books; and her research in both statistical theory and application appears in more than 60 scientific papers in refereed journals. The areas of her technical expertise include: design of complex experiments, Bayesian inference, spatial statistics and topological foundations for statistical theory. She has applied her expertise in statistical design and analysis of complex experiments and observational studies to a wide range of applications from physiology and medicine to engineering and sensors to social science applications in multi-observer scoring to ethical designs for clinical trials.