Institute of Educational Statistics
National Center for Education Statistics

National Institute of Statistical Sciences
Technical Expert Panel Report

CITING SIGNIFICANCE IN NCES DATA REPORTING

National Institute of Statistical Sciences

January 2019, updated March 2019

TABLE OF CONTENTS

NATIONAL INSTITUTE OF STATISTICAL SCIENCES

CITING SIGNIFICANCE IN NCES DATA REPORTING

EXECUTIVE SUMMARY

The vigorous national dialogue about how to describe the “significance” of research findings that an NCES-NISS Expert Panel addressed in 2018 has shifted. No longer is it about whether to move away from dichotomizing results into “significant” or “non-significant” (e.g., $p < 0.05$) but rather about how to do it and what information must now be supplied. That earlier panel determined that whether findings are reported as a data summary or as in-depth analyses, their importance must reflect both the magnitude and the associated uncertainty. In addition, interpretation depends on the subject matter context and the intended use of the information.

Therefore the specific charge to the current (second) panel was to recommend how to report the importance of results clearly and accurately in a manner that moves away from the p-value < 0.05 standard but is understandable to the public and acceptable to academic institutions. The panel responded that substituting alternative language could not be sufficient because the significance of a research finding must be judged in its substantive context. Magnitude and uncertainty address the questions: “What is the best estimate? What are the possible alternatives and how likely are they?” But the driving question is substantive: How much do any of these matter (in a substantive sense)?

The overall objective is clear, accurate, complete and transparent reporting of findings from NCES data. Accomplishing this requires providing more complete information (magnitude and uncertainty, at least) than a threshold.

Publication of NCES reports is unique in two important aspects that present specific challenges for NCES data reports. The first challenge arises from the breadth of the NCES readership. The second challenge for NCES reports is to make additional, more detailed information available about some of these complexities based on the finer scale data that NCES maintains in restricted files. Meeting these challenges requires accessible reporting that is credible at multiple technical levels from non-technical for the general public to technically clear for researchers in academia and outside.

Primary Recommendations

- Lead with *Magnitude* and its associated *Uncertainty*.
- Represent both magnitude and uncertainty everywhere, in every format: text, table, graph, figure, other visualization.
- Formulate an analysis plan in substantive terms with a narrative that drives selection of variables and factors selected, defines populations included/excluded.
- Support the analysis plan with appropriate statistical approach and methods.

- Publish complete results from all analyses corresponding to the analysis plan and give transparent access to statistical analysis process (since there is no direct access to restricted data for confirmation).
- Distinguish secondary and exploratory analyses clearly from planned analyses, also noting that uncertainty measures and p-values cannot be assumed to apply accurately.
- Present reports with equal depth and equal clarity in non-technical and technical language, with links to the underlying statistical analyses to permit validation.

Specific Recommendations for Implementation

- Revise NCES Standards and Guidelines to update and expand statistical methodology and to include modern data visualization.
- Educate NCES Staff and Contractors.
- Involve substantive and statistical experts in analysis plan and in report development.
- Align Review Process to new Standards and Guidelines.

PREFACE

In 2019, the National Center for Education Statistics (NCES) charged the National Institute of Statistical Sciences (NISS) with recommending how to report statistical significance clearly and accurately for a general audience as well as for more technically oriented readers. Whether findings are reported as a data summary or as in-depth analyses, their importance must reflect both the magnitude and the associated uncertainty, although interpretation depends on the subject matter context and the intended use of the information. This panel focused on changing report guidelines to meet this objective.

On 25-26 January 2019, the panel of technical experts met with NCES staff, followed by closed deliberation. The panel held further teleconferences during the preparation of this report.

NATIONAL INSTITUTE OF STATISTICAL SCIENCES TECHNICAL EXPERT PANEL REPORT

CITING SIGNIFICANCE IN NCES DATA REPORTING

I. BACKGROUND[1]

Perspective on Significance

There is a vigorous national dialogue across research disciplines about how to describe the “significance” of research findings. For research publication over the past century, significance has increasingly devolved into a probability statement labeled “statistically significant, p<0.05.”

What threshold determines a finding to be worthy of consideration?

The core of this question that has been pondered starting well before 1925, when Sir Ronald Fisher’s oft-cited book on statistical methods was published.

> *Even in the 19th century, we find people such as Francis Edgeworth taking values `like’ 5%—namely 1%, 3.25%, or 7%—as a criterion for how firm evidence should be, before considering a matter seriously.*[2]

The discussion has broadened and deepened, widely engaging scientific disciplines including biological and clinical sciences, social sciences, physical sciences and engineering as well as mathematical, computational and statistical sciences.

Three publications in particular in noted professional journals have addressed this topic recently, articulating broadly held views on significance and the use of p-values in such contexts.

There are proponents for simply changing to a more stringent threshold, as noted in a September 2017 article in *Nature Human Behaviour*:

> *We propose to change the default P-value threshold for statistical significance from 0.05 to 0.005 for claims of new discoveries.*[3]

In 2017, The American Statistical Association (ASA) opened discussion on statistical significance with the 2017 *Symposium on Statistical Inference*.

In September 2018 the first panel of technical experts, convened at request of the National Center for Education Statistics (NCES), focused on how significance of findings from data currently are and how they should be reported in NCES publications and presentations of data summaries. In February 2019 NCES

[1] BACKGROUND: Section updated March 2019

[2] Stigler, S., CHANCE 21:4, 2008: DOI: 10.1007/s00144-008-0033-3

[3] Benjamin, D.J., Berger, J.O., Johannesson, M. et al., Redefine statistical significance. *Nature Human Behaviour*. 2, 6–10 (2018) DOI:10.1038/s41562-017-0189-z.

convened the second panel of technical experts to consider the implementation of the recommendations for change made by the first panel.

In March 2019 the ASA discussion of p-values and alternative approaches culminated in a comprehensive treatment of the whole question of "significance" published in a special issue of *The American Statistician*.[4] In that same month Nature published a commentary with over 800 scientist signatories, with a different proposal: abandoning the threshold definition of statistical significance in favor of reporting findings with explicit statements of findings and their associated uncertainties or intervals.

> *. . . in line with many others over the decades, we are calling for a stop to the use of P values in the conventional, dichotomous way - to decide whether a result refutes or supports a scientific hypothesis.*[5]

The commentary then addresses the nature of (scientific) thought when premised on a dichotomy.

> *The trouble is human and cognitive more than it is statistical: bucketing results into 'statistically significant' and 'statistically non-significant' makes people think that the items assigned in that way are categorically different. The same problems are likely to arise under any proposed statistical alternative that involves dichotomization, whether frequentist, Bayesian or otherwise.*
>
> *Unfortunately, the false belief that crossing the threshold of statistical significance is enough to show that a result is 'real' has led scientists and journal editors to privilege such results, thereby distorting the literature. . .. discussion that focuses on estimates chosen for their significance will be biased.*[3]

This article continues by observing with disapproval the consequences for research of using dichotomization as a universal criterion.

> *. . . rigid focus on statistical significance encourages researchers to choose data and methods that yield statistical significance for some desired (or simply publishable) result, or that yield statistical non-significance for an undesired result, such as potential side effects of drugs - thereby invalidating conclusions.*[3]

NCES expert panels and these researchers concur in the essentials about significance of research findings.

Thus the national dialogue is shifting. No longer is it about whether to move away from dichotomizing results into "significant" or "non-significant," but rather about how to do it and what information must now be supplied.

Charge to Panel

The specific charge to this second panel was to develop recommended language for describing the significance and magnitude of findings in a manner that moves away from the p-value < 0.05 standard, but is understandable to the public and acceptable to academic institutions.

[4] The American Statistician, Vol 73, 2019: https://tandfonline.com/toc/utas20/73/sup1

[5] *Nature* **567**, 305-307 (2019) DOI: 10.1038/d41586-019-00857-9

The panel initially responded that substituting alternative language would be insufficient and that the significance of a research finding must be judged in its substantive context. This could best be accomplished by providing more complete information (magnitude and uncertainty, at least) than a threshold

II. CONTEXT

Findings from Panel on Significance and Non-Significance in NCES Reports

The 2018 panel's chief recommendations were first, to replace "significant" or "non-significant" with the magnitude of the findings and the uncertainty associated, so that importance could be interpreted substantively and confidence could be expressed quantitatively. Second, findings should be obtained by statistical best practices and not be limited to specific methods of either analysis or presentation. Third, reporting should be comprehensive for all planned analyses, rather than determined by a threshold for "statistical significance." The previous panel also considered the role of an analysis plan in determining which findings should be published.

The present panel was charged with considering how to implement these recommendations.

NCES Data Reports – Multiple Levels of Reports

Publication of NCES reports is unique in two important aspects that present specific challenges for NCES data reports. The first challenge arises from the breadth of the NCES readership which spans the full spectrum from the general public to the research community with wide-ranging interests. The NCES readership is equally broad in both its interests and its levels of quantitative skill. This does not imply that non-quantitatively minded readers are either uninterested or unprepared to deal with conceptual complexities in NCES data. Therefore, successful presentation of complicated relationships cannot rely solely on technical language and tools that would be natural in communicating with quantitative researchers.

The second challenge for NCES reports is to make additional, more detailed information available about some of these complexities based on the finer scale data that NCES maintains in restricted files. NCES does provide open access to the public use data that summarize finer scale information that is privacy-protected. These public data are necessarily limited to high-level aggregation and are not intended to probe complex relationships evident in the finer scale data. NCES reports and other publications based on NCES-permitted access to restricted files bridge this gap by providing to the public analyses that could only be done using data in the restricted files. Together, both aspects present a cogent argument for multiple levels of communication of each report. At the same time, NCES reports are not unique in comparison to research papers, technical publications, white papers and other documents that must meet a high standard for clarity, accuracy, and completeness of communication.

Scientific journals require that findings be verifiable, from data collection through interpretation, and some journals also require that the original data be accessible. Since NCES reports are based on unavailable restricted data, researchers cannot reproduce or confirm some analyses. Restrictions limit verification to examination of the process and intermediate computations that support the inferences and their interpretation.

Thus accessibility of information in NCES reports across the spectrum of quantitative skill across the NCES readership calls for both technical and non-technical presentations. The need for verifiability calls for linking reports to complete technical documentation of methodology and intermediate results.

III. DEFINING AND CONVEYING IMPORTANCE

Substantive Meaning

Even in the 19th century, discussion revolved around a criterion for the evidence needed "before considering a matter seriously." In the 21st century, the technological tools and enormity of the data often mean that allowing for random variation is not the primary issue. Considering a matter "seriously" still depends on the substantive implications of the evidence. These in turn depend upon i) the best estimate from the data and the uncertainty associated with this estimate, and upon ii) interest in the likely substantive implications.

Magnitude and Uncertainty

Magnitude and uncertainty address the questions: "What is the best estimate? What are the possible alternatives and how likely are they?" But the driving question is substantive: How much do any of these matter (in a substantive sense)?

Magnitude is a straightforward concept, and measurement is most often a point estimate, whether of a quantity, of a difference, or of a model parameter. In the case of a comparison, an inference, or a particular line of theory/inquiry, it is often possible (and useful!) to postulate in advance or even roughly approximate, the minimal magnitude 'worthy of serious consideration.' Note that for NCES reports this minimal magnitude most likely varies considerably among uses for and users of the information. In addition a single item may have more than one representation (e.g., global or average value for a population vs numbers of individuals or groups affected). Reporting actual values and their uncertainties allows users to make their own determinations.

Magnitude may also be a measure of the relative contribution of a factor to the overall uncertainty. For example, the simply calculated variation for a variable of interest may be partitioned into fractions that are assignable to each explanatory variable and to their interactions and a residual term (the "error variance").

Communicating uncertainty can be challenging because it most often arises from more than one source. Also the clearest way to present uncertainty can vary depending on circumstances and methodology. One source of uncertainty is random variation (random sampling); variation in measurement is a second source; adequacy of a model to fit the data is a third.

Part of the initial plan for a data collection or the outline for undertaking a summarization or analysis of data is determined by the precision needed for inference from the data, whether for an estimate to be valuable as an accurate descriptor or as the basis for an inference, a comparison, or a decision. For data collection, determining precision is an explicit part of the design process. Otherwise, as for secondary analysis, the attainable precision may be a consequence of the data available for analysis and may turn out to be either excessive or insufficient for the desired inference.

The "best way" to express uncertainty will depend on the methodology used to estimate magnitude. Whether uncertainty is expressed as standard error, coefficient of variation, log odds, Bayes' posterior distribution, AIC, R2 or some other quantity, the concept is the same: The likelihood of the data is greatest ("most compatible") at the "best estimate" and lower for alternative values. And the farther away an alternative value is from the best estimate, the lower the likelihood of the data.

When the inferential purpose is comparison, decision, or modeling (such as a time trend), importance depends on the relationship between magnitude and uncertainty. Essentially, the uncertainty defines the precision of measurement of the difference (or parameter estimate), and, in consequence, determines the sensitivity to any specified magnitude of the difference (or model parameter).

So the important message is: Convey both magnitude and uncertainty to discuss data meaningfully in clear, quantitative terms. The importance of the magnitude depends on the substantive context.

Indeterminate Language

The practice of dichotomizing data into statistically significant ($p<0.05$) or non-significant leads to a struggle for accuracy when a p-value lies close to the threshold (especially when it exceeds the threshold only very slightly). This often occurs when the sample size is insufficient for higher precision. The opposite is equally difficult, when the sheer volume of data yields such a high precision that everything is significant. A common response to either of these scenarios is to hedge descriptions of findings by employing non-technical, approximate, and imprecise, terminology.

Unfortunately this obscures rather than clarifies because the approximate expressions are used at the will of the writer and have no precise meaning. Citing the magnitude and the uncertainty provides more information and at the same time alleviates the problem of over-simplifying results. Standard statistical graphics (boxplots, for instance) can provide easily accessible displays of the spread of the data, whether for description or for comparison of subsets.

Reversion to non-specific language can also occur when the statistical analysis is incomplete and does not portray relationships adequately. Hence, findings are reported in common language that is not a re-expression of a term with technical meaning, making the credibility of the interpretation.

IV. ANALYSIS GOALS

Characterization of Goals

Determining importance of findings begins with the statement of the analysis goals. In the social and educational sciences, common practice is to write these as research questions that usually encompass the first line of analysis.

The research questions "Is . . .?" and "Does . . .?" beg for binary responses: "Yes" or "" No." Research questions "How . . .?" or "To what extent . . .? " ask for information that cannot be supplied with just a threshold test, which is applicable only to simplistic yes or no questions. The questions "Is . . .?" and "Does . . .?" may assure a "statistically significant ($p<0.05$)" finding but often these are posed when the answer is a foregone conclusion. For example, "Is there a difference in reading comprehension scores for 4th graders between English Language Learners (ELL) and other students (native or fluent English speakers)?"

The questions that expand substantive knowledge also explore differences in meaningful directions. To continue the example, “How big is the gap in reading comprehension between 4th grade ELL students (at different ELL levels) and non-ELL students?” “How different are the reading comprehension deficits for different native languages?” “How much does reading comprehension lead or lag mathematics based on ELL level?” “Which demographic and socioeconomic factors are most important in determining the degree of reading comprehension deficit for a 4th grade ELL student?” “What distinguishes higher-scoring 4th grade ELL students from lower-scoring ELL students at the same ELL level?”

Data analysis is undertaken to expand substantive knowledge; characterization of the analysis goals outlines the direction of that expansion and the nature of the information that is being sought. Formal analysis planning, with expert review, also reduces the opportunity for “p-hacking” –indiscriminate searching for any pattern that will pass the test of “statistically significant,” to persuade a journal editor to publish a manuscript.

Survey, Assessment and Administrative Data

Goals for sample-based data should identify and quantify the sources of uncertainty for estimates of effects and estimates of relationships among variables in the data. These include the randomness attributable to sampling, contributions from non-response and other adjustments, and instrument imprecision.

The term “administrative data” is used here for complete population records (a census). Although administrative data are subject to recording errors, these errors are generally not traceable or estimable. Analysis, therefore, consists either in simply reporting the data, (aggregated or not) or in modeling patterns that characterize these data. Uncertainty indicates how well the modeled patterns describe the data and is generally calculated from the residuals from fit (differences between model and data values).

For extremely large databases, the ordinary measures of uncertainty are often negligible. This poses the obvious problem for hypothesis testing or for calculating p-values; “everything is significant, whether it is of any importance or not.” One approach to inference is to describe patterns, as would be done for administrative data. However, it is also appropriate to question the homogeneity of the total population represented in the data and to examine the factors or patterns that illuminate any heterogeneity.

Statistical Basis

To achieve the substantive goals, a statistical plan is needed as well, at a minimum for the first line of research questions. The shape of the research inquiry together with the data properties will dictate the statistical tools of choice. These tools will, in turn, define how magnitude and uncertainty are expressed.

Even if specified only after the first line of research questions, the planning is the same for statistical analysis in the second line of inquiries. While it is not necessary - or even desirable – to plan in advance, it is necessary that the process be recounted as part of the final publication. Understanding the logic of the complete data analysis is essential to validating the inferences, and with restricted data, this may be the only validation possible.

Much NCES data are multivariate in nature and are only accurately interpreted when analysis methods are multidimensional and/or multivariate. This means that NCES data are a rich source of information about relationships among variables, both outcomes measures and covariates.

Proper analysis respects this complexity. However the lack of independence among multiple outcomes, factors, indicators, or covariates can challenge researchers. Multidimensional (or multivariate) statistical methods are called for when outcome measures are not independent of each other. Similarly, multivariate methods are needed to accommodate dependencies among measures of factors or covariates.

A second challenge is heterogeneity within the data. Beyond the aggregate analysis, important effects often appear only when population subsets are examined. The challenge of analysis is not to miss something important because of differential effects among definable subgroups.

V. ANALYSIS PLAN THROUGH PUBLICATION

Substantive Objective and Commitment to Publish

At the heart of this report is the goal of increasing substantive knowledge and substantiating new information by clearly stating findings together with estimates of the associated uncertainty.

The analysis needs to be grounded in the substantive context and needs to include the rationale for variable and factor selection, as well as to define the population or subpopulations included and excluded. A clear initial explanation of the rationale and approach also minimizes post hoc variable or factor selection

Implicit in this analysis plan is a commitment to publish all findings prescribed at the outset whether these are predictable or surprising and regardless of their magnitudes.

Initially Designed Analyses

Precision as well as determination of essential data elements figures into the initial design of every NCES survey or assessment. Design requirements are set and decisions are made starting with those selected data elements and the precision that is required and/or is attainable. Plans for reporting results is either an integral part (as specified for the data collection or for standard compendia of results) or a natural extension of inquiries predicted in advance.

By their formulation, all initial analyses are of substantive interest. Full reporting means including all results, whether remarkable or not, while always protecting privacy and personal information.

Subsequent and Secondary Analyses

Subsequent and secondary analyses are of several types. One includes analyses that are prompted by results from the initial planned analyses but are only planned subsequently. Another includes independent secondary analyses prompted by a researcher’s desire to pursue an earlier conjecture or theoretical proposition. Finally, researchers may conduct exploratory analyses with the goal of developing new conjectures. All these types of analyses constitute meaningful research, but the distinctions among them are important.

An initial planned analysis is defined by a narrative that drives the selection of variables and factors, the populations included and excluded, the relative importance of findings, and the statistical approach. All planned results are fully reported. A subsequent analysis differs from an initially planned analysis in that the narrative of a subsequent analysis indicates which results from the initial analysis motivated the selection of the secondary topic that was investigated. So the results of the secondary analysis are given in context with the primary analysis results. The conditionality of these secondary analyses means that probabilities calculated under (the usual) assumption of independence will not be accurate or appropriate.

Exploratory analyses (as distinct from "p-hacking") are freely driven by pattern discovery. These analyses provide conjectures for further investigation. The freedom of the researcher-explorer to proceed intuitively means that the meaning of calculated uncertainties is lost. Indicators of uncertainty may be useful strictly within the exploration, but they may also be misleading. While exploratory analyses may lead to provocative discussion, results must be clearly designated as explorations.

To be clear, "p-hacking" or searching for comparisons or patterns to satisfy a p-value threshold does not constitute exploratory analysis. Rather a researcher-explorer seeks to develop a conjecture or a narrative that reflects both evidence (data) and a rationale or theoretical premise.

VI. ILLUSTRATIONS FOR PARTICULAR CASES

Metrics

Some of the possible metrics for magnitude and for uncertainty are given here. While outside the charge to this panel, in many cases, the addition of well-chosen graphics often enhances understanding and may provide a more accessible description of results than text explanation alone.

Magnitude is most commonly reported as a point estimate, often an average or a median for a quantity or a difference. However, it may be the correlation between response variables, the percentage of a population, the rate of a trend, or another parameter of a model.

Magnitude may even be the set of values for specific population subsets or for a collection of interventions or other partitions of the information. On some occasions it is the distribution of values for the population studied.

Making a metric for uncertainty understandable is a greater challenge. For point estimates, one alternative is to define a likelihood-based interval. Terminology currently in vogue is "interval of compatibility," often shortened to "compatible interval" to encompass frequentist, Bayesian, fiducial and empirical intervals surrounding the point estimate.

In all these cases the likelihood is not equal across the interval, and this requires explanation. For example, "The likelihood for these data is greatest at the estimated value for . . . Other values could also give rise to these data, but the likelihood decreases with distance from the estimated value. At the ends of the interval the likelihood is only {'1 in 40' or 'one-tenth of the likelihood at the estimated value' - depending on the kind of interval}." Other possibilities include likelihood ratios, odds ratios (more intuitive than log odds), interquartile ranges ("results for half the population lie in the interval") and other intervals defined by quantiles.

Uncertainty for results expressed in percentages of individuals can be expressed as "margins of error," a term that is familiar to most of the population.

Parameter estimates have associated uncertainties. For models and for multi-way tables these include uncertainty about the model itself or about the relationships in a multi-way table. Goodness-of-fit is a metric for these kinds of uncertainty, and it calculated from the residuals from fit (defined by the differences between the model predictions and the actual data values). Variation accounted for is another measure of fit (R2) that is easily communicated (defined by the proportion of total variation that is assignable/explained by the model). Other scaled criteria that quantitate the amount of information accounted for by a model (AIC, BIC) are also commonly used indicators.

Different sample sizes can cause different issues. Three particular cases deserve special note. The first occurs with an enormous sample size. The problem – "Everything is significant" – arises because calculations of uncertainty lead to only tiny variation for an aggregate measure, such as a mean or a parameter value for a fitted model. In other words, the apparent precision gives sensitivity to distinctions far smaller than are of interest substantively. One choice is to treat the data similarly as for a census, and to provide the data distribution in conjunction with any aggregate estimates. An essential question in this case is whether the homogeneity assumption that underlies the calculation of the aggregate measure and its uncertainty holds true. This leads to subset analyses and decomposition of uncertainty into within-subset and between-subset components.

An equally problematic incompatibility of precision/sensitivity and magnitude of substantive interest occurs with small sample sizes. When the sample size is insufficient to attain desired precision, substantively interesting differences or values will not be significant. Consequently this case has plagued the threshold p-value approach to inference, and it occurs frequently when population subsets are small. Reporting magnitude and uncertainty makes understanding easier, however researchers also need to indicate what sensitivity is attainable and to note explicitly, for example, that "sample size acts as a limiting factor for the width of an interval."

The third case is administrative data that constitute a census since there is no sampling or random error on which to base uncertainty. Reporting is either a descriptive summary, the empirical data distribution itself, or a model of the patterns in the data. In the last case goodness-of-fit measures are based on the discrepancies of the data from the generalizations presented by the model.

p-Values

Another role for a p-value is as a (normalized) sliding-scale indicator of the distance between the data-based estimate and the postulated default (fixed-point) value or model-predicted value. In this way the p-value allows interpretation of comparative likelihoods based on the best estimate from the data and based on the default. The p-value does not fall afoul of making unreasonable distinctions (e.g., p=0.052 compared to p=0.048) when used as an indicator rather than a threshold determination of not significant vs significant.

Using a p-value makes sense when the default is bona fide. For example, in an analysis of variance to examine the relative contributions of covariates or factors to a model, the bona fide default is "only contributes noise." Similarly, the role of the p-value is to evaluate factors for inclusion in the model based on the information each factor contributes.

However, a p-value cannot by itself determine importance. Because it is normalized, the sensitivity of a p-value as an indicator depends on the normalizing coefficient which in turn is a function of sample size. Thus, an extraordinarily large sample will result in an indicator that is sensitive on a scale having little substantive interest, whereas a small sample will not reflect the minimal magnitude of substantive interest.

Therefore, the magnitude and the uncertainty are still essential to interpretation of the p-value in a substantive context.

VII. ACCOMPLISHING CHANGE: NCES CONTEXT

The issue at hand: How to communicate with clarity and accuracy the importance of a quantitative finding and in particular the magnitude and the implications of the precision with which the magnitude is measured or assessed. The importance of communicating the concepts of important and reproducible findings has grown in importance with the broadening audience for NCES data, statistical issues from big data, and increasing interest in results for small-to-very small data (sub)sets.

Probability-based “significance” originated more than a century ago in the need for a systematic basis for decision-making. Over time, the terms “significant” and a fortiori (non-significant) have expanded into other contexts where this binary partitioning is not only unnecessary, it also fails to convey important information.

The NCES context: The NCES databases are rich and deep. Many are long-term or have been serially collected over long periods of time, and they are high quality and thoroughly documented. As an information resource on education data, these data are unparalleled.

NCES data reporting is directed toward diverse communities that include both technical and lay audiences. The challenge is to communicate clearly and credibly to all, recognizing that both language and extent of technical detail need to differ.

NCES has been a leader among the federal statistical agencies, with the development of NCES Statistical Standards supported by Guidelines that cover implementation from sample design through data analysis. However, advances in statistical methodology, computational capacity, and information technology all underscore the need to update these standards and many of the Guidelines. The same statistical principles and issues for meaningful data reporting in the original Standards and Guidelines need to be addressed in a comprehensive revision.

NCES now seeks to make information more broadly available, not only from its public data resources, but especially expanding dissemination to the general public of results of reports based on analysis of restricted-access data. To facilitate this beyond NCES, IES is determined to make changes across IES.

NCES is equipped to innovate and to implement change. NCES has already embraced electronic and communication technologies to reach its many constituencies. NAEP funding and priorities have enabled investment in architecture and tools for graphical and interactive web delivery of information and for innovation in data visualization. In addition, NCES has invested careful thought in the design of each data collection, especially taking into account populations and factors of interest.

NCES has long relied on contractors for development as well as production work; contractors often author NCES reports. They, too, must engage in change. This includes roles for senior substantive and statistical experts in approving analysis plans, statistical methodology, inferences, and interpretations. Based on the long-standing relationship with contractors, NCES has acquired experience in training both staff and contractors, including in-person, webinar, video on demand, as well as traditional written materials.

The Challenge: Accomplishing change will require both commitment and time to implement, but NCES has the expertise to begin.

VIII. SUMMARY OF FINDINGS

Foundation for Recommendations

Objective: Clear, Accurate, Complete and Transparent reporting of findings from NCES data.

Requirement: Accessible reporting that is credible at multiple technical levels from non-technical for the general public to technically clear for researchers in academia and outside.

Principal Recommendations

- Lead with Magnitude and its associated Uncertainty.
- Represent both magnitude and uncertainty everywhere, in every format: text, table, graph, figure, other visualization.
- Support the analysis plan with appropriate statistical approach and methods.
- Publish complete results from all analyses corresponding to the analysis plan.
- Present reports with equal depth and equal clarity in non-technical and technical language, with links to the underlying statistical analyses to permit validation.

Language

- Abandon the terms "significant" and "significance."
- Eliminate vague terms in favor of quantitative, precise statements.
- Instead of tests of hypotheses, present comparisons in terms of magnitudes of differences and uncertainty.

Principal Elements

- Magnitude
  - Defined as what is meaningful and "worthy of serious consideration."
  - Estimates, differences, relative contributions of factors to outcomes or model.

- Uncertainty
  - Essential to understanding magnitude
  - Multiple possible presentations, preferably relating likelihood to the degree of closeness to stated magnitude value
- Statistical Basis
  - Statistical methodology/analysis) supports (substantive) analysis plan and follow-on analyses indicated by initial analysis.
    - Correct uncertainty calculation depends upon appropriate statistical methodology/analysis.

    - Much information in the NCES data files is multivariate in nature and is only accurately reflected when analysis methods are multidimensional and/or multivariate.
  - Transparency requires access to statistical analysis (since there is no access to restricted data for confirmation).
- Analysis Plan
  - Formulated in substantive terms with a narrative that drives selection of variables and factors selected, defines populations included/excluded, and aligns statistical approach to rationale
  - Complete reporting of all analyses planned at the outset as part of the survey/assessment design or planned subsequent analyses
  - Subsequent reports of secondary analyses with complete (subsequent) analysis plans
  - Clear indication of exploratory nature in reporting exploratory analyses., also noting that uncertainty measures cannot be assumed to apply accurately.

Implementation Requirements and Specific Recommendations for NCES

- Revise Standards and Guidelines
  - Rubrics of current Standards and Guidelines as starting point
    - Expand to include data visualization
    - Expand to encompass multidimensional, multivariate methods and modeling
  - Remove or replace references to p=0.05 or p<0.05 cited purely as a threshold: Where p-value supports inference based on magnitude and uncertainty, the p-value should be reported to several decimal places to be useful as a sliding-scale indicator
- Educate NCES Staff and Contractors
- Involve Experts in Report Development
  - Analysis plan approved by senior substantive expert (not necessarily NCES staff)
  - Statistical methods and results approved by senior statistician (author or expert)
- Align Review Process
  - Align to new Standards and Guidelines
  - Align to report types and intended audiences
  - Ensure linkage to statistical analysis

APPENDICES

Appendix A: Agenda

PCP 780

Thursday, January 24, 2019

8:30am	Arrival & Building Security
9:00am - 12:00pm	Welcome Introductions Commissioner's Remarks NCES Staff Presentations & Discussion
12:00pm - 1:00pm	Lunch (on your own)
1:00pm - 4:30pm	Panel Executive Session
4:30pm - 5:00pm	Clarification Requests of NCES from the Panel
5:30pm	Adjourn

Friday, January 25, 2019

8:30am	Arrival & Building Security
9:00am - 11:00am	Panel Executive Working Session
11:00am - 12:00pm	If Useful: NCES Staff Responses to Panel Requests
12:00pm - 3:30pm	Panel Executive Session & Working Lunch (Panel will purchase lunch & return to meeting room)
3:30pm - 5:00pm	Panel Feedback to NCES
5:00pm	Adjourn

Appendix B: Expert Panel Biosketches

Rajeev Darolia, *Ph.D.*

Title: Associate Professor, University of Kentucky

Rajeev Darolia is an Associate Professor of Public Policy and Economics (by courtesy) at the University of Kentucky. Professor Darolia teaches classes in causal research methods and program evaluation. His current research interests include questions about how public policy affects economic mobility and financial security, especially as it relates to education policy. Dr. Darolia publishes research across public policy, economics, and education journals, and his work has been funded by the National Science Foundation, the US Department of Labor, and the Association for Institutional Research, among others. Dr. Darolia is also a Visiting Scholar at the Federal Reserve Bank of Philadelphia, a Research Fellow at the IZA Institute of Labor Economics, and a 2018 National Academy of Education/Spencer Postdoctoral Fellow. He serves on the editorial boards of the Educational Evaluation and Policy Analysis, Journal of Higher Education, and Educational Researcher. Dr. Darolia received a PhD in Public Policy from George Washington University; he also holds a master's degree in economics and a bachelor's degree in finance.

Susanna Loeb*, Ph.D.*

Title: Professor, Brown University

Susanna Loeb is the Director of the Annenberg Institute and Professor of Education and International and Public Affairs at Brown University. Susanna's research focuses broadly on education policy and its role in improving educational opportunities for students. Her work has addressed issues of educator career choices and professional development, of school finance and governance, and of early childhood systems. Before moving to Brown, Susanna was the Barnett Family Professor of Education at Stanford University. She was the founding director of the Center for Education Policy at Stanford and codirector of Policy Analysis for California Education. Susanna led the research for both Getting Down to Facts projects for California schools. She has been a member of the National Board for Education Sciences, a senior fellow at the Stanford Institute for Economic Policy Research, and a faculty research fellow at the National Bureau of Economic Research.

Allen Schirm*, Ph.D.*

Title: Retired, Mathematica

Allen Schirm retired from Mathematica Policy Research in 2016 after more than 27 years, during which he held several positions, including Vice President, Director of Human Services Research, Director of Methods, and Senior Fellow. He is a fellow of the American Statistical Association, and was designated a National Associate of the National Academies of Sciences, Engineering, and Medicine "in recognition of extraordinary service" to the National Academies. Recently, he served as co-editor of a special issue of *The American Statistician* entitled "Statistical Inference in the 21st Century: A World Beyond 'P<0.05'," which was published in March 2019. Dr. Schirm received an A.B., *summa cum laude*, in statistics from Princeton University and a Ph.D. in economics from the University of Pennsylvania.

Mark Wilson*, Ph.D.*

Title: Professor, University of California, Berkeley

Mark Wilson is a professor of Education at UC, Berkeley, and also at the University of Melbourne. He received his PhD degree from the University of Chicago in 1984. His interests focus on measurement and applied statistics, and he has published over 120 refereed articles in those areas, and over 60 invited chapters. Recently he was elected president of the Psychometric Society, and also president of the US National Council on Measurement in Education (NCME); he is also a Member of the US National Academy of Education, a Fellow of the American Educational Research Association, and a National Associate of the US National Research Council. He is Director of the Berkeley Evaluation and Assessment Research (BEAR) Center. His research interests focus on the development and application of sound approaches for measurement in education and the social sciences, the development of statistical models suitable for measurement contexts, the creation of instruments to measure new constructs, and scholarship on the philosophy of measurement.

Linda J. Young*, Ph.D.*

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*, Ph.D.*

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. Her technical expertise includes 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.