DRAFT
December 15, 1996

Initial Report

Task Force on Statistical Inference

Board of Scientific Affairs
American Psychological Association

This report is the result of the initial two-day meeting of the Task Force on Statistical Inference held December 14-15, 1996 at the Newark Airport. The task force welcomes written reactions to this initial report from all interested parties, input that will inform our future deliberations. **The deadline for receiving reactions to this initial report is May 30, 1997**. Written responses should be sent to: Sangeeta Panicker, Liaison to the Task Force on Statistical Inference, APA - Science Directorate, 750 First Street (NE), Washington, DC 20002, e-mail: sxp.apa@email.apa.org

Members present:
Robert Rosenthal, PhD (Co-Chair)
Jacob Cohen, PhD (Co-Chair)
Leona S. Aiken, PhD
Mark Appelbaum, PhD
Gwyneth M. Boodoo, PhD
David A. Kenny, PhD
Helena C. Kraemer, PhD
Donald B. Rubin, PhD
Howard Wainer, PhD*
Leland Wilkinson, PhD

Members absent:
Robert Abelson, PhD (Co-Chair)
Bruce Thompson, PhD

APA Staff:
Christine R. Hartel, PhD*
Sangeeta Panicker, Liaison

* denotes partial attendance

This report reflects the initial deliberations of the task force and the first of our recommendations to the Board of Scientific Affairs (BSA). We address two issues. First, we consider the issue that brought the task force into existence, namely the role of null hypothesis significance testing in psychological research. Second, we consider the modification of current practice in the quantitative treatment of data in the science of psychology.

Null Hypothesis Significance Testing

Many have assumed the charge to this task force to be narrowly focused on the issue of null hypothesis significance testing and particularly the use of the *p* value. The charge this task force has accepted, however, is broader. It is the view of the task force that there are many ways of using statistical methods to help us understand the phenomena we are studying (e.g., Bayesian methods, graphical and exploratory data analysis methods, hypothesis testing strategies). We endorse a policy of inclusiveness that allows any procedure that *appropriately* sheds light on the phenomenon of interest to be included in the arsenal of the research scientist. In this spirit, the task force does not support any action that could be interpreted as banning the use of null hypothesis significance testing or *p* values in psychological research and publication.

The Broader Topics of Recommendations

Four broad topics of the quantitative treatment of research data in which the task force believes major improvements in current practice could and should be made were identified at this meeting. These topics are: (1) approaches to enhance the quality of data usage and to protect against potential misinterpretation of quantitative results, (2) the need for theory-generating studies, (3) the use of minimally sufficient designs and analytic strategies, and (4) issues with computerized data analyses.

(1) Approaches to enhance the quality of data usage and to protect against potential misinterpretation of quantitative results

Of these four topics, the first has so far received the greatest attention from the task force. With respect to this topic the task force has identified three issues that are particularly germane to current practice.

(a) we recommend that more extensive descriptions of the data be provided to reviewers and readers. This should include means, standard deviations, sample sizes, five-point summaries, box-and-whisker plots, other graphics, and descriptions related to missing data as appropriate.

(b) enhanced characterization of the *results of analyses* (beyond simple *p* value statements) to

include both direction and size of effect (e.g., mean difference, regression and correlation coefficients, odds-ratios, more complex effect size indicators) and their confidence intervals should be provided routinely as part of the presentation. These characterizations should be reported in the most interpretable metric (e.g., the expected unit change in the criterion for a unit change in the predictor, Cohen's *d*).

(c) the use of techniques to assure that the reported results are not produced by anomalies in the data (e.g., outliers, points of high influence, non-random missing data, selection, attrition problems) should be a standard component of all analyses.

(2) The need for theory-generating studies

In its recent history, psychology has been dominated by the hypothetico-deductive approach. It is the view of the task force that researchers have too often been forced into the premature formulation of theoretical models in order to have their work funded or published. The premature formulation of theoretical models has often led to the worst problems seen in the use of null hypothesis testing, such as misrepresentation of exploratory results as confirmatory studies, or poor design of confirmatory studies in the absence of necessary exploratory results. We propose that the field become more open to well formulated and well conducted exploratory studies with the appropriate quantitative treatment of their results, thereby enhancing the quality and utility of future theory generation and assessment.

(3) The use of minimally sufficient designs and analytic strategies

The wide array of quantitative techniques and the vast number of designs available to address research questions leave the researcher with the non-trivial task of matching analysis and design to the research question. Many forces (including reviewers of grants and papers, journal editors, and dissertation advisors) compel researchers to select increasingly complex ("state-of-the-art," "cutting edge," etc.)analytic and design strategies. Sometimes such complex designs and analytic strategies are necessary to address research questions effectively; it is also true that simpler approaches can provide elegant answers to important questions. It is the recommendation of the task force that the principle of parsimony be applied to the selection of designs and analyses. The minimally sufficient design and analysis is typically to be preferred because:

(a) it is often based on the fewest and least restrictive assumptions,

(b) its use is less prone to errors of application, and errors are more easily recognized, and

(c) its results are easier to communicate--to both the scientific and lay communities. This is not to say that new advances in both design and analysis are not needed, but simply that newer is not necessarily better and that more complex is not necessarily preferable.

(4) Issues with computerized data analysis

Elegant and sophisticated computer programs have increased our ability to analyze data with substantially greater sophistication than was possible only a short time ago. The ease of access to state-of-the-art statistical analysis packages, however, has not universally advanced our science. Common misuses of computerized data analysis include:

(a) reporting statistics without understanding how they are computed or what they mean,

(b) relying on results without regard to their reasonableness, or without verification by independent computation, and

(c) reporting results to greater precision than supported by the data, simply because they are printed by the program. The task force encourages efforts to avoid the sanctification of computerized data analysis. Computer programs have placed a much greater demand on researchers to understand and control their analysis and design choices.