Statistical Methods in Psychology Journals
Guidelines and Explanations

Leland Wilkinson
and
The Task Force on Statistical Inference
APA Board of Scientific Affairs

Introduction

In the light of continuing debate over the applications of significance testing in psychology journals and following the publication of Cohen (1994), the Board of Scientific Affairs (BSA) of the APA convened a committee called the Task Force on Statistical Inference (TFSI) whose charge was "to elucidate some of the controversial issues surrounding applications of statistics including significance testing and its alternatives; alternative underlying models and data transformation; and newer methods made possible by powerful computers." Robert Rosenthal, Robert Abelson, and Jacob Cohen (cochairs) met initially and agreed upon the desirability of having several types of specialists on the Task Force: statisticians, teachers of statistics, journal editors, authors of statistics books, computer experts, and wise elders. Nine individuals were subsequently invited to join and all agreed. These were: Leona Aiken, Mark Appelbaum, Gwyneth Boodoo, David A. Kenny, Helena Kraemer, Donald Rubin, Bruce Thompson, Howard Wainer, and Leland Wilkinson. In addition, Lee Cronbach, Paul Meehl, Frederick Mosteller and John Tukey served as Senior Advisors to the Task Force and commented on written materials.

The TFSI met twice in two years and corresponded throughout the period. After the first meeting, the Task Force circulated a preliminary report indicating its intention to examine issues beyond null hypothesis significance testing. The Task Force invited comments and used this feedback in the deliberations during its second meeting.

After the second meeting, the Task Force recommended several possibilities for further action, chief of which would be to revise the statistical sections of the APA Publication Manual. After extensive discussion, the BSA recommended that "before the TFSI undertook a revision of the APA Publication Manual, it might want to consider publishing an article in *American Psychologist*, as a way to initiate discussion in the field about changes in current practices of data analysis and reporting."

This article follows that request. The sections in italic are proposed guidelines that the TFSI recommends could be used for revising the APA Publication Manual or for developing other BSA supporting materials. Following each guideline are comments, explanations, or elaborations assembled by Leland Wilkinson for the Task Force and under its review. This paper is concerned with the use of statistical methods only and is not meant as an assessment of research methods in general. Psychology is a broad science. Methods appropriate in one area may be inappropriate in another.

The title and format of this paper are adapted from a similar article by Bailar and Mosteller (1988) that interprets the 1988 edition of the *Uniform Requirements for Manuscripts Submitted to Biomedical Journals*. That article should be consulted, since it overlaps somewhat with this one and discusses some issues relevant to research in psychology. Further detail can also be found in the publications on this topic by several committee members (Abelson, 1995, 1997; Rosenthal, 1994; Thompson, 1996; Wainer, 1999; see also articles in Harlow et al., 1997).

Method

Design

Make clear at the outset what type of study you are doing. Do not cloak a study in one guise to try to give it the assumed reputation of another. For studies that have multiple goals, be sure to define and prioritize those goals.

There are many forms of empirical studies in psychology, including case reports, controlled experiments, quasi-experiments, statistical simulations, surveys, observational studies, and studies of studies (meta-analyses). Some are hypothesis-generating: to explore data in order to form or sharpen hypotheses about a population for assessing future hypotheses. Some are hypothesis-testing: to assess specific *a priori* hypotheses or to estimate parameters by random sampling from that population. Some are meta-analytic: to assess specific *a priori* hypotheses or to estimate parameters (or both) by synthesizing the results of available studies.

Some researchers have the impression or have been taught to believe that some of these forms yield information that is more valuable or credible than others (see Cronbach, 1975 for a discussion). Occasionally proponents of some research methods disparage others. In fact, each form of research has its own strengths, weaknesses, and standards of practice.

Population

The interpretation of the results of any study depends on the characteristics of the population analyzed. Define the population (subjects, stimuli, or studies) clearly. If control or comparison groups are part of the design, present how they are defined.

Psychology students sometimes think that a statistical population is the human race or, at least, college sophomores. They also have some difficulty distinguishing a class of objects vs. a statistical population - that sometimes we make inferences about a population via statistical methods and other times we make inferences about a class through logical or other non-statistical methods. Populations may be sets of potential observations on people, adjectives, or even research papers. How a population is defined in a paper affects almost every conclusion in that paper.

Sample

Describe the sampling procedures, and emphasize any inclusion or exclusion criteria. If the sample is stratified (e.g., by site or gender) describe fully the method and rationale. Note the proposed sample size for each subgroup.

Interval estimates for clustered and stratified random samples differ from those for simple random samples. Statistical software is now becoming available for these purposes. If you are using a convenience sample (whose members are not selected at random), be sure to make that procedure clear to your readers. Using a convenience sample does not automatically disqualify a study from publication, but it harms your objectivity to try to conceal this by implying that you used a random sample. Sometimes the case for the representative-

ness of a convenience sample can be strengthened by explicit comparison of sample characteristics with those of a defined population across a wide range of variables.

Assignment

Random Assignment

For research involving causal inferences, the assignment of units to levels of the causal variable is critical. Random assignment (not to be confused with random selection) allows for the strongest possible causal inferences free of extraneous assumptions. If random assignment is planned, provide enough information to show that the process for making the actual assignments is random.

There is a strong research tradition and many exemplars for random assignment in various fields of psychology. Even those who have elucidated quasi-experimental designs in psychological research (*e.g.*, Cook and Campbell, 1979) have repeatedly emphasized the superiority of random assignment as a method for controlling bias and lurking variables.

"Random" does not mean "haphazard." Randomization is a fragile condition, easily corrupted: deliberately, as we see when a skilled magician flips a fair coin repeatedly to heads, or innocently, as we saw in the Vietnam draft lottery. As psychologists, we also know that human subjects are incapable of producing a random process (digits, spatial arrangements, etc.) or recognizing one. It is best not to trust the random behavior of a physical device unless you are an expert in these matters. It is safer to use the pseudo-random sequence from a well-designed computer generator or from published tables of random numbers. The added benefit of such a procedure is that you can supply a random number seed or starting number in a table that other researchers can use to check your methods later.

Nonrandom Assignment

For some research questions, random assignment is infeasible. In such cases, we need to minimize effects of variables that affect the observed relationship between a causal variable and an outcome. Such variables are commonly called confounds or covariates. The researcher needs to attempt to determine the relevant covariates, measure them adequately, and adjust for their effects either by design or by analysis. If the effects of covariates are adjusted by analysis, the strong assumptions that are made must be explicitly stated and, to the extent possible, tested and justified. Describe methods used to attenuate sources of bias, including plans for minimizing dropouts, non-compliance, and missing data.

Authors have used the term "control group" to describe, among other things, 1) a comparison group, 2) members of pairs matched or blocked on one or more nuisance variables, 3) a group not receiving a particular treatment, 4) a statistical sample whose values are adjusted *post-hoc* by the use of one or more covariates, or 5) a group for which the experimenter acknowledges bias exists and perhaps hopes that this admission will allow the reader to make appropriate discounts or other mental adjustments. None of these is an instance of a fully-adequate control group.

From this perspective, one can make a recommendation concerning editorial usage. If we can neither implement randomization nor approach total control of variables that modify effects (outcomes), then we should use the term "control group" cautiously. In particular, we should describe exactly which confounding variables have been explicitly controlled and speculate about which unmeasured ones could lead to incorrect inferences. In the absence of randomization, we should do our best to investigate sensitivity to various untestable assumptions.

Measurement

Variables

Explicitly define the variables in the study, show how they are related to the goals of the study, and explain how they are measured. The units of measurement of all variables, causal and outcome, should fit the language you use in the Introduction and Discussion sections of your report.

A variable is a method for assigning to a set of observations a value from a set of possible outcomes. For example, a variable called Gender might assign each of 50 observations to one of the values *male* or *female*. When we define a variable, we are declaring what we are prepared to represent as a valid observation and what we must consider as invalid. If we define the range of a particular variable (the set of possible outcomes) to be from 1 to 7 on a Likert scale, for example, then a value of 9 is not an outlier (an unusually extreme value). It is an illegal value. If we declare the range of a variable to be positive real numbers and the domain to be observations of reaction time in milliseconds to an administration of electric shock, then a value of 3,000 is not illegal; it is an outlier.

Naming a variable is almost as important as measuring it. We do well to select a name that reflects how a variable is measured. On this basis, the name "IQ test score" is preferable to "Intelligence" and "retrospective self-report of childhood sexual abuse" is preferable to "childhood sexual abuse." Without such precision, ambiguity in defining variables can give a theory an unfortunate resistance to empirical falsification. Being precise does not make us operationalists. It simply means that we try to avoid excessive generalization.

Editors and reviewers should be suspicious when they notice authors changing definitions or names of variables, failing to make clear what would be contrary evidence, or using measures with no past history, and thus no known properties. Researchers should be suspicious when codebooks and scoring systems are inscrutable or more voluminous than the research papers on which they are based. Everyone should worry when a system offers to code a specific observation in two or more ways for the same variable.

Instruments

If a questionnaire is used to collect data, summarize the psychometric properties of its scores with specific regard to the way the instrument is used and its intended population. Psychometric properties include measures of validity, reliability, and any other qualities affecting conclusions. If a physical apparatus is used, provide enough information (brand, model, design specifications) to allow another experimenter to replicate your measurement process.

There are many methods for constructing instruments and psychometrically validating scores from such measures. Traditional true-score theory and item-response test theory provide appropriate frameworks for assessing reliability and internal validity. Signal detection theory and various coefficients of association can be used to assess external validity. Messick (1989) provides a comprehensive guide to validity.

It is important to remember that a test is not reliable or unreliable. Reliability is a property of the scores on a test for a particular population of examinees (Feldt & Brennan, 1989). Thus, authors should provide reliability coefficients of the scores for the data being analyzed even when the focus of their research is not psychometric. Interpreting the size of observed effects requires an assessment of the reliability of the scores.

Besides showing that an instrument is reliable, we need to show that it does not correlate strongly with other key constructs. It is just as important to establish that a measure does *not* measure what it should not measure as to show that it *does* measure what it should.

Researchers occasionally encounter a measurement problem that has no obvious solution. This happens when they decide to explore a new and rapidly growing research area that is based on a previous researcher's well-defined construct implemented in a poorly developed psychometric instrument. Innovators, in the excitement of their discovery, sometimes give insufficient attention to the quality of their instruments. Once a defective measure enters the literature, subsequent researchers are reluctant to change it. In these cases, editors and reviewers should pay special attention to the psychometric properties of the instruments used, and they might want to encourage revisions (even if not by the scale's author) in order to prevent the accumulation of results based on relatively invalid or unreliable measures.

Procedure

Describe any anticipated attrition due to noncompliance, dropout, death, or other factors. Indicate how such attrition may affect the generalizability of the results. Clearly describe the conditions under which measurements are taken, e.g., format, time, place, personnel used to collect data. Describe the specific methods used to minimize experimenter bias, especially if you collected the data yourself.

Despite the long-established findings of the effects of experimenter bias (Rosenthal, 1966), many published studies appear to ignore or discount these problems. For example, some authors or their assistants with knowledge of hypotheses or study goals screen subjects (through personal interviews or telephone conversations) for inclusion in their studies. Some authors administer questionnaires. Some authors give instructions to subjects. Some authors perform experimental manipulations. Some tally or code responses. Some rate videotapes.

An author's self-awareness, experience, or resolve does not eliminate experimenter bias. In short, there are no valid excuses, financial or otherwise, for avoiding an opportunity to double-blind. Researchers looking for guidance on this matter should consult the classic book of Webb *et al*. (1966) and an exemplary dissertation (performed on a modest budget) by Baker (1969).

Power and Sample Size

Provide information on sample size and the process that led to sample size decisions. Document the effect sizes, sampling and measurement assumptions, as well as analytic procedures used in power calculations. Because power computations are most meaningful when done before data are collected and examined, it is important to show how effect-size estimates have been derived from previous research and theory in order to dispel suspicions that they might have been taken from data used in the study or, even worse, constructed to justify a particular sample size. Once the study is analyzed, confidence intervals replace calculated power in describing results.

Largely due to the work of Cohen (1969, 1988), psychologists have become aware of the need to consider power in the design of their studies, before they collect data. The intellectual exercise required to do this stimulates authors to take seriously prior research and theory in their field. And it gives an opportunity, with incumbent risk, for a few to offer the challenge that there is no applicable research behind a given study. If exploration were not disguised in hypothetico-deductive language, then it might have the opportunity to influence subsequent research constructively.

Computer programs that calculate power for various designs and distributions are now available. One can use them to conduct power analyses for some range of reasonable alpha values and effect sizes. Doing so reveals how power changes across this range and overcomes a tendency to regard a single power estimate as being absolutely definitive.

Many of us encounter power issues when applying for grants. Even when not asking for money, think about power. Statistical power does not corrupt.

Results

Complications

Before presenting results, report complications, protocol violations, and other unanticipated events in data collection. These include missing data, attrition, and nonresponse. Discuss analytic techniques devised to ameliorate these problems. Describe nonrepresentativeness statistically by reporting patterns and distributions of missing data and contaminations. Document how the actual analysis differs from the analysis planned before complications arose. 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 bias, attrition problems) should be a standard component of all analyses.

As soon as you have collected your data, before you compute *any* statistics, *look at your data*. Data screening is not data snooping. It is not an opportunity to discard data or change values in order to favor your hypotheses. However, if you assess hypotheses without examining your data, you risk publishing nonsense.

Computer malfunctions tend to be catastrophic: a system crashes, a file fails to import, data are lost. Less well-known are more subtle bugs that can be more catastrophic in the

long run. For example, a single value in a file may be corrupted in reading or writing (often in the first or last record). This circumstance usually produces a major value error, the kind of singleton that can make large correlations change sign and small correlations become large.

Graphical inspection of data offers an excellent possibility for detecting serious compromises to data integrity. The reason is simple: graphics broadcast, statistics narrowcast. Indeed, some international corporations that must defend themselves against rapidly evolving fraudulent schemes use real-time graphical displays as their first line of defense and statistical analyses as a distant second. The following example shows why.

Figure 1 shows a scatterplot matrix (SPLOM) of three variables from a national survey of approximately 3000 counseling clients (Chartrand, 1997). This display, consisting of pairwise scatterplots arranged in a matrix, is found in most modern statistical packages. The diagonal cells contain dot plots of each variable (with the dots stacked like a histogram) and scales used for each variable. The three variables shown are questionnaire measures of respondent's age (AGE), gender (SEX), and number of years together in current relationship (TOGETHER). The graphic in Figure 1 is not intended for final presentation of results; we use it instead to locate coding errors and other anomalies before we analyze our data. Figure 1 is a selected portion of a computer screen display that offers tools for zooming in and out, examining points, and linking to information in other graphical displays and data editors. SPLOM displays can be used to recognize unusual patterns in 20 or more variables simultaneously. We will focus on these three only.

There are several anomalies in this graphic. The AGE histogram shows a spike at the right end, which corresponds to the value 99 in the data. This coded value most likely signifies a missing value, since it is unlikely that this many people in a sample of 3000 would have an age of 99 or greater. Using numerical values for missing value codes is a risky practice (Kahn & Udry, 1986).

The histogram for SEX shows an unremarkable division into two values. The histogram for TOGETHER is highly skewed, with a spike at the lower end presumably signifying no relationship. The most remarkable pattern is the triangular joint distribution of TOGETHER and AGE. Triangular joint distributions often (but not necessarily) signal an implication or a relation rather than a linear function with error. In this case, it makes sense that the span of a relationship should not exceed a person's age. Closer examination shows that something is wrong here, however. We find some respondents (in the upper left triangular area of the TOGETHER-AGE panel) claiming that they have been in a significant relationship longer than they have been alive!

Had we computed statistics or fit models before examining the raw data, we would likely have missed these reporting errors. There is little reason to expect that TOGETHER would show any anomalous behavior with other variables, and even if AGE and TOGETHER appeared jointly in certain models, we may not have known anything was amiss, regardless of our care in examining residual or other diagnostic plots.

The main point of this example is that the type of "atheoretical" search for patterns that we are sometimes warned against in graduate school can save us from the humiliation of having to retract conclusions we might ultimately make on contaminated data. We are warned against fishing expeditions for understandable reasons, but blind application of models without screening our data is a far graver error.

Graphics cannot solve all our problems. Special issues arise in modeling when we have missing data. The two popular methods for dealing with missing data that are found in basic statistics packages - listwise and pairwise deletion of missing values - are among the worst

methods available for practical applications. Little and Rubin (1987) discuss these issues in more detail and offer alternative approaches.

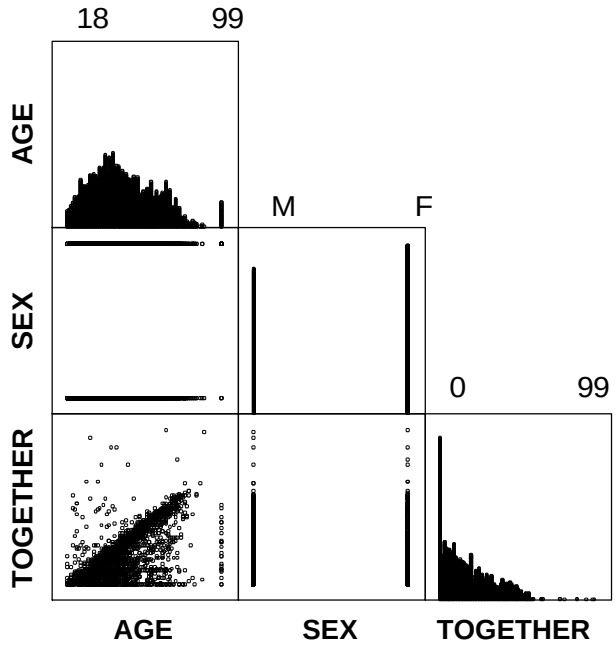

Figure 1 Scatterplot Matrix

Analysis

Choosing a minimally sufficient analysis

The enormous variety of modern quantitative methods leaves researchers with a non-trivial task of matching analysis and design to the research question. Although complex designs and state-of-the-art methods are sometimes necessary to address research questions effectively, simpler classical approaches often can provide elegant and sufficient answers to important questions. Do not choose an analytic method to impress your readers or to deflect criticism. If the assumptions and strength of a simpler method are reasonable for your data and research problem, use it. Occam's razor applies to methods as well as to theories.

We should follow the advice of Fisher (1935):

> Experimenters should remember that they and their colleagues usually know more about the kind of material they are dealing with than do the authors of text-books written without such personal experience, and that a more complex, or less intelligible, test is not likely to serve their purpose better, in any sense, than those of proved value in their own subject.

There is nothing wrong with using state-of-the-art methods, as long as you and your readers understand how they work and what they are doing. On the other hand, don't cling to

obsolete methods (e.g., Newman-Keuls or Duncan post-hoc tests) out of fear of learning the new. In any case, listen to Fisher. Begin with an idea. Then pick a method.

Computer programs

There are many good computer programs for analyzing data. More important than choosing a specific statistical package is verifying your results, understanding what they mean, and knowing how they are computed. If you cannot verify your results by intelligent "guesstimates," you should check them against the output of another program. You will not be happy if a vendor reports a bug after your data are in print (not an infrequent event). Do not report statistics found on a printout without understanding how they are computed or what they mean. Do not report statistics to a greater precision than supported by your data simply because they are printed by the program. Using the computer is an opportunity for you to control your analysis and design. If a computer program does not provide the analysis you need, use another program rather than let the computer shape your thinking.

There is no substitute for common sense. If you cannot use rules of thumb to detect whether the result of a computation makes sense to you, then you should ask yourself whether the procedure you are using is appropriate for your research. Graphics can help you to make some of these determinations; theory can help in other cases. But never assume that using a highly regarded program absolves you of the responsibility for judging whether your results are plausible. Finally, when documenting the use of a statistical procedure, refer to the statistical literature rather than a computer manual; when documenting the use of a program, refer to the computer manual rather than the statistical literature.

Assumptions

You should take efforts to assure that the underlying assumptions required for the analysis are reasonable given the data. Examine residuals carefully. Do not use distributional tests and statistical indexes of shape (e.g., skewness, kurtosis) as a substitute for examining your residuals graphically.

Using a statistical test to diagnose problems in model fitting has two shortcomings. Often significance tests based on summary statistics (such as tests for homogeneity of variance) are impractically sensitive. They lead us to reject fits that are relatively robust to these violations. Secondly, statistics such as skewness and kurtosis often fail to detect distributional irregularities in the residuals. Third, statistical tests depend on sample size, and as sample size increases, they often will reject innocuous assumptions. In general, there is no substitute for graphical analysis of assumptions.

Modern statistical packages offer graphical diagnostics for helping to determine whether a model appears to fit data appropriately. Most users are familiar with residual plots for linear regression modeling. Fewer are aware that John Tukey's paradigmatic equation *data = fit + residual* applies to a more general class of models and has broad implications for graphical analysis of assumptions. Stem-and-leaf plots, box plots, histograms, dot plots, spread/level plots, probability plots, spectral plots, autocorrelation and cross-correlation plots, co-plots, and trellises (Tukey, 1977; Chambers *et al.*, 1983; Cleveland, 1995) all

serve at various times for displaying residuals, whether they arise from ANOVA, nonlinear modeling, factor analysis, latent variable modeling, multidimensional scaling, hierarchical linear modeling, or other procedures.

Hypothesis tests

It is hard to imagine a situation in which a dichotomous accept-reject decision is better than reporting an actual p-value or, better still, a confidence interval. Never use the unfortunate expression, "accept the null hypothesis." Always provide some effect-size estimate when reporting a p-value.

Cohen (1994) has written on this subject in this journal. All psychologists would benefit from reading his insightful paper.

Effect sizes

Always present effect sizes for primary outcomes. If the units of measurement are practically meaningful (e.g., number of cigarettes smoked per day), then we should usually prefer an unstandardized measure (regression coefficient or mean difference) to a standardized measure (r or d). It helps to add brief comments that place these effect sizes in a practical and theoretical context.

The 1994 APA publication manual included an important new "encouragement" (p. 18) to report effect sizes. Unfortunately, empirical studies of various journals indicate that the effect size of this encouragement has been negligible (Kirk, 1996; Keselman et al., 1998; Thompson & Snyder, 1998). We must stress again that reporting and interpreting effect sizes in the context of previously reported effects is essential to good research. It enables readers to evaluate the stability of results across samples, designs, and analyses. Reporting effect sizes also informs power analyses and meta-analyses needed in future research.

Snyder and Lawson (1993), Fleiss (1994), Rosenthal (1994), and Kirk (1996) summarize various measures of effect sizes used in psychological research. Consult these articles for information on computing them. For a simple, general purpose display of the practical meaning of an effect size, see Rosenthal and Rubin (1982). Consult Rosenthal and Rubin (1994) for information on the use of "counternull intervals" for effect sizes as alternatives to confidence intevals.

Interval estimates

Interval estimates should be given for any effect sizes involving principal outcomes. Provide intervals for correlations and other coefficients of association or variation whenever possible.

Confidence intervals are usually available in statistical software; otherwise, confidence intervals for basic statistics can be computed from typical output. Comparing confidence intervals from a current study to intervals from previous, related studies helps focus attention on stability across studies (Schmidt, 1996). Collecting intervals across studies also helps in constructing plausible regions for population parameters. This practice should help

prevent the common mistake of assuming a parameter is contained in a confidence interval.

Multiplicities

Multiple outcomes require special handling. There are many ways to conduct reasonable inference when faced with multiplicity, e.g., Bonferroni correction of p-values, multivariate test statistics, empirical Bayes methods. It is your responsibility to define and justify the methods used.

Statisticians speak of "the curse of dimensionality." To paraphrase, multiplicities are the curse of the social sciences. In many areas of psychology, we cannot do research on important problems without encountering multiplicity. We often encounter many variables and many relationships.

One of the most prevalent strategies psychologists use to handle multiplicity is to follow an ANOVA with pairwise multiple-comparison tests. This approach is usually wrong for several reasons. First, pairwise methods such as Tukey's *HSD* procedure were designed to control a familywise error rate based on the sample size and number of comparisons. Preceding them with an omnibus *F*-test in a stage-wise testing procedure defeats this design, making it unnecessarily conservative. Second, researchers rarely need to compare all possible means to understand their results or assess their theory; by setting their sights large, they sacrifice their power to see small. Third, the lattice of all-possible pairs is a straightjacket; forcing themselves to wear it often restricts researchers to uninteresting hypotheses and induces them to ignore more fruitful ones.

As an antidote to the temptation to explore all pairs, imagine yourself restricted to mentioning only pairwise comparisons in the Introduction and Discussion sections of your report. Higher-order concepts such as trends, structures, or clusters of effects would be forbidden. Your theory would be restricted to first-order associations. This scenario brings to mind the illogic of the converse, popular practice of theorizing about higher-order concepts in the Introduction and Discussion sections and supporting that theorizing in the Results section with atomistic pairwise comparisons. If a specific contrast interests you, examine it. If all interest you, ask yourself why. For a detailed treatment of the use of contrasts, see Rosenthal, Rosnow, & Rubin (in press).

There is a variant of this preoccupation with all-possible-pairs that comes with the widespread practice of printing *p*-values or asterisks next to every correlation in a correlation matrix. Methodologists frequently point out that these *p*-values should be adjusted through Bonferroni or other corrections. One should ask instead why any reader would want this information. The possibilities are:

1) All the correlations are "significant." If so, this can be noted in a single footnote.

2) None of the correlations is "significant." Again, this can be noted once. We need to be reminded that this situation does not rule out the possibility that combinations or subsets of the correlations may be "significant." The definition of the null-hypothesis for the global test may not include other potential null-hypotheses that might be rejected if they were tested.

3) A subset of the correlations is "significant." If so, our purpose in appending asterisks would seem to be to mark this subset. Using "significance" tests in this way is really a highlighting technique to facilitate pattern recognition. If this is your goal in presenting results, then it is better served by calling attention to the pattern (perhaps by sorting the rows and columns of the correlation matrix) and assessing it directly. This would force you, as well,

to provide a plausible explanation.

There is a close relative of “all-possible-pairs” called “all-possible-combinations.” We see this occasionally in the publishing of higher-way factorial ANOVA’s that include all possible main effects and interactions. One should not imagine that placing asterisks next to conventionally “significant” effects in a 5-way ANOVA, for example, skirts the multiplicity problem. A typical 5-way fully factorial design applied to a reasonably large sample of random data has about an 80 percent chance of producing at least one “significant” effect by conventional *F*-tests at the .05 critical level (Hurlburt and Spiegel, 1976).

Underlying the widespread use of all-possible-pairs methodology is the legitimate fear among editors and reviewers that some researchers would indulge in “fishing expeditions” without the restraint of simultaneous test procedures. We should indeed fear the well-intentioned, indiscriminate search for structure more than the deliberate falsification of results, if only for the prevalence of wishful thinking over nefariousness. There are Bonferroni and recent related methods (*e.g.*, Benjamini & Hochberg, 1995) for controlling this problem statistically. Nevertheless, there is an alternative restraint. Reviewers should require writers to articulate their expectations well enough to reduce the likelihood of post-hoc rationalizations. Fishing expeditions are often recognizable by the promiscuity of their explanations. They mix ideas from scattered sources, rely heavily on common sense, and cite fragments rather than trends.

If, on the other hand, a researcher “fools” us with an intriguing result caught while indiscriminately fishing, we might want to fear this possibility less than we do now. The enforcing of rules to prevent chance results in our journals may at times distract us from noticing the more harmful possibility of publishing bogus theories and methods (ill-defined variables, lack of parsimony, experimenter bias, logical errors, artifacts) that are buttressed by evidently impeccable statistics. There are enough good ideas behind fortuitous results to make us wary of restricting them. This is especially true in those areas of psychology where lives and major budgets are not at stake. Let replications promote reputations.

Causality

Inferences of causality from non-randomized designs are fraught with pitfalls. Researchers using non-randomized designs have an extra obligation to explain the logic behind covariates included in their designs and to alert the reader to plausible rival hypotheses that might explain their results. Even in randomized experiments, attributing causal effects to any one aspect of the treatment condition requires support from additional experimentation.

It is sometimes thought that correlation does not prove causation but “causal modeling” does. Despite the admonitions of experts in this field, researchers sometimes use goodness-of-fit indices to hunt through thickets of competing models and settle on a plausible substantive explanation only in retrospect. McDonald (1997), in an analysis of an historical dataset, shows the dangers of this practice and the importance of substantive theory. Scheines *et al*. (1998; discussions following) offer similar cautions from a theoretical standpoint.

A generally accepted framework for formulating questions concerning the estimation of causal effects in social and biomedical science involves the use of "potential outcomes," with one outcome for each treatment condition. Although the perspective has old roots, including use by Fisher and Neyman in the context of completely randomized experiments analyzed by randomization-based inference (Rubin, 1990a), it is typically referred to as

"Rubin's Causal Model" or RCM (Holland, 1986). For extensions to observational studies and other forms of inference, see Rubin (1974, 1977, 1978). This approach is now relatively standard, even for settings with instrumental variables and multi-stage models or simultaneous equations.

The crucial idea is to set up the causal inference problem as one of missing data, as defined in Rubin (1976), where the missing data are the values of the potential outcomes under the treatment *not* received and the observed data include the values of the potential outcomes under the received treatments. Causal effects are defined on a unit level as the comparison of the potential outcomes under the different treatments, only one of which can ever be observed (we cannot go back in time to expose the unit to a different treatment). The essence of the RCM is to formulate causal questions in this way, and use formal statistical methods to draw probabilistic causal inferences, whether based on Fisherian randomization-based (permutation) distributions, Neymanian repeated-sampling randomization-based distributions, frequentist superpopulation sampling distributions, or Bayesian posterior distributions (Rubin, 1990b).

If a problem of causal inference cannot be formulated in this manner (as the comparison of potential outcomes under different treatment assignments), it is not a problem of inference for causal effects, and the use of "causal" should be avoided. To see the confusion that can be created by ignoring this requirement, see the classic "Lord's Paradox" and its resolution by the use of the RCM in Holland and Rubin (1983).

The critical assumptions needed for causal inference are essentially always beyond testing from the data at hand because they involve the missing data. Thus, especially when formulating causal questions from nonrandomized data, the underlying assumptions needed to justify any causal conclusions should be carefully and explicitly argued, not in terms of technical properties like "uncorrelated error terms," but in terms of real world properties, such as how the units received the different treatments.

The use of complicated "causal modeling" software rarely yields any results that have any interpretation as causal effects. If such software is used to produce anything beyond an exploratory description of a data set, the bases for such extended conclusions must be carefully presented and not just asserted based on imprecise labeling conventions of the software.

Tables and figures

While tables are commonly used to show exact values, well-drawn figures need not sacrifice precision. Figures attract the reader's eye and help convey global results. Because individuals have different preferences for processing complex information, it often helps to provide both tables and figures. This works best when figures are kept small enough to allow space for both formats. Avoid complex figures when simpler ones will do. In all figures, include graphical representations of interval estimates whenever possible.

Bailar and Mosteller (1988) offer helpful information on improving tables in published reports. Many of their recommendations (e.g., sorting rows and columns by marginal averages, rounding to a few significant digits, avoiding decimals when possible) are based on the clearly-written tutorials of Ehrenberg (1975, 1981).

A common deficiency of graphics in psychological publications is their lack of essential information. In most cases, this information is the shape or distribution of the data. Whether

from a negative motivation to conceal irregularities or from a positive belief that "less is more," omitting shape information from graphics often hinders scientific evaluation. Chambers *et al*. (1983) and Cleveland (1995) offer specific ways to address these problems. The following examples do this using two of the most frequent graphical forms in psychology publications.

Figure 2 shows plots based on data from 80 graduate students in a Midwestern university psychology department collected from 1969 through 1978. The variables are scores on the psychology advanced test of the Graduate Record Examination (GRE), the undergraduate grade point average (GPA), and whether or not a student completed a Ph.D. in the department (PhD). The left panel of Figure 2 shows a format appearing frequently in psychology journal articles: two regression lines, one for each group of students. This graphic conveys nothing more than four numbers, the slopes and intercepts of the regression lines. Because the scales have no physical meaning, seeing the slopes of lines (as opposed to reading the numbers) adds nothing to our understanding of the relationship.

The right panel of Figure 2 shows a scatterplot of the same data with a locally-weighted scatterplot smoother for each PhD group (Cleveland & Devlin, 1988). This robust curvilinear regression smoother (called LOESS) is available in modern statistics packages. Now we can see some curvature in the relationships. (When a model that includes a linear and quadratic term for GPA is computed, the apparent interaction involving the PhD and NoPhD groups depicted in the left panel disappears.) The graphic in the right panel tells us many things. We note the unusual student with a GPA of less than 4.0 and a psychology GRE score of 800, we note the less surprising student with a similar GPA but a low GRE score (both of whom failed to earn doctoral degrees), we note the several students who had among the lowest GRE scores but earned doctorates, and so on. We might imagine these kinds of cases in the left panel (as we should in any dataset containing error), but their location and distribution in the right panel tells us something about this specific dataset.

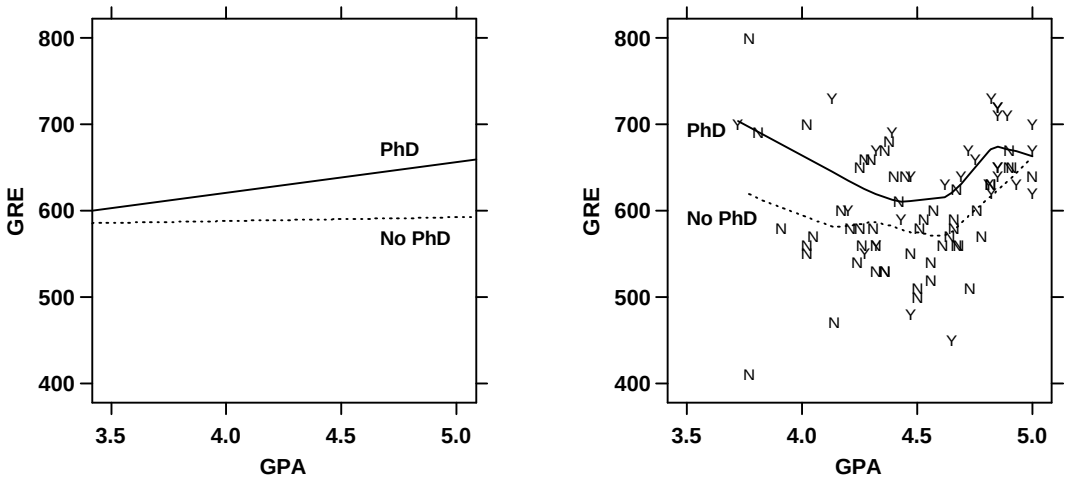

Figure 2 Graphics for Regression

The left panel of Figure 3 shows another popular format for displaying data in psychology journals. It is based on the dataset used for Figure 2. Authors frequently use this format for displaying the results of *t*-tests or ANOVAs. For factorial ANOVAs, this format gives

authors an opportunity to represent interactions by using a legend with separate symbols for each line. In more laboratory-oriented psychology journals (e.g., animal behavior, neuroscience), authors sometimes add error bars to the dots representing the means.

The right panel of Figure 3 adds to the line graphic a dot plot representing the data and 95 percent confidence intervals on the means of the two groups (using the *t*-distribution). The graphic reveals a left-skewness of GRE scores in the PhD group. While this skewness may not be severe enough to affect our statistical conclusions, it is nevertheless noteworthy. It may be due to ceiling effects (although note the 800 score in the NoPhD group) or some other factor. At the least, the reader has a right to be able to evaluate this kind of information.

There are other ways to include data or distributions in graphics, including box plots and stem-and-leaf plots (Tukey 1977) and kernel density estimates (Silverman, 1986; Scott, 1992). Many of these procedures are found in modern statistical packages. It is time for authors to take advantage of them and for editors and reviewers to urge authors to do so.

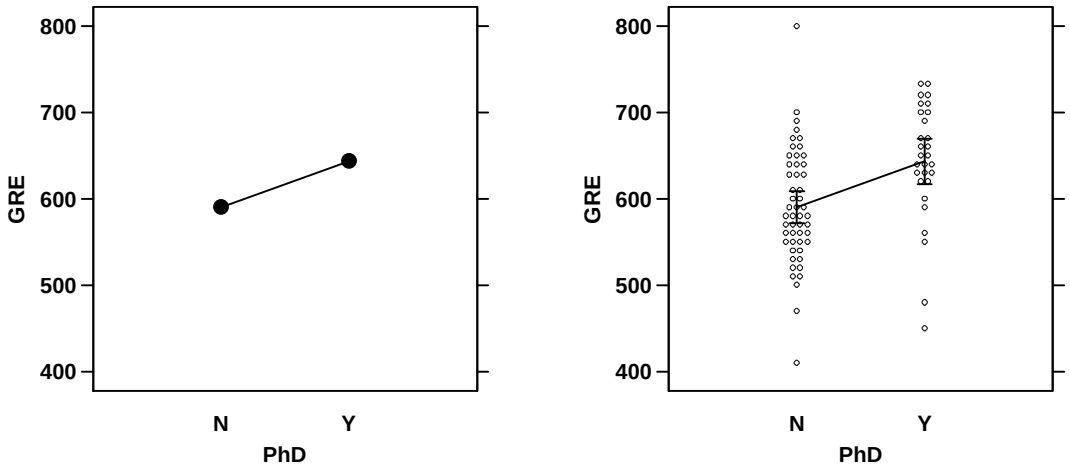

Figure 3 Graphics for Groups

Discussion

Interpretation

When you interpret effects, think of credibility, generalizability, and robustness. Are the effects credible, given the results of previous studies and theory? Do the features of the design and analysis (e.g., sample quality, similarity of the design to designs of previous studies, similarity of the effects to those in previous studies) suggest the results are generalizable? Are the design and analytic methods robust enough to support strong conclusions?

Novice researchers err either by over-generalizing their results or, equally unfortunately, over-particularizing. Explicitly compare the effects detected in your inquiry with the effect sizes reported in related previous studies. Do not be afraid to extend your interpretations to

a general class or population if you have reasons to assume that your results apply. This general class may consist of populations you have studied at your site, or other populations at other sites, or even more general populations. Providing these reasons in your discussion will help you to stimulate future research for yourself and others.

Conclusions

Speculation may be appropriate, but use it sparingly and explicitly. Note the shortcomings of your study. Remember, however, that acknowledging limitations is for the purpose of qualifying results and avoiding pitfalls in future research. Confession should not have the goal of disarming criticism. Recommendations for future research should be thoughtful and grounded in present and previous findings. Gratuitous suggestions ("further research needs to be done ...") waste space. Do not interpret a single study's results as having importance independent of the effects reported elsewhere in the relevant literature. The thinking presented in a single study may turn the movement of the literature, but the results in a single study are important primarily as one contribution to a mosaic of study effects.

Some had hoped that the Task Force would vote to recommend an outright ban on the use of significance tests in psychology journals. Although this might eliminate some abuses, the committee thought that there were enough counterexamples (e.g., Abelson, 1997) to justify forbearance. Furthermore, the committee believed that the problems raised in its charge went beyond the simple question of whether to ban significance tests.

The committee hopes instead that this report will induce editors, reviewers, and authors to recognize practices that institutionalize the thoughtless application of statistical methods. Distinguishing statistical significance from theoretical significance (Kirk, 1996) will help the entire research community to publish more substantial results. Encouraging good design and logic will help improve the quality of conclusions. And promoting modern statistical graphics will improve the assessment of assumptions and display of results.

More than fifty years ago, Hotelling *et al*. (1948) wrote, "Unfortunately, too many people like to do their statistical work as they say their prayers -- merely substitute in a formula found in a highly respected book written a long time ago." Good theories and intelligent interpretation advance a discipline more than rigid methodological orthodoxy. If editors keep in mind Fisher's words quoted above, then there is less danger of methodology substituting for thought. Statistical methods should guide and discipline our thinking, but should not determine it.

Notes

Jacob Cohen died on January 20, 1998. Without his initiative and gentle persistence, this report most likely would not have appeared. Grant Blank provided the Kahn and Udry (1986) reference. Gerard Dallal and Paul Velleman offered helpful comments.

References

Abelson, R.P. (1995). *Statistics as Principled Argument.* Hillsdale, NJ: Lawrence Erlbaum.

Abelson, R.P. (1997). On the surprising longevity of flogged horses: Why there is a case for the significance test. *Psychological Science, 23*, 12-15.

Bailar, J.C., & Mosteller, F. (1988). Guidelines for statistical reporting in articles for medical journals: Amplifications and explanations. *Annals of Internal Medicine, 108*, 266-273.

Baker, B..L. (1969). Symptom treatment and symptom substitution in enuresis. *Journal of Abnormal Psychology, 74*, 42-49.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society* (Series B), 57, 289-300.

Chambers, J., Cleveland, W., Kleiner, B., & Tukey, P. (1983). *Graphical methods for data analysis.* Monterey, CA: Wadsworth.

Chartrand, J.M. (1997). National sample survey. Unpublished raw data.

Cleveland, W.S. (1995). *Visualizing Data*. Summit, NJ: Hobart Press.

Cleveland, W.S., & Devlin, S. (1988). Locally weighted regression analysis by local fitting. *Journal of the American Statistical Association, 83*, 596-640.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences.* New York: Academic Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences. (2nd Ed.)*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1994) The earth is round ($p<.05$). *American Psychologist, 49*, 997-1003

Cook, T.D., & Campbell, D.T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally.

Cronbach, L.J. (1975). Beyond the two disciplines of psychology. *American Psychologist, 30,* 116-127.

Ehrenberg, A.S.C. (1975). *Data reduction: Analyzing and interpreting statistical data*. New York: John Wiley & Sons.

Ehrenberg, A.S.C. (1981). The problem of numeracy. *The American Statistician, 35*, 67-71.

Feldt, L.S., & Brennan, R.L. (1989). Reliability. In R.L. Linn (Ed.), *Educational measurement* (3rd Ed., pp. 105-146). Washington, DC: American Council on Education.

Fisher, R.A. (1935). *The design of experiments*. Edinburgh: Oliver and Boyd.

Fleiss, J.L. (1994). Measures of effect size for categorical data. In Cooper, H., and Hedges, L.V. (Eds.), *The handbook of research synthesis* (pp. 2245-260). New York: Russell Sage Foundation.

Harlow, L.L., Mulaik, S.A., & Steiger, J.H. (1997). *What if there were no significance tests?* Hillsdale, NJ: Lawrence Erlbaum Associates.

Holland, P.W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*, 945-960.

Holland, P. W., & Rubin, D. B. (1983). On Lord's paradox. In H. Wainer & S. Messick (Eds.), *Principals of modern psychological measurement* (pp. 3-25). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hotelling, H., Bartky, W., Deming, W.E., Friedman, M., & Hoel, P. (1948). The teaching of statistics. *The Annals of Mathematical Statistics, 19*, 95-115.

Hurlburt, R.T., & Spiegel, D.K. (1976). Dependence of F ratios sharing a common denominator mean square. *The American Statistician, 20*, 74-78.

Kahn, J.R., & Udry, J.R. (1986). Marital coital frequency: Unnoticed outliers and unspecified interactions lead to erroneous conclusions. *American Sociological Review, 51*, 734-737.

Keselman, H.J., Huberty, C.J., Lix, L.M., Olejnik, S., Cribbie, R., Donahue, B., Kowalchuk, R.K., Lowman, L.L., Petoskey, M.D., Keselman, J.C., & Levin, J.R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research, 68*, 350-386.

Kirk, R.E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56*, 746-759.

Little, R.J.A., & Rubin, D.B. (1987). *Statistical analysis with missing data.* New York: John Wiley & Sons.

McDonald, R.P. (1997). Haldane's lungs: A case study in path analysis. *Multivariate Behavioral Research, 32*, 1-38.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational Measurement* (3rd. Ed., pp. 13-103). Washington, DC: American Council on Education.

Rosenthal, R. (1966). *Experimenter Effects in Behavioral Research.* New York: Appleton-Century-Crofts.

Rosenthal, R. (1994). Parametric measures of effect size. In Cooper, H., & Hedges, L.V. (Eds.), *The handbook of research synthesis* (pp. 231-244). New York: Russell Sage Foundation.

Rosenthal, R., & Rubin, D.B. (1982). A simple general purpose display of magnitude of experimental effect. *Journal of Educational Psychology, 74*, 166-169.

Rosenthal, R., & Rubin, D.B. (1994). The counternull value of an effect size: A new statistic. *Psychological Science, 5*, 329-334.

Rosenthal, R., Rosnow, R.L., & Rubin, D.B. (in press). *Contrasts and effect sizes in behavioral research: A correlational approach*. New York: Cambridge University Press.

Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*, 688-701.

Rubin, D.B. (1976). Inference and Missing Data. *Biometrika, 63,* 581-592.

Rubin, D.B. (1977). Assignment of treatment group on the basis of a covariate. *Journal of Educational Statistics, 2*, 1-26.

Rubin, D.B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics, 6*, 34-58.

Rubin, D.B. (1990a) Neyman (1923) and Causal Inference in Experiments and Observational Studies." *Statistical Science, 5,* 4, pp. 472-480.

Rubin, D.B. (1990b). Formal Modes of Statistical Inference for Causal Effects. *Journal of Statistical Planning and Inference, 25*, 279-292.

Scheines, R., Spirites, P., Glymour, C., Meek, C., & Richardson, T. (1998). The TETRAD Project: Constraint based aids to causal model specification. *Multivariate Behavioral Research, 33,* 65-117.

Schmidt, F. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. *Psychological Methods, 1*, 115-129.

Scott, D.W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization.* New York: John Wiley & Sons.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. New York: Chapman & Hall.

Snyder, P., & Lason, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. *Journal of Experimental Education, 61*, 334-349.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher, 25*(2), 26-30.

Thompson, B., & Snyder, P.A. (1998). Statistical significance and reliability analyses in recent *JCD* research articles. *Journal of Counseling and Development, 76*, 436-441.

Tukey, J.W. (1977). *Exploratory data analysis.* Reading, MA: Addison-Wesley.

Wainer, H. (1999). One cheer for null hypothesis significance testing. *Psychological Methods, 4*(2), xx-xx.

Webb, E.J., Campbell, D.T., Schwartz, R.D., & Sechrest, L. (1966). *Unobtrusive measures: Nonreactive research in the social sciences.* Chicago: Rand McNally.