

CORRESPONDENCE

Dear Editor:

In the May 2000 issue of *JSA*, Thomas Gray (Gray, 2000) calls our attention to a report by L. Wilkinson (Wilkinson, 1999) that represents the deliberations of the distinguished 12-member Task Force on Statistical Inference, American Psychological Association Board of Scientific Affairs. Some of the issues raised by Wilkinson are the weakness of the null hypothesis (H_0) and the improper use and interpretation of it, and the use of other procedures, such as the confidence interval, that do a better job. Gray points out that there has been a slight improvement over time in *JSA* with regard to the reporting of reliability statistics and an unfortunate increase in statistical significance testing (SST, also known as null hypothesis significance testing, NHST) in assessing research. Because of *JSA*'s use of SST, it is necessary to understand the limitations of SST and to examine alternatives. Before I proceed, let it be said that not everyone agrees with the Wilkinson committee, including some of its members. SST and NHST are used interchangeably in the following discussion.

A major concern in the Wilkinson article is the nature of H_0 and the failure to realize what it cannot do. Kirk (1996) provides the three most often cited criticisms: "Null hypothesis significance testing and scientific inference address different questions. In scientific inference, what we want to know is the probability that the null hypothesis (H_0) is true given that we have obtained a set of data (D); that is $p(H_0/D)$. What null hypothesis significance testing tells us is the probability of obtaining these data or more extreme data if the null hypothesis is true, $p(D/H_0)$." The one, $p(D/H_0)$, does not imply the other, $p(H_0/D)$. "A second criticism of null hypothesis significance testing is that it is a trivial exercise. . . . Because the null hypothesis is always false, a decision to reject it simply indicates that the research design had adequate power to detect a true state of affairs, which may or may not be a large effect or even a useful effect. . . . A third criticism of null hypothesis significance testing is that by adopting a fixed level of significance, a researcher turns a continuum of uncertainty into a dichotomous reject-do-not-reject decision" (pp. 747-748). The first and third criticisms seem clear, but the second is something many of us have never considered. It requires elaboration, and the reader is directed to Kirk (1996). All of the following material concerns SST and NHST (i.e., the null hypothesis, H_0).

Some experts have vehement feelings about the use of SST. Rozeboom, (1997, p. 335) is a colorful example: "Null-hypothesis testing is surely the most bone-headedly misguided procedure ever institutionalized in the rote training of science students." Others (e.g., Cohen, 1997) have similar opinions that are probably as strongly felt, but Rozeboom surely gets our attention. Since I also am addicted to hyperbole, I enjoy Rozeboom's comments.

However, we might profit from reading one 10-line paragraph on the history of SST in Rothman (1986, pp. 115-116), after which we might not see it as so boneheaded. Even Cohen (1990) acknowledges support for H_0 when discussing its origins in agriculture. He writes (p. 1307), "[It] offered a decision scheme, mechanical, and objective, independent of content, and led to clear-cut yes/no decisions. . . . The outcome of an experiment can quite properly be a decision to use this rather than that amount of manure. . . . But we do not deal in manure, at least not knowingly." Yates (1951) offers a calm and competent analysis of the problems of statistical significance testing (p. 32), ending with, "Tests of significance are preliminary and/or ancillary" (p. 33), implying that much more is required.

The adjuncts and alternatives offered to SST are confidence intervals, effect size, power analysis, and Bayesian Inference. Strangely, all of these depend to some extent on "misguided . . . institutionalized . . . rote training" (i.e., $\alpha = 0.05$, or some such value). The Wilkinson committee wants us to be judicious in our use of α and NHST. Allow me to discuss each of these alternatives in a bit more detail.

Confidence intervals

The confidence interval (CI) suggests a solution because it implies that additional samples will produce the parameter of interest (e.g., the sample mean) within the limits of the interval ($1 - \alpha$), where α could be 0.05, hence the CI would be 0.95. We do not know where in the CI the parameter is located, or if in fact it is there at all, but we are "confident" to the extent of 95%. If the interval does not contain zero, we can interpret this as the traditional SST. Concisely, the advantages of CIs are: "Confidence intervals convey information about magnitude and precision of effect simultaneously, keeping these two aspects of measurement closely linked. The use of p values, or 'significance' testing, blurs these concepts so that the focus on measurement is lost" (Rothman, 1986, p. 121). The combination of NHST and confidence interval uses the same units of measurement as the data and therefore make trivial results difficult to ignore (Kirk, 1996).

Effect size

Effect size (Cohen, 1962, 1988) controls for the fact that a large enough sample will reject H_0 even when the effect is so small that it has neither social nor scientific importance. The effect size (ES) thus emphasizes social and scientific importance. The ES is the alternative hypothesis. It specifies in advance how large a result (e.g., the difference between a treatment and a con-

trol) will be acceptable to the researcher. Also, if an ES is provided in a report, readers can then decide how important the result is to them.

How does one determine effect size? Personal experience, an examination of the literature, meta-analysis, intuition, estimations from theory and accepted convention about what represents small, medium and large effects are the general approaches (Cohen, 1988, 1997; Murphy and Myers, 1998). For example, Cohen (1988, 1992) suggests ESs of $d = .2, .5$ and $.8$ as conventions for small, medium and large effect sizes for t tests. The merits of effect size are discussed by Kirk (1996, pp. 749-750), including the interpretation of the conventions and other uses of them, such as estimating sample size. Particularly helpful are Kirk's formulas for computing "effect magnitude" (defined below) from the t and F tests (Kirk, Table 3, columns 2 and 4). Misinterpretation of the meaning of rejection of H_0 has been endemic among users of t and F and a source of severe criticisms of H_0 .

Also very accessible is Cohen's (1992) "A Power Primer" article in which he laments the lack of power analyses in research literature, gives an overview of the effect size construct and then provides two tables along with descriptive text. Table 1 presents computation of eight (8) ES indexes (d, r, q, g, h, w, f and f^2) and their values for small, medium and large effects. In addition, Cohen's Table 2 provides the required sample sizes for Power = .80 at three levels of α (.01, .05, .10) and three levels of effect size (small, medium, large) (see "Power Analysis" below). Cohen's Table 2 provides the above information for each of the eight tests of Table 1, along with 1 through 6 degrees of freedom for chi-square, 2 through 7 groups for ANOVA and 2 through 8 independent variables for multiple correlation.

Kirk (1996) discusses a related matter, *effect magnitude*, which includes measures of effect size, strength of association and some others. He lists a large number of such indicators and provides algebraic conversions among various measures of effect magnitude and guides to their interpretation. It should be noted that new students are often taught statistical "tests," such as the Pearson Product Moment Correlation r and the t test. In reality, r is an estimate of strength of association (and can be considered an ES), whereas the t test is the statistical test used to evaluate whether the r is itself "significant"; and when we report the results of a t test, it would behoove us to provide an ES (r, d , something) or confidence limits. Kirk (pp. 748-749) points out the need for strength of association measures, citing Peters and Van Voorhis (1940): "The F and z tests employed with the analysis of variance do not directly indicate the strength of the relation that is present, but only its reliability." In other words, the problem has been known for a long time (see Keppel, 1991; Yates, 1951).

Strength of association may be thought of as the amount of variance of the dependent variable accounted for by the independent variable or, more generally, the variance shared among variables. One measure of this is the analyses of variance components (Scheffé, 1959), which for simple experimental designs, even outside of agricultural experiments, are easy to calculate and interpret (Keppel, 1991; Scheffé, 1959).

"Big effects are more impressive than small effects" (Abelson, 1995, p. 46), and effects expressed in the raw response units can have immediate meaning to the investigator. Raw response units are those in which the original measurements were made (e.g., inches, percent, mg/dl). There are, of course, disadvantages to

raw response units in expressing effect sizes, in which case the standardized unit is preferred. An example of a standardized unit would be the difference between the control and the treated divided by the common standard deviation (see Kirk, 1996, pp. 750-751, for definitions of the standard deviation). Thus, the standardized effect size is independent of the original response scale, and can be compared to other response scales similarly transformed. The difference between original and standardized units is the difference between concrete and abstract and which is used depends on one's goal.

Power analysis

Power is defined as $1 - \beta$, where β is the probability of making a Type II error (acceptance of H_0 when it is false). Thus, Power is the probability of avoiding a Type II error (Power "is the long-run frequency of acceptance of H_1 if H_1 is true"; Sedlmeier and Gigerenzer, 1989, p. 309). H_1 is the alternative hypothesis. Power analysis is usually performed at the start of research but it is also useful afterwards. It has some distinct advantages, one of which is that it can be used to test for minimum effects, thereby avoiding Kirk's second criticism of H_0 and also it can be used to test H_0 . The minimum-effect hypothesis tests the smallest treatment difference that the investigator will accept. Power is itself related to sensitivity, effect size and decision criterion. Any three of these four variables can be used to determine the missing element. Sensitivity is a function of sample size (e.g., number of subjects), quality of measures, and research design. Sample size is generally cited as the most important: the larger the sample, the more likely the study is to find significant effects. Power also increases as effect size increases. Decision criteria refers to the "level. The more lenient the decision criteria (e.g., 0.05 is more lenient than 0.01), the greater the power. For a detailed presentation of power analysis see Murphy and Myers (1998), and for an accessible and easily tracked presentation, as well as a ready made table for quick power evaluations at the conventionally acceptable power level of .80, see Cohen (1992). Another criterion measures the percentage of variance accounted for by the independent variable (PV). There are a number of these, appropriate to different statistics. Murphy and Myers suggest that power should be above .50 and power of .80 or above is adequate. Nevertheless, these power settings are arbitrary, as is $\alpha = .05$. Their Table 2.2 (p. 47) and the accompanying discussion provide approximate equivalents of their PV for small, medium and large effects in terms of three conventional criteria, r, d and f^2 . $R^2, \text{adj. } R^2$ and ω^2 also are measures of the proportion of variance accounted for (Maxwell and Delaney, 1990).

Bayesian inference

"Bayesian statistical methods support inferences without reference to either significance tests or confidence intervals. . . . This class of methods entails the use of prior information and empirical data to generate posterior distributions that in turn serve as the basis for statistical inference" (Pruzek, 1997, p. 287). Pruzek offers an elementary example and walks us through calculations that are simple and understandable. The Bayesian analysis requires three steps: (1) generation of a prior distribution; (2) collection of empirical data; (3) combining the prior distribution with

the likelihood distribution (based on the empirical data) to form a posterior distribution, which is used to make inferences about the parameter (e.g., mean) of interest. The posterior distribution represents the locus of the parameter being investigated. The posterior distribution has a *credible interval*, two points on the horizontal axis, that looks like a confidence interval except that the credible interval represents the location of the *population* parameter with a determined (e.g., 95%) probability (Rindskopf, 1997, p. 322). The confidence interval, on the other hand, tells us where the *sample* estimate of a parameter lies given our idea of the value of the population parameter with a 95% probability (Mulaik et al., 1997, p. 71). It seems to me that the *credible interval* tells us where the population parameter lies, while the *confidence interval* tells us where future sample estimates will be.

Bayesian statistics have always left me mystified and doubtful. The doubt comes from the use of prior information, which is described as subjective. The subjective idea seems to run counter to the purpose of statistics, which is to supply objectivity in forming conclusions about the outcome of research. But the subjective part does not have to be too subjective. Pruzek's analysis suggests that the source of one's subjective opinion might be something along the lines of effect size, prior experience with similar work, expert opinion, meta-analysis and so forth. It is what is done with this information that is different. One estimates values for the parameter of interest and for each of two constants which represent the range of the parameter. The mean of the subjective distribution and its variance are provided by simple algebra from these estimates. The prior distribution (subjective estimates of the parameters) is combined with the likelihood distribution (the empirical data), again by simple algebra, which leads to the posterior distribution, with its new mean and variance. The posterior distribution is supposed to supply better estimates than results based solely on the original empirical data. A true and impressive application of the Bayesian approach appears in Sontag et al. (1998, pp. 82-85): a nuclear bomb lost in the Mediterranean Sea was found when traditional approaches failed.

Harlow (1997) presents a survey of opinions put forth in the 14 chapters of *What If There Were No Significance Tests?* Her survey sounds like the development of the prior distribution. She then summarizes: "Focusing on a dichotomous decision would contribute little to the development of strong theories or sound judgement; lacks the precision of either confidence interval, effect sizes or power calculations; is less informative than goodness of approximation assessment (procedures used in structural equation modeling, SEM) or the use of specific, realistic and nonzero hypotheses; and is less thorough than either replication, meta-analysis or Bayesian inference. . . . It seems, the overriding view on this issue is that NHST may be overused and unproductive, particularly when used as a simple dichotomous decision rule" (p. 12).

The Other Side of the Argument

Once upon a time the chairman of a Department of Sociology said to me, "We can always depend on Professor X to come down firmly on both sides of any issue." The opinions of the NHST opponents are not universal. Abelson, who was a member of the group that produced the Wilkinson report, points out (1997a) that NHST is useful in tests of goodness of fit to models and

should not be abandoned for this application. (For more on various issues related to NHST, see Abelson, 1995.) Abelson (1997b, p. 117) states succinctly, "(1) Although bad practice certainly has characterized some significance testing, many of the critics of significance tests overstate their case by concentrating on such bad practice, rather than providing a balanced analysis; (2) Proposed alternatives to significance testing, especially meta-analysis, have flaws of their own; (3) Significance tests fill an important need in answering some key research questions, and if they did not exist they would have to be invented."

Hagen (1997, p. 22) puts it this way: "The logic of the NHST is elegant, extraordinarily creative, and deeply embedded in our methods of statistical inference. It is unlikely that we will ever be able to divorce ourselves from that logic even if someday we decide that we want to. . . . The NHST has been misinterpreted and misused for decades. This is our fault, not the fault of NHST. I have tried to point out that the NHST has been unfairly maligned; that it does, indeed, give us useful information; and that the logic underlying statistical significance testing has not been successfully challenged." But that is not the challenge to NHST. The challenge is to the abbreviated, single use of it, and its consequent misinterpretation.

Others have arguments against the elimination of NHST, for example Chow (1988). Rindskopf (1997), despite being a Bayesian, says that the null hypothesis continues to be used because researchers are "testing approximately the right thing under many real circumstances, even if most researchers do not know the rationale" (p. 321). This statement may be damning with faint praise, but to me it is more effective than Rozeboom's histrionics.

It is often suggested that the solution to these problems is computers. Computers are wonderful for reducing computational drudgery. There are, however, several difficulties with the computer solution: (1) one still needs to understand one's statistics; (2) one still has to choose the appropriate analysis; (3) one still needs to understand the statistical packages; and (4) one still needs to prove to oneself that the statistical program is correct. In the end (5) one is still obligated to interpret one's results correctly. On this point, it has been my experience as faculty member and referee that some researchers, including psychologists, use very simple experimental designs, but do not analyze them correctly. This may be a function of using computer packages. (The writers of computer manuals write to impress their colleagues, not to make the packages easy to use.) For example, psychologists like to use the counterbalanced design (called the crossover design by Cochran and Cox, 1950), which they often analyze as a simple two-way, repeated measure design. Such an analysis vitiates the values of the counterbalancing and leads to incorrect results when the reason for counterbalancing proves to be significant. The rule is that every design has a fixed and appropriate analysis. Deviation from that analysis can only be successfully done when one understands the design and knows what one is doing. There are criteria for deviations (see Hays, 1988).

The *Journal of Studies on Alcohol* has a range of disciplines larger than that projected by the Wilkinson Report. The report emphasizes correlational analysis, because the committee's interest is in social psychology and in areas of direct application, such as clinical psychology. Only passing reference is made to experimental research. Very often the social significance of experimental research is of only distant concern to the researcher. That comes

later, when the phenomenon being studied is well described, and there is enough data to obtain an effect size by any of the methods suggested. Some of these disciplines from which the *Journal of Studies on Alcohol* receives manuscripts have their own needs and traditions in the use of statistics, all of which need to be considered in evaluating their research.

Conclusion

The Wilkinson Report covers a large number of issues, but attention here was limited to SST because of its complexity, the long-running controversy surrounding it and its importance in assessing research results. Having said that, we should attend seriously to what the Wilkinson committee proposes and the arguments of those who are against the elimination of SST. It is true that users of SST can polish up their act by a few simple operations. With this in mind we make the following recommendations and suggest that associate editors and referees insist on compliance when monitoring manuscripts:

- NHST should not be abandoned but should be accompanied by confidence limits. Effect size, power analysis and Bayesian posterior analysis are alternative measures and always acceptable when properly used.
- Where possible, measures of the strength of association should be reported.
- When effect sizes are used in the planning of research, their source and justification should be reported.
- Whichever index of effect size is used, it should be interpreted briefly and correctly for the reader.
- This last, although not covered in detail in this letter, is an important point made by Gray (2000) and should be included here in recommendations: Reliability statistics (psychometric properties) should be evaluated on the research sample for assessments and tests used in the research.

Finally, two caveats are worth quoting: "No statistical method always works" and "Statistical methods are better conceived as options than as commandments" (Abelson, 1997a, p. 14).

References

- ABELSON, R.P. *Statistics as Principled Argument*, Mahwah, NJ: Lawrence Erlbaum, 1995.
- ABELSON, R.P. On the surprising longevity of flogged horses: Why there is a case for the significance test. *Psychol. Sci.* **8**: 12-15, 1997a.
- ABELSON, R.P. A retrospective on the significance test ban of 1999 (If there were no significance tests, they would be invented). In: HARLOW, L.L., MULAİK, S.A. AND STEIGER, J.H. (Eds.) *What If There Were No Significance Tests?* Mahwah, NJ: Lawrence Erlbaum, 1997b, pp. 117-141.
- CHOW, S.L. Significance test or effect size? *Psychol. Bull.* **103**: 105-110, 1988.
- COCHRAN, W.G. AND COX, G.M. *Experimental Designs*, New York: John Wiley & Sons, 1950.
- COHEN, J. The statistical power of abnormal-social psychological research: A review. *J. Abnorm. Social Psychol.* **65**: 145-153, 1962.
- COHEN, J. *Statistical Power Analysis for the Behavioral Sciences*, 2nd Edition, Mahwah, NJ: Lawrence Erlbaum, 1988.
- COHEN, J. Things I have learned (so far). *Amer. Psychol.* **45**: 1304-1312, 1990.
- COHEN, J. A power primer. *Psychol. Bull.* **112**: 155-159, 1992.
- COHEN, J. The earth is round ($p < .05$). In: HARLOW, L.L., MULAİK, S.A. AND STEIGER, J.H. (Eds.) *What If There Were No Significance Tests?* Mahwah, NJ: Lawrence Erlbaum, 1997, pp. 21-35.
- GRAY, B.T. Correspondence. *J. Stud. Alcohol* **61**: 487-488, 2000.
- HAGEN, R.L. In praise of the null hypothesis statistical test. *Amer. Psychol.* **52**: 15-24, 1997.
- HARLOW, L.L. Significance testing introduction and overview. In: HARLOW, L.L., MULAİK, S.A. AND STEIGER, J.H. (Eds.) *What If There Were No Significance Tests?* Mahwah, NJ: Lawrence Erlbaum, 1997, pp. 1-17.
- HAYS, W.L. *Statistics*, 4th Edition, New York: Holt, Rinehart and Winston, 1988.
- KEPPEL, G. *Design and Analysis: A Researcher's Handbook*, 3rd Edition, Upper Saddle River, NJ: Prentice Hall, 1991.
- KIRK, R.E. Practical significance: A concept whose time has come. *Educ. Psychol. Meas.* **56**: 746-759, 1996.
- MAXWELL, S.W. AND DELANEY, H.D. *Designing Experiments and Analyzing Data: A Model Comparison Perspective*, Belmont, CA: Wadsworth, 1990.
- MULAİK, S.A., RAJU, N.S. AND HARSHMAN, R.A. There is a time and a place for significance testing. In: HARLOW, L.L., MULAİK, S.A. AND STEIGER, J.H. (Eds.) *What If There Were No Significance Tests?* Mahwah, NJ: Lawrence Erlbaum, 1997, pp. 65-115.
- MURPHY, K.R. AND MYORS, B. *Statistical Power Analysis: A Simple and General Model for Traditional and Modern Hypothesis Testing*, Mahwah, NJ: Lawrence Erlbaum, 1998.
- PETERS, C.C. AND VAN VOORHIS, W.R. *Statistical Procedures and their Mathematical Bases*, New York: McGraw-Hill, 1940.
- PRUZEK, R.M. An introduction to Bayesian inference and its applications. In: HARLOW, L.L., MULAİK, S.A. AND STEIGER, J.H. (Eds.) *What If There Were No Significance Tests?* Mahwah, NJ: Lawrence Erlbaum, 1997, pp. 287-318.
- RINDSKOPF, D.M. Testing "small," not null, hypotheses: Classical and Bayesian approaches. In: HARLOW, L.L., MULAİK, S.A. AND STEIGER, J.H. (Eds.) *What If There Were No Significance Tests?* Mahwah, NJ: Lawrence Erlbaum, 1997, pp. 319-332.
- ROTHMAN, K.J. *Modern Epidemiology*, Boston: Little, Brown, 1986.
- ROZEBOOM, W.W. Good science is adductive, not hypothetico-deductive. In: HARLOW, L.L., MULAİK, S.A. AND STEIGER, J.H. (Eds.) *What If There Were No Significance Tests?* Mahwah, NJ: Lawrence Erlbaum, 1997, pp. 335-391.
- SCHEFFÉ, H. *The Analysis of Variance*, New York: John Wiley & Sons, 1959.
- SEDLMEIER, P. AND GIGERENZER, G. Do studies of statistical power have an effect on the power of studies? *Psychol. Bull.* **105**: 309-316, 1989.
- SONTAG, S., DREW, C. AND DREW, A.L. *Blind Man's Bluff: The Untold Story of American Submarine Espionage*, New York: Public Affairs, 1998.
- WILKINSON, L. Statistical methods in psychology journals: Guidelines and explanations. *Amer. Psychol.* **54**: 594-604, 1999.
- YATES, F. The influence of *Statistical Methods for Research Workers* on the development of the science of statistics. *J. Amer. Stat. Assoc.* **46**: 19-34, 1951.

JOHN A. CARPENTER, PH.D.
Center of Alcohol Studies
Rutgers University
Piscataway, NJ

Editor's note: Dr. Carpenter, who served as both editor and editor emeritus to the *Journal of Studies on Alcohol*, passed away on February 27, 2001.