Journal of Organizational Behavior
J. Organiz. Behav. **28**, 647–659 (2007)
Published online 27 April 2007 in Wiley InterScience
(www.interscience.wiley.com) **DOI**: 10.1002/job.448

Invited Essay

Comments and recommendations regarding the hypothesis testing controversy

DOUGLAS G. BONETT[1]* AND THOMAS A. WRIGHT[2]

[1]*Departments of Statistics and Psychology, Iowa State University, Ames, Iowa 50011, U.S.A.*
[2]*Managerial Sciences Department, University of Nevada, Reno, Nevada 89557-0206, U.S.A.*

Summary Hypothesis tests are routinely misinterpreted in scientific research. Specifically, the failure to reject a null hypothesis is often interpreted as support for the null hypothesis while the rejection of a null hypothesis is often interpreted as evidence of an important finding. Many of the most frequently used hypothesis tests are ''non-informative'' because the null hypothesis is known to be false prior to hypothesis testing. We discuss the limitations of non-informative hypothesis tests and explain why confidence intervals should be used in their place. Several examples illustrate the use and interpretation of confidence intervals. Copyright © 2007 John Wiley & Sons, Ltd.

Introduction

''Whenever possible, the basic statistical report should be in the form of a *confidence interval*.''
William W. Rozeboom, *Psychological Bulletin*, 1960, Vol. 57: 426.

The value of hypothesis testing methods has long been questioned (Boring, 1919), with the frequency and intensity of these criticisms increasing in recent years (Harlow, Mulaik, & Steiger, 1997; Kline, 2004; Morrison & Henkel, 1970). At the extreme, this debate has culminated in recommendations to completely ban hypothesis testing from scientific journals (Gardner & Altman, 1986; Hunter, 1997). This hypothesis testing controversy has been more contentious in some academic disciplines than others. For instance, medical researchers quickly accepted, with little controversy or debate (Altman, Machin, Bryant, & Gardner, 2000, p. 7), the recommendations to report confidence intervals rather than hypothesis tests, while the controversy rages on among social scientists. Currently, there is considerable variability in the editorial policies of social science journals regarding the reporting of hypothesis tests and *p*-values.

In comparison to both medical and social science researchers, business researchers have been mostly apathetic with regard to the hypothesis testing controversy, with article acceptance decisions in leading

* Correspondence to: Douglas G. Bonett, Department of Statistics, Iowa State University, Ames, Iowa 50011, U.S.A.
E-mail: dgbonett@iastate.edu

Received 2 February 2006
Revised 28 November 2006
Accepted 20 February 2007

management journals continuing to rely on the results of hypothesis testing methods. Not withstanding the sage advice from Rozeboom (1960) almost 50 years ago, many management researchers remain unfamiliar with the use and interpretation of confidence intervals. For instance, out of more than 130 empirical articles published in 2003 and 2004 in both the *Academy of Management Journal* and *Administrative Science Quarterly*, only one article reported a confidence interval. The reliance on statistical hypothesis testing methods in scientific research is a serious problem. Rothman (1986, p. 445) well articulated both the seriousness and deep-seated nature of this problem in noting that, ''Testing for significance continues today not on its merits as a methodological tool but on the momentum of tradition. Rather than serving as a thinker's tool, it has become for some a clumsy substitute for thought, subverting what should be a contemplative exercise into an algorithm prone to error.''

Many of the arguments given by many social scientists favoring the continued use of hypothesis testing are based on various misunderstandings. Foremost among these involve misunderstandings regarding populations and population parameters, a failure to distinguish between informative and non-informative hypothesis tests, and confusion regarding the interpretation of confidence intervals. In this article we will: (1) review the basic concepts of a population and a population parameter; (2) make a distinction between informative and non-informative hypothesis tests; (3) explain how to interpret a confidence interval; (4) give several examples to illustrate how confidence intervals may be used in place of hypothesis testing methods; (5) discuss recently proposed ''solutions'' to the hypothesis testing problem that should be avoided; (6) recommend publication guidelines regarding the effective use of confidence intervals and the inappropriate use of hypothesis testing. We begin our discussion with some basic definitions and ideas regarding populations and population parameters.

Populations

A population is any well-defined collection of units. In business applications, the units are typically employees, customers, products, or firms. For instance, a population could be 5214 employees in a particular organization, or 24 648 customers who purchased a specific product, or 2000 cordless phones produced in a recent production run, or the set of Fortune 1000 companies. Management and organizational researchers often study populations of employees, market researchers often study populations of customers, operations managers often study populations of products, and economists often study populations of firms. Since many readers of the *Journal of Organizational Behavior* will be primarily interested in populations of employees, employee populations will be used in all of our examples.

In applications where it is impractical or too costly to measure every unit in the population, it might be feasible to measure a random sample of units from the population. The population from which the random sample is taken is called the *study population*. The study population might be all 920 managers who work for an insurance company in California, all 1000 CEOs of Fortune 1000 companies, or all 958 sales workers who work for a large manufacturing firm in Atlanta, Georgia. Using inferential statistical methods, information from a random sample can then be used to make certain types of statements about the study population. Logical arguments are then needed to make inferences from the study population to other populations of interest.

Population Parameters

A population parameter is an unknown number that describes some characteristic of a study population. The mean, median, proportion, variance, correlation, and slope are commonly used to describe a study

DOI: 10.1002/job

population. Let θ (theta) denote the numerical value of a study population parameter. For instance, θ might represent the *mean* job performance score in a study population of sales workers, or the *proportion* of all human resource managers in a study population who are dissatisfied with their job, or the Pearson *correlation* between employee training and job performance in a study population of first-year employees. In applications where the population may be partitioned into two or more *subpopulations*, we let θ_j denote a population parameter for subpopulation j. For instance, θ_1 and θ_2 might represent the mean job performance ratings for a subpopulation of 1950 employees who have less than 2 years of college education and a subpopulation of 1340 employees who have 2 or more years of college education, respectively. In applications where a single sample has been taken from the study population, and then randomly divided into two or more groups with each group receiving a different treatment, θ_j then represents the parameter value that would occur if *all* members of the study population had received treatment j. Note that the interpretation of θ_j is conceptually different for subpopulations and treated populations and this must be taken into consideration when interpreting a confidence interval that involves θ_j.

Different types of hypotheses may be stated regarding the unknown values of the study population parameters and a wide variety of statistical methods are available to test these hypotheses. Although there is controversy surrounding the use of hypothesis tests, not all hypothesis tests have serious limitations and it is necessary to make a distinction between non-informative and informative hypothesis tests.

Non-Informative Hypothesis Tests

Non-informative hypothesis tests do not provide new information regarding the unknown parameter values of a study population. Most researchers will be surprised to learn that many commonly used hypothesis tests are non-informative tests. A few examples of non-informative hypothesis tests are the F-tests in ANOVA, ANCOVA, and regression models, and the chi-square tests of homogeneity, independence, and goodness-of-fit. In general, F-tests with numerator degrees of freedom greater than one and chi-square tests with degrees of freedom greater than one are used in non-informative hypothesis tests.

To better understand why certain hypothesis tests are non-informative, consider the following null and alternative hypotheses that are tested using a one-way ANOVA:

$$H_0 : \theta_1 = \theta_2 = \ldots = \theta_k$$

$$H_1 : \theta_j \neq \theta_{j'} \quad \text{for at least one } j, j' \text{ pair}$$

where θ_j is a subpopulation mean or a population mean under treatment j. For instance, suppose θ_j is the mean job performance if all employees in the study population had been trained using one of four goal-setting techniques. The null hypothesis for this specific example is: H_0: $\theta_1 = \theta_2 = \theta_3 = \theta_4$. In other words, this null hypothesis states that all four population parameter values are *exactly* equal. In our example where job performance is compared across four treatment conditions, we are virtually certain that the values of θ_1, θ_2, θ_3 and θ_4 are not all exactly equal. In applications such as this one, we are virtually certain that H_0 must be false and therefore we are virtually certain that H_1 must be true. The ANOVA F-test, which would be used in this example to decide if H_0 should be rejected, is a non-informative test because we already know that H_0 is false. Savage (1957, p. 324) was referring to non-informative hypotheses when he said: ''Null hypotheses of no difference are usually known to be

DOI: 10.1002/job

false before the data are collected . . . when they are, their rejection or acceptance simply reflects the size of the sample and the power of the test, and is not a contribution to science.'' All of the examples of non-informative hypothesis tests listed above can be expressed in a form in which the null hypothesis is a hypothesis of no difference.

Not all hypothesis tests are non-informative. For instance, consider the following null and alternative hypotheses:

$$H_0 : \theta \leq h$$

$$H_1 : \theta > h$$

where h is some specified number and θ is a study population parameter. It is common to set $h = 0$ when θ is a correlation or a slope coefficient. Hypotheses of this general type are referred to as *directional hypotheses* and are described in most introductory statistics texts. For instance, suppose that θ is the population correlation between organizational commitment and job performance and the researcher wants to test a theory that predicts a positive correlation between these two variables. In this application, the researcher could test H_0: $\theta \leq 0$ against H_1: $\theta > 0$. In typical applications the researcher does not know with certainty if $\theta \leq 0$ or if $\theta > 0$. A statistical test that allows the correct selection of H_1 with high confidence would provide useful information. As an additional example, suppose θ is the proportion of staff employees at a large hospital who are dissatisfied with their job and h is set by the researcher at 0.25. The researcher does not know with certainty if $\theta \leq 0.25$ or if $\theta > 0.25$ and a statistical test that allows the correct selection of H_1 with high confidence would provide useful information.

Another type of informative hypothesis test is the *directional two-sided test* (Kaiser, 1960). Consider the following hypotheses

$$H_0 : \theta_1 - \theta_2 = h$$

$$H_1 : \theta_1 - \theta_2 > h$$

$$H_2 : \theta_1 - \theta_2 < h$$

where h is some specified number (usually zero) and θ_j is a study population parameter. Suppose that θ_1 represents the mean problem solving score for a study population of female managers, θ_2 represents the mean problem solving score for a study population of male managers, and h is set at zero. Although we know in advance that H_0 is almost certainly false, we do know with complete certainty if H_1 is true or if H_2 is true. A statistical test that allows us to correctly select H_1 or H_2 with high confidence would provide useful information. Many commonly used tests, such as the one-sample t-test, independent-samples t-test, paired-samples t-tests, and t-tests for regression coefficients, may be used to test directional two-sided hypotheses. However, if only the p-value of these tests is reported, the test becomes non-informative because the p-value alone does not allow one to select H_1 or H_2.

Misinterpretation Errors in Non-Informative Hypothesis Testing

Tests of non-informative hypotheses are routinely misinterpreted. In studies where H_0 is not rejected, researchers often interpret the result as evidence that H_0 is true. For instance, in our example above, where θ_j is the study population mean job performance score for goal-setting technique j, a failure to reject H_0 does not indicate that the four goal-setting techniques are equally effective. In fact, a failure to reject H_0 will occur with high probability if the sample size is too small. This type of misinterpretation is very common in diagnostic tests of normality and equality of variance, where a failure to reject the null hypothesis is often interpreted to imply that the normality or equal variance assumption has been

DOI: 10.1002/job

satisfied. Other common misinterpretations of this type are the goodness-of-fit tests in covariance structure models and contingency tables where a failure to reject H_0 is often interpreted as support for the model implied by the null hypothesis.

A second common type of misinterpretation of non-informative tests occurs when the test rejects H_0 because the *p*-value is less than some value, such as 0.05. As a consequence, the results are often declared to be ''significant'' or ''highly significant'', depending on the size of the *p*-value. The results are then interpreted as though an important finding has been obtained. However, a rejection of H_0 simply indicates that the null hypothesis is false, a fact that was known *prior* to hypothesis testing. Furthermore, because the null hypothesis in a non-informative test is almost certainly false, the *p*-value can be made as small as desired by taking a sufficiently large sample size.

A Confidence Interval Example

We now present a simple hypothetical example to illustrate the type of useful information that can be obtained from a confidence interval, but cannot be obtained from a hypothesis test (either informative *or* non-informative). Suppose a random sample of 50 supervisors is obtained from a study subpopulation of 3150 college-educated supervisors, and a second random sample of 50 supervisors is obtained from a study subpopulation of 5,872 supervisors with high-school only education. All 100 supervisors are given an organizational commitment questionnaire that is scored on a 1–7 scale. The mean commitment score for the college-educated supervisors is 3.8 ($SD = 1.2$) and the mean score for high-school only supervisors is 4.6 ($SD = 1.0$). An independent-samples *t*-test could be used to analyze the sample data. The results of this hypothetical study might be reported as: ''Supervisors with only a high-school education exhibit greater commitment to the organization than supervisors with college educations, $t(48) = 2.45$, $p < 0.05$.''

The results of this directional, two-sided test suggest that supervisors with a college education are, on average, less committed to their organization than supervisors who do not have a college education. However, the *p*-value and sample means do not describe the *magnitude* of the difference in the two study *subpopulations*. A two-sided confidence interval can provide this critical information. A two-sided confidence interval result might be reported as: ''We are 95 per cent confident that the mean organizational commitment score for all 3150 supervisors without a college education is 0.5 to 3.8 points greater than the mean organizational commitment score for all 5872 supervisors who have a college education.'' The *t*-test provides information about the direction of the difference in population means whereas the confidence interval provides information about both the direction and the magnitude of the difference in population means.

The additional information provided by a confidence interval can be made clear by considering several different patterns of confidence interval results. Suppose that a difference less than 0.25 between any two population means on the organizational commitment scale is considered by experts to be trivial and scientifically unimportant. Consider the following three situations and how the confidence interval results would differ from hypothesis testing results. *Situation 1*: suppose the 95 per cent confidence interval ranged from 0.03 to 0.18. The hypothesis testing result would be ''statistically significant'' (because the confidence interval does *not* include zero), but the confidence interval result would indicate that the difference in population means is small and unimportant. *Situation 2*: now suppose the 95 per cent confidence interval ranged from 1.1 to 1.9. The hypothesis testing result would again be ''statistically significant'' but the confidence interval results would indicate that the difference in population means is large and important. *Situation 3*: suppose the 95 per cent confidence interval ranged from -0.04 to 0.16. The hypothesis testing result would be ''inconclusive'' (because the

DOI: 10.1002/job

confidence interval does include 0), but the confidence interval results would clearly indicate that the difference in population means is small and unimportant.

Interpreting A Confidence Interval

A two-sided confidence interval consists of a lower limit and an upper limit that captures the true value of a population parameter with a specified level of confidence. To understand the interpretation of these limits, imagine that a $100(1 - \alpha)$ per cent confidence interval for θ has been computed from many different random samples from the same study population. Statistical theory tells us that about $100(1 - \alpha)$ per cent of these confidence intervals will capture the true value of θ, assuming that certain assumptions have been satisfied. Of course, in practice only one random sample will be obtained from the study population, but given the known theoretical behavior of a confidence interval, we can be $100(1 - \alpha)$ per cent confident that our one confidence interval has captured the true value of θ. This idea generalizes to differences in population parameters or functions of two or more population parameters. We reported the confidence interval result in the previous example as: ''We are 95 per cent confident that the mean organizational commitment score for all 3150 supervisors without a college education is 0.5 to 3.8 points greater than the mean organizational commitment scores for all 5872 supervisors who have a college education.'' Imagine repeating this study many thousands of times in the same two study subpopulations. We know from statistical theory that about 95 per cent of the confidence intervals for the difference in subpopulation means will capture the true difference in subpopulation means. Knowing this, we say that we are 95 per cent confident that our *one* confidence interval (0.5, 3.8) has captured the difference in the two subpopulation means. Our stated level of confidence is a subjective *degree of belief* and not a relative frequency probability interpretation. This subjective degree of belief can be made more clear in the context of a simple physical example.

Suppose a jar contains 1000 marbles in which 950 are blue and 50 are red. All marbles are the same size, weight and texture and have been thoroughly mixed. Close your eyes and remove one marble. Knowing that 95 per cent of the marbles are blue and that every marble has the same chance of being selected, with eyes still closed, you would form a degree of belief regarding the color of the marble in your hand. Your degree of belief that the marble in your hand is blue should be identical to your degree of belief that the 95 per cent confidence interval in the commitment example (0.5, 3.8) has captured the difference in the two subpopulation means. We now give some hypothetical examples to illustrate how hypothesis testing results can be replaced with confidence interval results for several elementary types of statistical analyses. In some examples, the hypothesis testing results are ''non-significant'' and in other examples the hypothesis testing results are ''significant''. In each example, the use of a confidence interval, when contrasted with the more traditional hypothesis testing approach, provides important additional information.

Illustrating Use of Confidence Intervals

Example 1

A random sample of 250 managers was obtained from a study population of 3845 retail store managers. Each of the 250 managers completed a job satisfaction questionnaire and their most recent quarterly job

DOI: 10.1002/job

performance rating was also obtained. If the null hypothesis of a zero correlation can be rejected, the hypothesis testing results might be reported as: ''A statistically significant positive correlation between job satisfaction and job performance was found ($r = 0.38$, $p < 0.01$).'' This hypothesis testing result simply indicates that the population correlation is greater than 0, which may not be useful information. Furthermore, the sample correlation of 0.38 contains sampling error and could differ considerably from the correlation in the study population. An alternative analysis could report the following confidence interval result: ''A 95 per cent confidence interval for the correlation between job satisfaction and job performance in the study population of 3845 managers ranges from 0.269 to 0.481.''

Example 2

A random sample of 425 lawyers was obtained from a study population of 7945 public service lawyers in Arizona. Each of the 425 lawyers completed a nine-item emotional exhaustion scale and reported an average number of hours worked per week. If the null hypothesis of a zero correlation cannot be rejected, the results might be reported as: ''No statistically significant correlation between emotional exhaustion and hours worked per week was detected ($r = 0.09$, $p > 0.05$).'' It is important to remember that failure to reject the null hypothesis cannot be used as evidence that these two variables are uncorrelated in the study population. To show that the relation between emotional exhaustion and hours worked is small and unimportant, the following confidence interval result could be reported: ''A 95 per cent confidence interval for the correlation between emotional exhaustion and hours/week in the study population of 7945 Arizona public service lawyers ranges from -0.005 to 0.184. This result indicates that the correlation between emotional exhaustion and hours worked per week is at most 0.184, which suggests that the linear relation between these two variables is very weak.''

Example 3

A random sample of 50 employees was obtained from a study population of 6680 unionized assembly-line workers and a second random sample of 50 employees was obtained from a study population of 5922 non-unionized assembly-line workers. The 100 workers were given a 10-item (Agree–Disagree) questionnaire to measure their level of job stress. The sample mean stress score was 7.45 ($SD = 3.9$) for unionized workers and 6.22 ($SD = 3.7$) for non-unionized workers. The results of an independent-samples *t*-test might be reported as: ''Non-unionized workers exhibit greater levels of job stress than unionized workers, $t(48) = 3.20$, $p < 0.01$.'' However, the hypothesis testing result and sample means do not provide critical information about the magnitude of the difference in the study *subpopulation* means. Alternatively, a confidence interval result could be reported as: ''We can be 95 per cent confident that the mean stress score for the 5922 non-unionized workers is 0.47 to 1.99 greater than the mean stress score for the 6680 unionized workers.''

Example 4

A random sample of 500 unemployed managers was obtained from an executive employment agency that has over 18 000 clients. Each person in the sample was given a structured interview to determine age, degree of social support (measured on a 1–7 scale), and degree of financial hardship (measured on a 1–7 scale). Depression was measured on a 0–63 scale. Hypothesis testing results might be reported as: ''A multiple linear regression analysis was conducted with depression as the outcome variable and age,

DOI: 10.1002/job

social support, and financial hardship as the explanatory variables. The overall test for the predictive ability of the three explanatory variables was statistically significant, $F(3, 497) = 4.237$, $p = 0.006$.'' This is a non-informative test because the alternative hypothesis, which states that at least one of the three explanatory variables has a non-zero regression coefficient in the population, is almost certainly true.

A confidence interval for the population squared multiple correlation might be more useful and could be reported as: ''We are 95 per cent confident that differences in age, social support, and financial hardship are associated with 0.2 per cent to 5.1 per cent of the variance of depression scores in this population of approximately 18 000 unemployed managers.'' The confidence interval results show that the explanatory variables are at best weakly associated with depression, even though the hypothesis testing results were ''significant''.

Confidence intervals for the unstandardized regression coefficients also provide useful information. For instance, the effect of social support could be reported as: ''The 95 per cent confidence interval for the social support regression coefficient ranges from -0.08 to -0.72. This result suggests that any one-point decrease in social support is associated with a 0.08–0.72 increase in depression scores.'' Recall that depression is measured on a 0–63 scale, while social support is measured on a 1–7 scale and thus the 95 per cent confidence interval for the social support coefficient indicates that the association between social support and depression would be considered trivial.

Example 5

Two human resource managers were trained in the use of a new interviewing method for screening applicants for a large number of anticipated customer service vacancies. After training, the two managers independently interviewed a random sample of 15 applicants from a pool of 475 applicants to assess the reliability of each manager's rating. In studies such as these it is common to simply report point estimates of the reliability coefficients. For instance: ''The reliability of a single manager rating is 0.89.'' However, this point estimate contains sampling error and it is important to report a confidence interval for the study population reliability coefficient. For instance, the results could be stated as: ''The reliability of a single manager rating is 0.89 (95 per cent CI: 0.71–0.96) in our study population of 475 applicants.''

Example 6

A directory of about 120 000 professors from doctoral-granting universities was compiled and a random sample of 300 professors was obtained from this list. The professors in the sample were contacted and asked to describe their current level of job satisfaction as: (1) highly dissatisfied, (2) somewhat dissatisfied, (3) satisfied, or (4) highly satisfied. They were then asked to indicate if they are: (1) actively looking for another job, (2) thinking about other job possibilities, or (3) not considering a job change any time in the near future. A typical hypothesis testing result might be reported as: ''A chi-square test of independence revealed a statistically significant relation between job satisfaction and job seeking behavior, $\chi^2(6) = 187.5$, $p < 0.001$.'' This is another example of a non-informative hypothesis test because we know, prior to hypothesis testing, that job satisfaction and job seeking behavior are not completely independent.

Unlike the hypothesis testing result, a confidence interval for some measure of association will provide useful information regarding *both* the direction and strength of the dependency. In this case, if both variables are measured on an ordinal scale, then an ordinal measure of association would be most appropriate. Furthermore, an asymmetric measure of ordinal association is appropriate if we consider

DOI: 10.1002/job

one variable (e.g., job satisfaction) to be an explanatory variable and the other variable (e.g., job seeking) to be an outcome variable. Thus, a confidence interval for Somers' d (an asymmetric ordinal measure of association) could be computed and the result might be reported as: ''A 95 per cent confidence interval for Somers' d ranged from 0.49 to 0.64 and indicates that the ordinal association between job satisfaction and job seeking behavior is positive and moderately strong.''

Example 7

A random sample of 40 supervisors was obtained from a study population of 655 supervisors who work at one of three California branch offices of a health insurance company. The 40 supervisors were randomly divided into four groups. The first two groups received leadership training delivered in a classroom format and the other two groups received an on-line leadership training course. The first and third groups also received management internship assignments; the second and fourth groups did not receive management internship assignments. After training, the leadership potential of all 40 supervisors was assessed using a method that assigned scores of 1–20 to each supervisor, where higher scores reflect higher potential. A typical analysis for this 2×2 factorial experiment might be reported as: ''Statistically significant main effects of Training Method ($F(1, 36) = 4.98,\ p = 0.032$) and Internship ($F(1, 36) = 4.31, p = 0.045$) were detected. Training Method did not interact with Internship ($F(1, 36) = 1.33,\ p = 0.256$).''

The main effect results, as reported, are non-informative because they simply indicate that the population main effects are nonzero, a fact that was known prior to hypothesis testing. Furthermore, the above claim that ''Training Method did not interact with Internship'' is misleading. Recall that a failure to reject the null hypothesis cannot be used as evidence to support the null hypothesis. Confidence interval results are more informative and can be used to describe the magnitude of the main effects and the interaction effect. To accurately interpret the confidence intervals, it is necessary to first understand the meaning of the study population parameters. The four parameters θ_1, θ_2, θ_3 and θ_4 represent the mean leadership score for all 655 supervisors if they had all received: (1) classroom training with an internship; (2) classroom training without an internship; (3) on-line training with an internship; and 4) on-line training without an internship, respectively.

The main effect of Training is defined as $(\theta_1 + \theta_2)/2 - (\theta_3 + \theta_4)/2$ and its 95 per cent confidence interval ranges from -2.76 to -0.13 which suggests that the mean leadership score would be 0.13–2.76 points higher if all 655 supervisors were given on-line leadership training rather than classroom leadership training. The main effect of Internship is defined as $(\theta_1 + \theta_3)/2 - (\theta_2 + \theta_4)/2$ and its 95 per cent confidence interval ranges from 0.05 to 2.65 which suggests that the mean leadership score would be 0.05 to 2.65 points higher if all 655 supervisors received a management internship rather than no management internship. The two-way interaction effect is defined as $(\theta_1 - \theta_2) - (\theta_3 - \theta_4)$ and its 95 per cent confidence interval ranges from -1.11 to 4.11 which suggests that the effect of Internship with classroom leadership training $(\theta_1 - \theta_2)$ could be as much as 1.11 smaller than or as much as 4.11 greater than the effect of Internship with on-line leadership training $(\theta_3 - \theta_4)$. The confidence interval for the interaction effect is wide, in part, because the sample size is too small. A larger sample size is needed to accurately assess the magnitude of the interaction effect. The difference between the hypothesis testing results and confidence interval results in this example is striking and illustrates the main point made by Rothman (1986) quoted in our introduction.

Example 8

Covariance structure models (also known as causal models or structural equation models) are frequently used in management and organizational research. A particular causal hypothesis will imply

DOI: 10.1002/job

that some of the model parameters should be very small while other model parameters should be meaningfully large. Let Θ_1 be a $k \times 1$ vector of parameters that should be small under the hypothesized causal model and let Θ_2 be a $j \times 1$ vector of model parameters that should be meaningfully large under the hypothesized model. The so-called ''goodness-of-fit test'' is simply a test of H_0: $\Theta_1 = 0$ and a chi-square test with k degrees of freedom is typically used to test this null hypothesis. The chi-square goodness-of-fit test is a non-informative test of a null hypothesis that the k parameters all exactly equal zero against an alternative hypothesis that at least one of the parameters is not exactly equal to zero. It is common, although inappropriate, to interpret a ''non-significant'' chi-square goodness-of-fit test as evidence that the hypothesized model is correct.

Suppose that a particular causal hypothesis implies that variable *A* causes variable *B*, variable *B* causes variable *C*, and variable *C* causes variable *D*. This hypothesis implies that there will be meaningfully large path coefficients from variables *A* to *B*, *B* to *C*, and *C* to *D and* that the path coefficients from variables *A* to *C*, *A* to *D*, and *B* to *D* all should be small. A typical, but inappropriate, interpretation of a chi-square goodness-of fit test for this model might be stated as ''The chi-square goodness-of-fit test ($\chi^2(3) = 6.15$, $p > 0.10$) was non-significant and thus provides support for our hypothesized causal model.'' A much better way to assess the hypothesized model involves the computation of simultaneous (Bonferroni) confidence intervals for all theoretically important path coefficients (*i.e.*, all $j + k$ parameters in Θ_1 and Θ_2). If the simultaneous confidence intervals suggest that all path coefficients that should be small are small and all path coefficients that should be large are large, then this would provide support for the hypothesized model. It may be difficult to specify ''small'' and ''large'' values of the path coefficients because their values depend on the variance of the predictor variables. However, standardized path coefficients less than about 0.2 could be considered small and standardized path coefficients greater than about 0.5 could be considered large in most applications. Confidence intervals for standardized path coefficients may be obtained using bootstrap methods. In practice, a causal model will often be ''partially supported'' because not all path coefficients that should be large will be large and not all path coefficients that should be small will be small.

Some Methods to Avoid

To address the limitations of non-informative hypothesis tests, some social science journals recommend reporting a point estimate of a standardized measure of effect size along with *p*-values to describe the magnitude of the effect. Cohen's *d*, which is the difference between two sample means divided by a pooled standard deviation estimate, is a commonly used measure of effect size. Values of *d* equal to 0.2, 0.5, and 0.8 are claimed by Cohen (1988, p. 25) to represent ''small'', ''medium'', and ''large'' effects. However, reporting the sample value of Cohen's *d* can be very misleading. For instance, a researcher might report the results of a two-sample *t*-test as follows: ''Females were found to perform better than males on the teamwork task ($t(18) = 2.15$, $p < 0.05$). The gender effect appears to be large ($d = 0.96$).'' However, the value $d = 0.96$ describes the *sample* and not the study *population* of interest. The study population value of *d* could be much larger or smaller than 0.96. A 95% confidence interval for the population value of *d* in this example ranges from 0.02 to 1.88, indicating that the standardized effect size could be virtually zero or perhaps very large in the study population. Reporting the sample value of a standardized measure of effect size along with a *p*-value does not solve the problems associated with non-informative tests. If a standardized measure of effect size is of interest, a

DOI: 10.1002/job

confidence interval for the study population value of the standardized effect size should be reported along with the sample value of the standardized effect size.

Post hoc power analysis has been recommended as a way of interpreting a ''non-significant'' statistical test. This method involves computing the power of the test for the observed value of the test statistic. The result, referred to as the ''observed power'', can be obtained as an output option in the current version of SPSS. Proponents of this method argue that the observed power provides evidence that the null hypothesis is true if the statistical test fails to reject the null hypothesis and the observed power is sufficiently large. Hoenig and Heisey (2001) clearly explain the fallacy of this argument and show that observed power is simply a function of the *p*-value and therefore adds nothing to the interpretation of the results. Researchers should discontinue the practice of reporting observed power as evidence to support the null hypothesis.

Psychological Science recently added a new recommendation for contributors to the journal. Authors are now encouraged to report a value called p_{rep} proposed by Killeen (2005). The p_{rep} value is an estimate of the probability that an effect observed in a given sample could be replicated in another random sample of the same size from the same (or virtually identical) study population. The interpretation of p_{rep} is superficially attractive partly because of legitimate arguments regarding the scientific importance of ''replication'' as characterized in the physical sciences. In physical science experiments, results obtained in one laboratory are not generally accepted by the scientific community until the obtained results can be replicated in another laboratory. In these physical science experiments, statistical inference is not used to make an inference from a sample to a study population, but rather (using the terminology adopted here) the results of the experiment apply directly to a specific ''study population''. Thus, ''replication'' in the physical sciences is the generalization of results from one ''study population'' to other ''populations'' which is an essential step in the scientific process. However, the p_{rep} value does not estimate the probability that an effect observed in the current sample can be replicated in any *new* population; it simply estimates the probability that an effect can be replicated in another *random sample* of the same size from the *same* study population. The generalization of results from one random sample to another random sample from the same study population *is not* useful scientific information. In contrast, the generalization of results from a random sample to the study population *is* useful information and this is exactly what is provided by a confidence interval. Furthermore, p_{rep} is simply a function of the *p*-value and therefore it adds no scientific value beyond that provided by the *p*-value.

Publication Guideline Recommendations

Rozeboom (1960, p. 428) argued many years ago that ''... the stranglehold that conventional null hypothesis significance testing has clamped on publication standards must be broken.'' This change will not occur until journal gatekeepers provide tangible guidelines and support for authors seeking to implement this change. We suggest that journal editors and reviewers take the initiative in helping to change the way researchers present the statistical results of their studies. As one option, the ''Instructions to Authors'' could clearly recommend the use of confidence intervals and discourage the use of non-informative tests. Ideally, every journal should have at least one associate editor and several editorial board members willing to work with authors of conditionally accepted papers to insure that the appropriate confidence intervals have been reported and correctly interpreted.

The role of the *p*-value in editorial decisions also needs to be reconsidered. Almost 40 years ago Lykken (1968, p. 159) implored editors to ''... be bold enough to take responsibility for deciding which

DOI: 10.1002/job

studies are good and which are not, without resorting to letting the *p* value of the significance tests determine this decision.'' Publication criteria should not focus on the size of the *p*-value but rather on far more important criteria such as the quality of the sampling design, the quality of the research design, the psychometric reliability and validity of the variables under investigation, the theoretical or practical importance of the results, and the widths of the confidence intervals. Having challenged management researchers to mend their ways, we now look to editors, reviewers and authors to take the necessary steps to bring our scientific practices–with respect to confidence intervals and statistical tests–up to standard.

Acknowledgements

The authors gratefully thank Denise Rousseau for her help and insight.

Author biographies

Douglas G. Bonett is a professor of Statistics and Psychology at Iowa State University. He received his PhD from the University of California, Los Angeles in 1983. He has taught a wide variety of statistics courses over the years and most of his research focuses on the development of new statistical procedures.

Thomas A. Wright, formally a professor of Organizational Behavior at the University of Nevada, Reno, is now the Jon Wefald Leadership Chair in Business Administration and a professor of management at Kansas State University. He received his PhD from the University of California, Berkeley. Similar to the Claude Rains character from the classic movie, *Casablanca*, he has published in many of the ''usual suspects.'' The highlight of his professional career has been publishing a number of articles on business ethics with his late father, Vincent P. Wright.

References

Altman, D. G. , Machin, D., Bryant, T. N. , & Gardner, M. J. (Eds.). (2000). *Statistics with confidence (*2nd ed.). London: BMJ Books.

Boring, E. G. (1919). Mathematical versus statistical importance. *Psychological Bulletin*, *16*, 335–338.

Cohen, J. (1988). *Statistical power analysis for the behavior sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Gardner, M. J., & Altman, D. G. (1986). Confidence intervals rather than *P* values: Estimation rather than hypothesis testing. *British Medical Journal*, *292*, 746–750.

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

Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *American Statistician*, *55*, 19–24.

Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science*, *8*, 3–7.

Kaiser, H. F. (1960). Directional statistical decisions. *Psychological Review*, *67*, 160–167.

Killeen, P. R. (2005). An alternative to null-hypothesis significance testing. *Psychological Science*, *16*, 345–353.

DOI: 10.1002/job

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: APA.
Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin*, *70*, 151–159.
Morrison, D. E. , & Henkel, R. E. (Eds.), (1970). *The significance test controversy.* Chicago: Aldine.
Rothman, K. J. (1986). Significance questing. *Annals of Internal Medicine*, *105*, 445–447.
Rozeboom, W. W. (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletin*, *57*, 416–428.
Savage, I. R. (1957). Nonparametric statistics. *Journal of the American Statistical Association*, *52*, 331–344.

DOI: 10.1002/job