Judgment and Decision Making, Vol. 2, No. 5, October 2007, pp. 317–325

A note on determining the number of cues used in judgment analysis studies: The issue of type II error

Jason W. Beckstead*
University of South Florida College of Nursing

Abstract

Many judgment analysis studies employ multiple regression procedures to estimate the importance of cues. Some studies test the significance of regression coefficients in order to decide whether or not specific cues are attended to by the judge or decision maker. This practice is dubious because it ignores type II error. The purposes of this note are (1) to draw attention to this issue, specifically as it appears in studies of self-insight, (2) to illustrate the problem with examples from the judgment literature, and (3) to provide a simple method for calculating post-hoc power in regression analyses in order to facilitate the reporting of type II errors when regression models are used.

Keywords: judgment analysis, self-insight, multiple regression, post-hoc power.

1 Introduction

For decades judgment analysts have successfully used multiple regression to model the organizing cognitive principles underlying many types of judgments in a variety of contexts (see Brehmer & Brehmer, 1988; Cooksey, 1996; Dhami, et al., 2004, for reviews). Most often these models depict the individual judge or decision maker as combining multiple differentially weighted pieces of information (cues) in a compensatory manner to arrive at a judgment. Further, these analyses portray those who have acquired expertise on a judgment task as applying their judgment model or "policy" with regular, although less than perfect, consistency. The ability of linear regression models to accurately reproduce such expert judgments under various conditions has been discussed in detail (e.g., Dawes, 1979; Dawes & Corrigan, 1974; Einhorn & Hogarth, 1975). If one accepts the proposition that people's judgments can be modeled as though they are multiple regression equations, questions arise such as: 1) How many of the available cues does the individual use? and 2) How should the number of cues used be determined?

Too many researchers blindly apply statistical significance tests to inform them — in a kind of deterministic manner — whether judges did or did not attend to specific cues. If the t-test calculated on a cue's weight is significant, then the cue is counted as being attended to by the judge. Relying on p values in this way is a problem because these values are affected by the number of cues and number of cases presented to the judge during the task and by how well the overall regression equation fits the total set of responses.

This issue is discussed in this note which is organized as follows: First, examples from the judgment literature are reviewed to illustrate the existence of the problem. Second, notation commonly used by judgment analysts when describing regression procedures is introduced. Third, using this notation, a method for calculating the post-hoc power of t-tests on regression coefficients based on the noncentral t distribution is described. Fourth, this method is applied to estimate the number of cases necessary for statistical significance in order to illustrate how the investigator's conclusions about the number of cues attended to in a judgment task should be informed by considerations of type II error. Finally, an SPSS program for performing the calculations is described and provided in the Appendix.

2 Some examples in the judgment literature

Although it is reasonable to conclude that a "significant" cue is important to the judge and reliably used as he or she makes judgments, the converse does not follow. When a cue's weight (regression coefficient, standardized regression coefficient, or squared semipartial correlation) is not significant, it does not necessarily mean that the cue is unimportant; there may simply be insufficient statistical power to produce a significant test result. Determining the number of cues to which an individual attends is an important issue from both practical and theoretical viewpoints. In a practical sense, informing poorly performing

*Address: Jason W. Beckstead, University of South Florida College of Nursing, 12901 Bruce B. Downs Boulevard MDC22, Tampa, Florida 33612. Email: jbeckste@health.usf.edu

https://doi.org/10.1017/S1930297500000632 Published online by Cambridge University Press

judges that they should attend to more (or different) cues than they apparently do can improve their accuracy (see Balzer, et al., 1989, for review of cognitive feedback). Theories of cognitive functioning have long considered determining the amount of information we process to be a relevant question (e.g., Gigerenzer & Goldstein, 1996; Hammond, 1966; Miller 1956).

In the typical judgment analysis the problem of type II error is overlooked. I know of no studies in the judgment analysis literature that report the power of the significance tests on cue weights when these tests are relied upon to determine the number of cues being used by a judge. While an exhaustive review of the empirical literature is beyond the scope of this note, a few examples are presented to illustrate the problem.

Phelps and Shanteau's (1978) purportedly determined the number of cues used by expert livestock judges in making decisions using two different experimental ("controlled" and "naturalistic") designs. The same seven livestock judges rated the breeding quality of gilts (female breeding pigs) in two completely within-subject experiments. The controlled design used a partial factorial design in which each judge made 128 judgments of gilts described on 11 orthogonal cues. The naturalistic design used eight photographs of gilts. In this experiment the judges first rated the breeding quality of the gilt in each photo and then rated each photo on the same 11 cues used in controlled design. This procedure was repeated, resulting in a total of 16 judgments per judge. The authors then used significance tests to determine whether specific cues were being used by each judge in the two experiments. An important finding was that the judges used far more cues (mean = 10.1) in the controlled design than they did in the naturalistic design (mean = 0.9). The relevant data are summarized in Table 1. Using the *F* statistics reported in their Tables 1 and 2 to calculate estimates of effect sizes (η^2) reveals some paradoxical results; many of the cues showed stronger relationships to judgments in the naturalistic design. Because of the lower statistical power in the naturalistic design (the controlled design presented 128 cases whereas the naturalistic design presented only 16) fewer cues were counted as significant and it was concluded that less information was being used by all judges under the naturalistic design.

When comparing the results of the two experiments the authors attributed the difference in the amount of information used by the experts to the stimulus configuration, "...the source of the discrepancy seems to be in the intercorrelations among the characteristics and not in the statistical analysis" (Phelps & Shanteau, 1978, p.218). Although Phelps and Shanteau pointed out that the *F* statistics they report could easily be expressed as estimates of effect size they did not do so. If they had, they may have come to a different conclusion about the influences of naturalistic and controlled cue configurations in their judgment tasks.

One area of research particularly sensitive to the problem at hand is the study of self-insight into decisions. The assessment of self-insight in social judgment studies has traditionally compared statistical weights (derived via regression equations) with subjective weights. A widely accepted finding is that people have relatively poor insight into their judgment policies (see Brehmer & Brehmer, 1988; Harries, et al., 2000; Slovic & Lichtenstein, 1971, for reviews). In most studies assessing insight, judges are required to *produce* subjective weights (e.g., distributing 100 points among the cues). "It was the comparison of statistical and subjective weights that produced the greatest evidence for the general lack of self-insight" (Reilly, 1996, p. 214). Another robust finding from this literature is that people report using more cues than are revealed by regression models. "A cue is considered used if its standardized regression coefficient is significant" (Harries, et al., 2000, p. 461).

Two influential studies on insight by Reilly and Doherty (1989, 1992) asked student judges to *recognize* their judgment policies among those from several other judges. In the first study seven of eleven judges were able to identify their own policies. In contrasting this finding to previous studies the authors noted "These data reflect an astonishing degree of insight" (Reilly & Doherty, 1989, p. 125). In the second study the number of cues and the stimulus configuration were manipulated. Overall, 35 of 77 judges were able to identify their own policies. The authors reconciled this encouraging finding with the prevailing literature on methodologic grounds, arguing that the lack of insight shown in previous studies might be related to people's inability to articulate their policies. "There is the distinct possibility that while people have reasonable self-insight on judgment tasks, they do not know how to express that insight. Or pointing the finger the other way round, while people do have insight we do not know how to measure it" (1992, p. 305).

In both these studies, when judges were presented with policies, each judge's set of cue weights (squared semipartial correlations in this case) was rescaled to sum to 100, and importantly, cues which did not account for significant ($p < .01$) variance were represented as zeros. The authors noted the majority of judges (in both studies) indicated that they had relied on the presence or absence of zeros as part of the search strategy used to recognize their own policies. The use of significance tests to assign specific cues a rescaled value of zero in these studies is problematic for two reasons. First, the power of a significance test on a squared semipartial correlation in multiple regression is affected by the value of the multiple R^2. As R^2 increases, smaller weights are more likely to be significant. Second, the power of these significance tests

https://doi.org/10.1017/S1930297500000632 Published online by Cambridge University Press

Table 1: Summary of results from Phelps and Shanteau (1978) with addition of effect size estimates.

	No. of significant cues		Median η^2		No. cues with larger η^2 in naturalistic
Judge	Controlled	Naturalistic	Controlled	Naturalistic	
1	10	2	0.205	0.365	5
2	9	0	0.321	0.310	7
3	10	0	0.158	0.024	3
4	9	3	0.264	0.333	8
5	11	1	0.177	0.200	5
6	11	0	0.376	0.167	2
7	11	0	0.162	0.184	5

is affected by the number of predictors in the regression equation. The net result was that the criterion used to assign zero to a specific cue was not constant across judges. Only when all judges are presented with the same number of cues and all have equal values of R^2 for their resultant policy equations could the criterion be consistently applied.

To illustrate, Reilly and Doherty (1989) presented 160 cases containing 19 cues to each judge. Consider two judges with different values of R^2 based on 18 of the cues, say .90 and .50. The minimum detectable effect (i.e., smallest weight that the 19th cue could take and still be significant) for the first judge is .008 but .039 for the second judge. The same problem exists in the 1992 study that used 100 cases and is compounded by the fact that the authors manipulated the number of cues presented to the judges; half the sample rated cases described by six cues and the other half rated cases described by twelve cues. In the recognition portion of both studies the useful pattern of zeros in the cue profiles was an artifact introduced arbitrarily by the use of significance tests. Had the authors used $p < .05$ rather than $p < .01$ to assign zeros, their conclusions about insight might have been astonishingly different.

Harries et al. (2000, Study 1), examining the prescription decisions of a sample of 32 physicians, replicated the finding that people are able to select (recognize) their policies among those from several others. This study followed up on the participants in a decision making task (Evans, et al., 1995) in which 100 cases constructed from 13 cues were judged and regression analysis was used to derive decision policies. Judges also provided subjective cue weights, first indicating the direction (sign) of influence, then rating how much (0–10 scale) the cue had bearing on their decisions. When comparing tacit to stated policies (i.e., regression weights to subjective weights) Harries et al. (2000) described a "triangular pattern of self-insight": a) cues that had significant weights were the ones that the judge indicated he or she used, b) where the judge indicated that a cue was not important it did not have a significant weight, and c) there were cues that the judge indicated were important but which did not have significant weights. The authors' choice of p value for determining whether a cue was attended to in the tacit policies had influence on all three sides of this triangular pattern.

Approximately 10 months following the decision-making task, participants were presented with sets of decision policies in the form of bar charts rather than tables of numbers. Cues with statistically significant weights were presented as darker bars. With only four cues having significant effects on decisions (Harries et al., 2000, p. 457), it is possible that physicians used the presence or absence of lighter bars in the same way that Reilly and Doherty's students made use of zeros in their recognition strategies. Had more cues been classified and presented as significant, the policy recognition task might have proved more difficult.

Other examples exist in the applied medical judgment literature. Gillis et al. (1981) relied extensively on p values of beta weights for describing the judgment policies of 26 psychiatrists making decisions to prescribe haloperidol based on 8 symptoms (see their Table 4). Averaged across judges, the number of cues used was 2.4, 1.9, or 1.0 depending on the p value employed (.05, .01, or .001, respectively). Had the investigators chosen to compare the number of cues used with self-reported usage, which of the three p values ought they have relied upon? Had the investigators rescaled and presented policies to participants for recognition (via Reilly & Doherty), their choice of p value could have affected the difficulty of the recognition task.

More recently, in a judgment analysis of 20 prescribing decisions made by 40 physicians and four medical guideline experts, Smith et al. (2003) reported "The number of significant cues ... varied between doctors, ranging from

https://doi.org/10.1017/S1930297500000632 Published online by Cambridge University Press

0 to 5" (p. 57), and among the experts "The mean number of significant cues was 1.25" (p. 58). It is noteworthy that this study presented doctors with a relatively small number of cases thus leaving open the meaning of "significant." Had Smith et al. presented more than 20 cases, they may have concluded (based on p values) that doctors and guideline experts attended to more information when making prescribing decisions.

Other models of judgment, known as "fast and frugal heuristics" have recently been proposed as alternatives to regression models (see Gigerenzer, 2004; Gigerenzer & Kurzenhäuser, 2005; Gigerenzer, et al., 1999). A hallmark of fast and frugal models is that they are purported to rely on far fewer cues than do judgment models described by regression procedures. When comparing these classes of models, the number of cues the judge uses is one way of differentiating the psychological plausibility of these models (see Gigerenzer, 2004). Studies comparing regression models with fast and frugal models have implied that significance testing is *the* method of determining the number of cues used despite the fact that the developers of these methods (e.g., Stewart, 1988) made no such claim and currently advise against it (Stewart, personal communication, July 2, 2007).

In a study comparing regression with fast and frugal heuristics, Dhami and Harries (2001) fitted both types of models to 100 decisions made by medical practitioners. They report that number of cues attended to was significantly greater when modeled by regression than by the matching heuristic. According to the regression models the average number of cues used was 3.13 and the average for the fast and frugal models was 1.22. "In the regression model a cue was classified as being used if its Wald statistic was significant ($p < .05$) . . . " (Dhami & Harries, 2001, p. 19). In the heuristic model, the number of cues used was determined by the percentage of cases correctly predicted by the model; significance tests were not used. At issue is not the fact that different criteria were used to count the cues used under the two types of models (although this is a problem when evaluating their results), but rather, that the authors relied on a significance test known for some time to be dubious[1], and their choice of p value for counting cues may have biased their data to favor the psychological plausibility of the fast and frugal model. Had they used $p < .01$ rather than $p < .05$, the average number of cues used according to the regression procedure would presumably have been lower, and perhaps not different than the average found for the matching heuristic.

In the last few paragraphs, examples from the literature have been presented that highlight the problems associated with using significance tests to determine the number of cues used in judgment tasks. Tests of significance on regression coefficients or R^2 are really not very enlightening for distinguishing the "best" judgment model from among a set of competing models. The true test of which model (among a set of contenders) is the best is the ability of the equation to predict the judgments made in some future sample of cases, the data from which were not used to estimate the regression equation. The remaining sections of this note formally present the regression model as used in judgment analysis and discuss a method for assessing the power of significance tests so as to provide more information to judgment analysts who use them.[2]

3 Notation

Following Cooksey (1996), let the *k* cues be denoted by subscripted X's (e.g., X_1 to X_k). In a given judgment analysis a series of m profiles or cases is constructed where each case is comprised of *k* cues. The judge or subject makes m responses Y_s to these cases. The resulting multiple regression equation representing the subject's judgment policy is of the general form

$$Y_s = b_0 + b_1X_1 + b_2X_2 + ... + b_kX_k + e \qquad (1)$$

where b_0 represents the regression constant and the remaining b_i represent regression coefficients for each cue where each coefficient indicates the amount by which the prediction of Y_s would change if its associated cue value changed by one unit while holding all other cue values constant, and e represents residual or unmodeled influences.

Tests of significance may be employed to assess the null hypothesis that the value of b_i in the population is zero, thus H_0: $b_i = 0$ against the alternative H_1: $b_i \neq 0$. The ratio b_i/SEb_i is distributed as a t statistic with degrees of freedom (*df*) = m - *k* - 1. The SEb_i is found as

$$SEb_i = \frac{sd_{Ys}}{sd_{Xi}}\sqrt{\frac{1-R^2_{Y_s}}{m-k-1} \times \frac{1}{R^2_{Xi}}} \qquad (2)$$

where sd_{Ys} and sd_{Xi} are, respectively, the standard deviations for the judgments and for the *i*th cue's values; $R^2_{Y_s}$ is the squared multiple correlation for the judgment equation; and R^2_{Xi} is the squared multiple correlation

[1]Hauck and Donner (1977) found that the Wald test behaves in an aberrant manner. Jennings (1986) has also questioned the adequacy of the Wald test for making statistical inferences. Hosmer and Lemeshow (2000) recommend using the likelihood-ratio test instead.

[2]The utility of statistical significance and hypothesis testing as a general approach has been questioned by researchers in the social sciences (e.g., Armstrong, 2007; Nickerson, 2000; Rozeboom, 1960). I believe that many of us are likely to continue to rely on this approach for some time. It is therefore important that we fully understand the assumptions, mechanics, and limitations of this approach.

https://doi.org/10.1017/S1930297500000632 Published online by Cambridge University Press

from a regression analysis predicting the *i*th cue's values from the values of the remaining $k-1$ cues. In standard multiple regression it can be shown that the significance test of b_i ($t = b_i/SEb_i$) is equivalent to testing significance of the standardized regression coefficient β_i and the squared semipartial correlation associated with X_i (see Pedhazur, 1997). This is fortunate because most commercially available statistics packages routinely print values for SEb_i but not for $SE\beta_i$.[3]

4 Post-hoc power analysis on t-test of regression coefficients

Having analyzed data from a judgment analysis using multiple regression it is rather simple to calculate the statistical power associated with the t-test of each regression coefficient. All that is needed from the analysis is the observed value of t, its *df*, and the a priori specified value of α. To obtain the power of the t-test that H_0: $b_i = 0$ for $\alpha = .05$, one may employ the *noncentral* distribution of the t statistic (see Winer et al., 1991, pp. 863–865), here denoted t', which is actually a family of distributions defined by *df* and a noncentrality parameter δ, hence $t'(df; \delta)$. In the present context $\delta = b_i / SEb_i$. The power of the t-test on the regression coefficient may then be determined as

$$Prob(t') > t_{df,\,1-\alpha/2} | \delta = b_i/SEb_i) = 1 - Prob(\text{type II error}) \quad (3)$$

Thus the probability that the noncentral t' will be greater than the critical value of t, given the observed value of $t = b_i / SEb_i$, is equal to the power of the test that H_0: $b_i = 0$ for $\alpha = .05$. For example, consider the following result from an illustrative judgment analysis involving $k = 6$ cues and $m = 30$ cases provided by Cooksey (1996, p.175). The unstandardized regression coefficient for a particular cue is $b = 0.267$, ($\beta = .295$) its standard error is 0.146, thus $t = 0.267/0.146 = 1.829$. The critical value for t with $df = 30 - 6 - 1 = 23$, and $\alpha = .05$ for a two-tailed test is 2.069; consequently the null hypothesis is not rejected and it might be concluded that this cue is unimportant to the judge. Using the information from this significance test and the noncentral distribution of $t'(df = 23; \delta = 1.829)$ we find that the probability of type II error = .582, and thus the power to reject the null is only .418. To claim that this cue is "unimportant to the judge," or "is not being attended to by the judge" does not seem justifiable in light of the rather high probability of type II error.

[3]The method presented here is also directly applicable to standardized regression coefficients when their corresponding standard errors are available.

5 Estimating the number of cases necessary for significant t-test of regression coefficients

Faced with such a nonsignificant result, as in the example presented above, the judgment analyst may wish to know the extent to which this outcome was related to the study design. In particular, how was the nonsignificant t-test of the cue weight affected by his or her decision to present m cases to the judge instead of some larger number m^*? To address this question we must first clarify the types of the stimuli used in judgment studies.

Brunswik (1955) argued for preserving the substantive properties (content) of the environment to which the investigator wishes to generalize in the stimuli presented during the experimental task. Hammond (1966), in attempting to overcome the difficulties inherent in such representative designs, distinguished between "substantive" and "formal" sampling of stimuli. Formal stimulus sampling concerns the relationships among environmental stimuli (with content ignored). The following discussion is limited to studies employing formal stimulus sampling. When taking the formal approach to stimulus sampling, the investigator's focus is on maintaining the statistical characteristics of the task environment (e.g., *k*, sd_{Xi} and R^2_{Xi}) in the sample of stimuli presented to the participant. These characteristics of the environment may be summarized as a covariance matrix, $\mathbf{\Sigma}$. If the investigator obtains a sample of m stimuli from the environment, the covariance matrix $\mathbf{S}_\text{m}$, may be computed from the sample and compared with $\mathbf{\Sigma}$. The basic assumption of formal stimulus sampling may then be stated as $\mathbf{S}_\text{m} \approx \mathbf{\Sigma}$. Whether probability or nonprobability sampling is used, it is possible for the investigator to construct an alternative set of m^* cases such that $\mathbf{S}_{\text{m}^*} = \mathbf{S}_\text{m}$. Under the condition that $\mathbf{S}_{\text{m}^*} = \mathbf{S}_\text{m} \approx \mathbf{\Sigma}$, it is possible to estimate SEb_i*, the standard error of the regression coefficient based on the larger sample of cases m^*. Inspection of Eq. (2) reveals that SEb_i becomes smaller as the number of cases m becomes larger. Holding all other terms in Eq. (2) constant, SEb_i* may be found as

$$SEb_i^* = \frac{SEb_i}{\sqrt{\frac{m^*-k-1}{m-k-1}}} \quad (4)$$

Substituting SEb_i* in place of SEb_i when calculating t-test on b_i allows us to estimate the impact of increasing m to m^* on type I error in the same judgment analysis. Making the same substitution in Eq. (3) allows us to estimate the impact of this change on type II error and power.

Stewart (1988) has discussed the relationships among *k*, R^2_{Xi}, and m and recommends $m = 50$ as a minimum for reliable estimates of cue weights when *k* ranges from

https://doi.org/10.1017/S1930297500000632 Published online by Cambridge University Press

4 — 10 and $R^2_{Xi} = 0$. He points out that as the intercorrelations among the cues increases the number of cases will need to be increased in order to maintain reasonably small values of SEb_i. Of course the investigator's choice of m should also influenced by his or her sense of subject burden. Stewart notes from empirical evidence that most judges can deal with making between "40 to 75 judgments in an hour, but the number varies with the judge and the task" (Stewart, 1988, p.46). In discussing the design of judgment analysis studies Cooksey (1996) has suggested that the optimal number of cases may be closer to 80 or 90. Reilly and Doherty (1992) reported the average time for 77 judges to complete 100 12-cue cases was 1.25 hours. In a recent study by Beckstead and Stamp (2007) 15 judges took on average 32 minutes (range 20–47) to respond to 80 cases constructed from 8 cues.

For the example given in the previous section, if the investigator had used m^* = 40, rather than m = 30, Eq. (4) indicates that SEb_i* would have been 0.122 and the resulting value for the t-test would have been 2.191 with p = .036. The point here is that had the investigator presented 10 more cases (sampled from the same population), he or she might have come to a different conclusion about the number of cues attended to by this judge.

6 An SPSS program for calculating post-hoc power in regression analysis

The calculations for determining post-hoc power for tests of regression coefficients as used in judgment analysis studies and estimating SEb_i* are straightforward and based on statistical theory, however detailed tables of noncentral t distributions are hard to come by. The author has written an SPSS program for performing these calculations that is provided in the Appendix. To illustrate the program, consider another cue taken from the same example found in Cooksey (1996, p.175) where b = -0.423, SEb = 0.386, and k = 6 for m = 30. Inserting these values into the program and specifying that the number of cases increase to 90 by increments of 10, produces the result shown in Table 2.

As m^* increases, the estimated values of SEb^* decrease and the values of the t-statistic increase. According to these estimates, the t-test on this cue would have been significant had approximately 85 cases been used in the judgment task. The program can be "rerun" specifying a smaller increment in order to refine this estimate. The results provided by such an analysis could also be very useful in the planning of subsequent judgment studies.

Table 2: Illustration of the influence of the number of cases $m*$ on t-tests of regression coefficient

m^*	SEb^*	t-test	p-value
40	0.322	1.313	.198
50	0.282	1.498	.141
60	0.254	1.664	.102
70	0.233	1.814	.074
80	0.217	1.952	.055
90	0.203	2.082	.040

Note: The regression coefficient b = -0.423 and SEb = 0.386 for m = 30. The t-test of this coefficient was t = -1.096, p = .284; post-hoc power of the t-test is given by the Eq. (3) as .182. Due to negative sign of regression coefficient, resulting t-test values are negative; the sign has been omitted for clarity of presentation.

7 Summary and recommendations

In this note the issue of type II error has been raised in the context of determining whether or not a cue is important to a judge in judgment analysis studies. Some of the potential pitfalls of relying on significance tests to determine cue utilization have been pointed out and a simple method for calculating post-hoc power of such tests has been presented. A short computer program has been provided to facilitate these analyses and encourage the calculation (and reporting) of statistical power when judgment analysts rely on significance tests to inform them as to the number of cues attended to in judgment tasks.

As a tool for understanding the individual's cognitive functioning, regression analysis has proved to be quite useful to judgment researchers for over 40 years. In this role I believe that its true value lies in its descriptive, not its inferential, facility. Like any good tool, if we are to continue our reliance upon it we must insure that it is in proper working order and not misuse it.

There are alternative models of judgment being advocated (e.g., probabilistic models proposed by Gigerenzer and colleagues) that do not fall prey to the problems associated with regression analysis. However, as judgment researchers develop, test, and apply these models, questions about the amount of information (i.e., the number of cues) individuals use when forming judgments and making decisions are bound to arise. The strongest evidence for the veracity of any judgment model is its ability to predict the outcomes of future decisions.

The practice of post-hoc power calculations as an aid in the interpretation of nonsignificant experimental results is not without its critics (e.g., Hoenig & Heisey,

https://doi.org/10.1017/S1930297500000632 Published online by Cambridge University Press

2001; Nakagawa & Foster, 2004). Hypothesis testing is easily misunderstood but when applied with good judgment it can be an effective aid to the interpretation of experimental data (Nickerson, 2000). Higher observed power does not imply stronger evidence for a null hypothesis that is not rejected (see Hoenig & Heisey, 2001 for discussion of the power approach paradox). Some researchers have argued for abandoning the use hypothesis testing altogether and relying instead on the confidence interval estimation approach (Armstrong, 2007; Rozeboom, 1960). I tend to agree with Gigerenzer and colleagues who put it succinctly, "As long as decisions based on conventional levels of significance are given top priority . . . theoretical conclusions based on significance or nonsignificance remain unsatisfactory without knowledge about power" (Sedlmeier & Gigerenzer, 1989, p. 315).

References

Armstrong, J. S. (2007). Significance tests harm progress in forecasting. *International Journal of Forecasting, 23*, 321–327.

Balzer, W. K., Doherty, M. E., & O'Connor, R. Jr. (1989). Effects of cognitive feedback on performance. *Psychological Bulletin, 106*, 410–433.

Beckstead, J. W., & Stamp, K. D. (2007). Understanding how nurse practitioners estimate patients' risk for coronary heart disease: A judgment analysis. *Journal of Advanced Nursing, 60*, 436–446.

Brehmer, A., & Brehmer, B. (1988). What have we learned about human judgment from thirty years of policy capturing? In B. Brehmer & C. R. B. Joyce (Eds.), *Human judgment: The SJT view*, (pp. 75–114). Amsterdam: Elsevier Science Publishers.

Brunswik, E. (1955). Representative design and probabilistic theory in functional psychology. *Psychological Review, 62*, 193–217.

Cooksey, R. W. (1996). *Judgment analysis: Theory, methods, and applications*. San Francisco: Academic Press.

Dawes, R. M., (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34*, 571–582.

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin, 81*, 95–106.

Dhami, M. K., Hertwig, R., & Hoffrage, U. (2004). The role of representative design in an ecological approach to cognition. *Psychological Bulletin, 130*, 959–988.

Dhami, M. K., & Harries, C., (2001). Fast and frugal versus regression models of human judgment. *Thinking and Reasoning, 7*, 5–27.

Einhorn, H. J., & Hogarth, R. M. (1975).Unit weighting schemes for decision making. *Organizational Behavior and Human Performance, 13*, 171–192.

Evans, J. St. B. T., Harries, C., Dennis, I., & Dean, J. (1995). General practitioners' tacit and stated policies in the prescription of lipid lowering agents. *British Journal of General Practice, 45*, 15–18.

Gigerenzer, G. (2004). Fast and frugal heuristics: The tools of bounded rationality. In D. J. Koehler and N. Harvey (Eds.), *Blackwell handbook of judgment and decision making*, (pp. 62–88). Oxford: Blackwell Publishing.

Gigerenzer, G. & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review, 103*, 650–669.

Gigerenzer, G., & Kurzenhäuser, S. (2005). Fast and frugal heuristics in medical decision making. In R. Bibace, J. D. Laird, K. D. Noller, and J. Valsiner (Eds.), *Science and Medicine in Dialogue: Thinking through Particulars and Universals*, (pp. 3–15). Westport CN: Praeger.

Gigerenzer, G., Todd, P. M., & the ABC Research Group (Eds.) (1999). *Fast and frugal heuristics: The adaptive toolbox*.

Gillis, J. S., Lipkin, J. O., & Moran, T. J. (1981). Drug therapy decisions. *Journal of Nervous and Mental Disease, 169*, 439–437.

Hammond, K. R. (1966). Probabilistic functionalism: Egon Brunswik's integration, of the history, theory, and method of psychology. In K. R. Hammond (Ed.), *The psychology of Egon Brunswik*. New York: Holt Rinehart & Winston, (pp. 15–80).

Harries, C., Evans, J. St. B. T., & Dennis, I. (2000). Measuring doctors's self-insight into their treatment decisions. *Applied Cognitive Psychology, 14*, 455–477.

Hauck, W. W. & Donner, A. (1977). Wald's test as applied to hypotheses in logit analysis. *Journal of the American Statistical Association, 82*, 1110–1117.

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.

Hosmer, D. W. & Lemeshow, S. (2000). *Applied logistic regression, 2nd Ed.* New York: John Wiley & Sons, Inc.

Jennings, D. E. (1986). Judging inference adequacy in logistic regression. *Journal of the American Statistical Association, 81*, 987–990.

Miller, G. A. (1956). The magical number seven plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*, 81–97.

Nakagawa. S. & Foster, T. M. (2004). The case against retrospective statistical power analyses with an introduction to power analysis. *Acta Ethologica, 7*, 103–108.

Nickerson, R. S. (2000). Null hypothesis significance

testing: A review of an old and continuing controversy. *Psychological Methods, 5,* 241–301.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and prediction*, 3rd ed. Fort Worth: Harcourt Brace College Publishers.

Phelps, R. H, & Shanteau, J. (1978). Livestock judges: How much information can an expert use? *Organizational Behavior and Human Performance, 21*, 209–219.

Reilly, B. A. (1996). Self-insight, other-insight, and their relation to interpersonal conflict. *Thinking and Reasoning, 2*, 213–222.

Reilly, B. A., & Doherty, M. E. (1989). A note on the assessment of self-insight in judgment research. *Organizational Behavior and Human Decision Processes, 44*, 123–131.

Reilly, B. A., & Doherty, M. E. (1992). The assessment of self-insight in judgment policies. *Organizational Behavior and Human Decision Processes, 53*, 285–309.

Rozeboom, W. W. (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletin, 57*, 416–428.

Sedlmeier, P. & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105*, 309–316.

Slovic, P. & Lichtenstein, S. (1971). Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance, 6*, 649–744.

Smith, L., Gilhooly, K., & Walker, A. (2003). Factors influencing prescribing decisions in the treatment of depression: A Social Judgment Theory approach. *Applied Cognitive Psychology, 17*, 51–63.

Stewart, T. R. (1988). Judgment analysis: Procedures. In B. Brehmer & C. R. B. Joyce (eds.) *Human judgment: The SJT view*, (pp. 41–74). Amsterdam: Elsevier Science Publishers.

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical principles in experimental design*, 3rd ed. New York: McGraw-Hill, Inc.

Appendix

The following is an SPSS program to calculate post-hoc power of t-test on regression coefficients and to estimate sample size needed for significance of such tests. After typing the commands into a syntax window and supplying information specific to your analysis, simply run the program to obtain results similar to those found in Table 2.

Color Key: commands, comments, information to be supplied by the user.

```
**------------------------------------------------------------------------------.
**ENTER NECESSARY INFORMATION FROM MULTIPLE REGRESSION ANALYSIS HERE*.
DEFINE @STUFF ().
COMPUTE b = -0.423                          /*unstandardized regression coefficient */.
COMPUTE SEb = 0.386                       /*standard error of regression coefficient */.
COMPUTE k = 6                           /*number of predictors in regression equation */.
COMPUTE N = 30                                /*number of observations or cases */.
COMPUTE alpha = .05                                    /*type I error criterion */.
COMPUTE maxN = 90                       /*maximum value of N for table of estimates */.
COMPUTE incN = 10                           /*increment in N for table of estimates */.
!ENDDEFINE .
**------------------------------------------------------------------------------.
**CALCULATING POST-HOC POWER for t-TEST of REGRESSION COEFFICIENT.
NEW FILE.
INPUT PROGRAM.
@STUFF.
COMPUTE t = ABS(b/SEb)                 /*confirming t-test on b found in reg output */.
COMPUTE df = N-k-1                            /*degrees of freedom for t-test on b */.
COMPUTE tcrit = IDF.T(1-(alpha/2),df)          /*critical value of t for desired α */.
COMPUTE t_prob = 2*(1-CDF.T(t,df))            /*this is obs p value for t-test on b */.
COMPUTE Power = 1-NCDF.T(tcrit,df,t)           /*post-hoc power for obs t-test on b */.
END CASE.
END FILE.
END INPUT PROGRAM.
FORMAT N k DF (F3.0) t_prob t b SEb Power (F8.3).
LIST b SEb t k N t_prob power.
**ESTIMATING SAMPLE SIZE NECESSARY FOR t-TEST OF b TO BE SIGNIFICANT.
NEW FILE.
INPUT PROGRAM.
@STUFF.
LOOP newN = N+incN TO maxN BY incN.
COMPUTE SEbStar = SEb/SQRT((newN-k-1)/(N-k-1))             /*est of SEb under new N */.
COMPUTE tcritN = IDF.T(1-(alpha/2),newN-k-1)          /*crit t value for desired α */.
COMPUTE tstar = ABS(b/SEbStar)                              /*est of t under new N */.
COMPUTE t_probN = 2*(1-CDF.T(tstar,newN-k-1))           /*est of p-value for tstar */.
COMPUTE powerN = 1-NCDF.T(tcritN,newN-k-1,tstar)   /*estd power of test under new N */.
END CASE.
LEAVE b SEb k N alpha.
END LOOP.
END FILE.
END INPUT PROGRAM.
FORMAT newN (F5.0) SEbStar powerN tstar t_probN b (F5.3).
LIST newN b SEbStar tstar t_probN powerN.
```

https://doi.org/10.1017/S1930297500000632 Published online by Cambridge University Press