UNDERSTANDING STATISTICS, 3(1), 101–134

TEACHING ARTICLE

Some Myths and Legends in Quantitative Psychology

Dave Grayson
School of Psychology
University of Sydney

The aim of this article is to discuss some of the limitations of some common quantitative methods used in psychology. The main reason for so doing is to attempt to challenge the "golden rule" aspect that such methods frequently acquire among researchers, with the aim of emphasizing that quantitative methods are tools with limitations that need evaluation, and with rationales that require justification, in each single scientific application.

causality, Cronbach's alpha, Cohen's Kappa, normality, power, family-wise error, MANOVA, variable importance

As the recent contretemps in psychology about null hypothesis significance testing indicates (see e.g., Cohen, 1994; Harlow, Mulaik, & Steiger, 1997; Nickerson, 2000; Wilkinson & The APA Task Force on Statistical Inference, 1999), there is an unfortunate aspect of "institutionalization" associated with the use of statistical and psychometric techniques in psychology. Significance testing has a long and controversial history in psychology, and in statistics for that matter (see Gigerenzer et al., 1989; Oakes, 1986), and one might doubt that this most recent debate has resolved, or even addressed, the more fundamental issues involved (e.g., about the very meaning of "probability"; see Grayson, 1998, and references therein). This article is not, however, a further offering about significance testing or the meaning of p values. This article addresses some commonly used quantitative methods, arguing that their use is also somewhat "institutionalized," and that their proper use in any given application demands due attention to their rationales.

Requests for reprints should be sent to Dave Grayson, School of Psychology, University of Sydney, Sidney NSW 2006 Australia. E-mail: dgrayson@psych.usyd.edu.au

Good scientific research is a creative and original process, and "golden rules" are anathema to creativity. In any single scientific application, one would hope a good researcher would "tailor" or craft their methods and procedures to that application. A given quantitative technique may be very useful in one situation, but be of rather limited use in another. Making and using such intelligent discrimination requires that the user be aware of the tools they use, their rationales and their limitations. No carpenter would ever make a statement such as "hammers are good tools"; they are for some applications, but not for others.

In any event, whether one agrees with this perspective, the aim of this article is to discuss some commonly used quantitative techniques, and some common-held beliefs about them, attempting to provide insight into how and when they may *not* be appropriate (and thus when they are). My motivation for doing this comes from the frequent interactions I have had with (a) researchers believing that some universal, "expert" justification exists for their chosen golden rule; (b) researchers being forced by some journal referee to perform some analysis (because it is that referee's golden rule) for which they see no need; and (c) researchers seeking the solace of expertise when in fact a difficult empirical evaluation, justification, and decision on *their* part is what is required.

The techniques discussed next seem to me to be uncontroversial in their mathematical rationales, and these mathematical underpinnings do not seem open to debate or discussion. Any debate should revolve around a *particular empirical application* of that technique, about what state-of-affairs indeed applies—and thus about the *rationale for the technique* in that application. The techniques are what they are, tools, and thus are neither good nor bad. So, what is discussed next should in no way be taken as a criticism of any tool per se. On the other hand, every tool easily can be inappropriate to use in some application, and this is the aspect of the scientific process to which the current offering is aimed. Even these statements, however, are my *opinions*, and we all must form our own. What would seem to be not ideal is to simply accept someone else's (some expert's) opinion without understanding *why*; and this seems to happen far too often in psychology in relation to quantitative methods.

Nine numbered issues are discussed next, only one of which (the fourth, normality and multiple linear regression [MLR]) reflects a misunderstanding—frequent among "users," in my experience—that would *not* appear in any textbook. An appendix presents some relevant underlying technical material for those who are interested, numbered accordingly.

PSYCHOMETRIC MODELS AND CAUSALITY

There are at least two rationales for measuring a psychological "construct" by accumulating the responses to a set of component items, or questions in a questionnaire. They can be distinguished by the direction of causality involved, and might be re-

ferred to as (a) measurement by the *accumulation of effects* and (b) measurement by the *accumulation of causes*. These two situations are represented graphically as Figures 1a and 1b, respectively. (For a similar distinction between "cause indicators" and "effect indicators," see Bollen, 1989, pp. 64–67.)

In Situation (a) variation in each item or observed component is thought to result *because of* variation in the underlying construct. Hence, when the underlying construct varies (across a population of subjects), the items simultaneously vary as a result of this; hence the items covary together, and it is from this covariation (or correlation) that we infer the existence of the underlying construct. A typical example of this (causal) model would be the responses to symptoms of (i.e., results of) the construct of depression. This causal model is the one that underlies (all?) psychometric models, and that is why the techniques of psychometrics are correlational in essence (interitem correlations, Cronbach's α, factor analysis, etc.).

In Situation (b), the construct is measured by accumulating its *causes*. A good example of this rationale would be a Life Events scale. The construct being measured is "stress-due-to-life-events," and the items accumulated are those stressors that bring about this stress (i.e., that cause it), for example, recent divorce, recent job loss, recent promotion, recent death of spouse, recent house move, and so on. There is no reason why these recent stressors should *not* be independent and hence uncorrelated (the "slings and arrows" of recent "fate"), and hence the correlative, psychometric models are inappropriate, and using such psychometric models in cases like this is equally inappropriate.

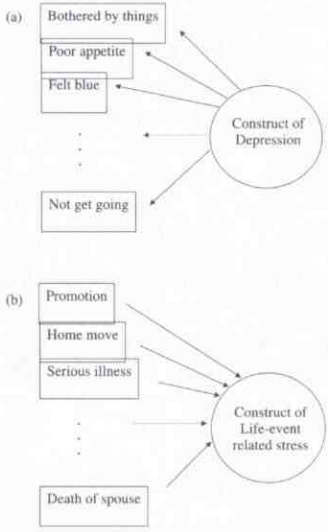

FIGURE 1 Psychometric (a) and nonpsychometric (b) models for accumulating item responses.

CRONBACH'S ALPHA

Cronbach's α is regularly used as an index of item "homogeneity," or of "unidimensionality," among the component items of a test. Typically, the test is a new test, and a moderate to high value of α on some sample is cited as evidence that all the items in the test are measuring the same (intended) psychological "construct." Unfortunately, for tests with a small number of items (7 or so, say), this reasoning is invalid: Moderate to high values of α can easily obtain when the test is *not* unidimensional. Indeed, Cronbach (1951) introduced this index to be used as a measure of reliability on items already presumed to be unidimensional: "Tests divisible into distinct subtests should be so divided before using the formula" (p. 297).

Table 1 shows three different sets of data, all yielding the same moderate-to-high value for Cronbach's α (α = 0.8). In each case, seven items are involved, and each item has a variance of 10 units. The cases differ in their patterns of covariance among the items. The formula for α is:

$$\alpha = \left(\frac{p}{p-1} \right) \frac{Covs}{Vars + Covs},$$

where p is the number of items (7 in the cases in Table 1), "*Covs*" is the sum of all the (off-diagonal) item–item covariances, and "*Vars*" is the sum of all the (diagonal) item variances.

Example A illustrates that a high value Cronbach's α will occur when all the items are indeed homogeneous in terms of having the same, moderate item–item correlations. However, Examples B and C illustrate that equally well a high Cronbach α occurs when two distinct subtests exist—Items 1 through 3 and Items 4 through 7—such that items *within* a subtest intercorrelate highly, whereas the intercorrelations *between* subtests are small (Example B: $r = 0.5$ within subtest, $r = 0.25$ between subtest), or even zero (Example C: $r = 0.85$ and 0 within and between, respectively).

Clearly, then, with a small set of items, a high Cronbach α indeed confirms the possibility that the items are "homogeneous" or "unidimensional," but it simultaneously confirms the possibility that the items are not "homogeneous" or "unidimensional"; and thus it can hardly be cited as substantial evidence supporting "unidimensionality." Hence the caution quoted previously from Cronbach's (1951) original paper.

COHEN'S KAPPA

When participants, or patients, require classification into a dichotomous state—for example, sick (1) versus well (0)—such a judgement is frequently made by a rater.

TABLE 1
Different Configurations of Item Variances and Covariances

Item	*1*	*2*	*3*	*4*	*5*	*6*	*7*
A. One construct							
1	10	3.6	3.6	3.6	3.6	3.6	3.6
2	3.6[a]	10	3.6	3.6	3.6	3.6	3.6
3	3.6	3.6	10	3.6	3.6	3.6	3.6
4	3.6	3.6	3.6	10	3.6	3.6	3.6
5	3.6	3.6	3.6	3.6	10	3.6	3.6
6	3.6	3.6	3.6	3.6	3.6	10	3.6
7	3.6	3.6	3.6	3.6	3.6	3.6	10
B Two moderately correlated constructs							
1	10	5	5	2.5	2.5	2.5	2.5
2	5	10	5	2.5	2.5	2.5	2.5
3	5	5	10	2.5	2.5	2.5	2.5
4	2.5	2.5	2.5	10	5	5	5
5	2.5	2.5	2.5	5	10	5	5
6	2.5	2.5	2.5	5	5	10	5
7	2.5	2.5	2.5	5	5	5	10
C. Two uncorrelated constructs							
1	10	8.5	8.5	0	0	0	0
2	8.5	10	8.5	0	0	0	0
3	8.5	8.5	10	0	0	0	0
4	0	0	0	10	8.5	8.5	8.5
5	0	0	0	8.5	10	8.5	8.5
6	0	0	0	8.5	8.5	10	8.5
7	0	0	0	8.5	8.5	8.5	10

Note. All with Cronbach's $\alpha = .80$.

[a]The correlation between Items 1 and 2 in this configuration would be 3.6/{SQRT(10).SQRT(10)} = 0.36.

When ratings are made twice on the same set of subjects, the data can be arranged in a 2×2 table. The two ratings might reflect independent ratings on two occasions by the same rater (addressing intrarater reliability), or they might reflect independent ratings made by two raters (addressing interrater reliability). The data can be arranged as in the top panel of Table 2, where p_{10} is the proportion of subjects who simultaneously are rated as 1 by rating 1 and 0 by rating 2, and so on. Cohen's κ is a measure of the concordance between the ratings, the degree to which (1,1) and (0,0) pairs of ratings dominate (1,0) and (0,1) ratings, taking into account that (1,1) and (0,0) will occur by chance alone even when the two ratings are made randomly. The formula is:

$$\kappa = \frac{p_{\text{observed}} - p_{\text{chance}}}{1 - p_{\text{chance}}}, \text{ where}$$

$$p_{\text{observed}} = p_{11} + p_{00} \text{ and } p_{\text{chance}} = p_{1*}p_{*1} + p_{0*}p_{*0} \text{ and } p_{1*} = p_{11} + p_{10}, \text{ etc.}$$

TABLE 2
Different Configurations

Rating 1	*Rating 2*		
	Well (0)	*Sick (1)*	
Notation for κ			
Well (0)	p_{00}	p_{01}	p_{0*}
Sick (1)	p_{10}	p_{11}	p_{1*}
	$\kappa = \frac{(p_{11} + p_{00}) - (p_{1*}p_{*1} + p_{0*}p_{*0})}{1 - (p_{1*}p_{*1} + p_{0*}p_{*0})}$		
Example A			
Well (0)	0.363	0.137	0.5
Sick (1)	0.137	0.363	0.5
	0.5	0.5	1
Example B			
Well (0)	0.35	0	0.35
Sick (1)	0.3	0.35	0.65
	0.65	0.35	1

Note. All with Cohen's $\kappa = 0.45$.

When we summarize data like that in each 2×2 table of Table 2 with a single summary index, such as κ, many distinct configurations of data will underlie the same summary value. In a 2×2 table of joint proportions (constrained to add to 1), three functionally independent parameters are required to unambiguously represent the original data. Hence the same κ value will apply to many different data configurations. However, even the same data configuration may be underlain by quite distinct processes.

Figure 2 attempts to illustrate these comments. Consider the data in Example A, the middle panel of Table 2. These data yield a κ of 0.45, apparently indicating moderate ratings' reliability. However, such data can arise when both ratings are equally moderately reliable, or when one rating is "perfect," but the other relatively unreliable. For instance, if we imagine a process where there exists an underlying latent trait, with unit normal distribution in the population, and each rating can be characterized by an ogive curve imposed on this latent trait: For a given trait value on the x-axis, the curve height (the y coordinate) represents the probability of the rating deciding that subjects with that trait value are sick rather than well. A rating that does not treat all subjects with the same latent trait value in the same way—as either all sick or all well—is behaving unreliably at that latent trait level; and a step-function therefore indicates perfect rating behavior at all latent trait levels. Thus, the slope of a rating's ogive characteristic curve in such a model represents the overall reliability of that rating. Figure 2(i) shows a case where both rating characteristic curves are the same, and data as in Example A are

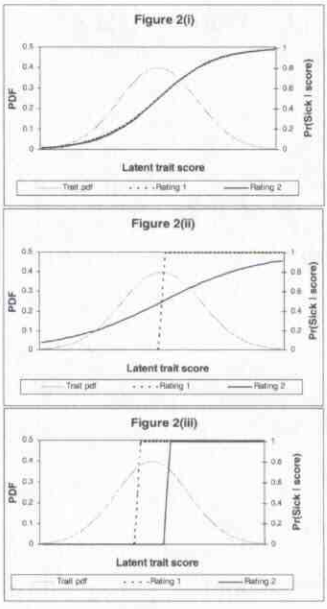

FIGURE 2 Ratings' characteristic curves.

generated. However, the same data are generated by the situation in Figure 2(ii), where Rating 2 is quite noisy, but Rating 1 is perfect: participants with trait value above zero being unambiguously identified as sick, whereas those below zero are all identified as well. Both these scenarios yield the same 2×2 data, and thus the same value of $\kappa = 0.45$.

However, such a κ value, of 0.45, can even arise when both ratings are in some sense perfect. Consider the data in Table 2 Example B, where $\kappa = 0.45$ as well: The absence of any participants at all who simultaneously are judged well by Rating 1 and judged sick by Rating 2 indicates a complete absence of errors of misclassification. These data were generated by the situation displayed in Figure 2(iii): Both ratings are perfect, in the sense of having "step-function" characteristics, but they are operating at different levels of *severity* on the underlying trait, and hence the zero entry in the 2×2 table.

These examples show that simply knowing the value of κ tells us little if anything about which rating, if either, is better or worse, or even good or poor. (Thompson & Walter, 1988, made much the same points.)

NORMALITY AND MLR

The linear regression model underlies many of the (fixed effects) statistical techniques that we use in psychology (e.g., *t* test, analysis of variance [ANOVA]). In

MLR we wish to model a dependent y-variable with a linear function of p x-predictors and an r-residual. In the population of interest, we presume:

$$y = \alpha + \beta_1 x_1 + \ldots + \beta_p x_p + r.$$

If we have a limited set of data that we presume is randomly sampled from subpopulations defined by holding the x-predictors constant, then we can use the sample data (y, x_1, ... , x_p) to estimate this linear equation, using ordinary least squares (OLS):

$$y = \hat{\alpha} + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_p x_p + \hat{r}.$$

To perform OLS we assume in the population:

(a) $E(r|x_1, \ldots, x_p) = E(r|xs) = 0$, for every profile of x-scores;

that is, given any distinct profile of x predictors in the population, the average (or expectation) of the residuals is zero. The OLS process of estimation process *ensures* that our estimates will be unbiased for the true population parameters, over imaginary repeated samplings and estimations, a valuable property all on its own. (No assumption of normality underlies this useful property. This is just as well, as in psychology, it is rare to have any variable that is in truth even genuinely continuous, let alone normally distributed!)

Further assumptions need to be made to enable us to obtain the distributions of how these $\hat{\beta}$ estimates vary over repeated samples, and so to obtain p values allowing us to perform statistical tests of hypotheses. Not surprisingly, these assumptions involve the distributions in the population of the r residuals: (b) $r|xs \sim N(0, \sigma^2_{xs})$, for each profile of x scores (the assumption of normality), and (c) $\sigma^2_{xs} = \sigma^2$, constant, over all profiles of x predictors (the assumption of "homoscedasticity," homogeneity of variance); that is, in the population, given any fixed profile of the x predictors, we assume that the r residuals for that x profile are normally distributed, and that the variances of these normal distributions are the same over all x profiles.

Now the important thing to notice about the distributional assumptions in (b) and (c) is that they concern the r residuals (conditional on x profiles), and they do *not* concern the (unconditional) distribution of y other than indirectly, and even then y distributions within subgroups of subjects with the same x profile. Thus very little can be said about the truth or falsity of these assumptions from a simple inspection of the (unconditional) y distribution in our sample of data, and whether that distribution is normal in shape. Consider the following example: Suppose we are interested in the effect on depression (y) of a particular therapy, and we perform a trial with treatment

($x = 0$) and control ($x = 1$) groups. Suppose that we have equal numbers of subjects in each group, and that in the (theoretical) parent populations, conditions (b) and (c) are met: the residuals in the control population are normal with variance σ^2, and the residuals in the treatment population are normal with the same variance. Finally suppose that the effect of treatment on y is to reduce depression scores by β. Then the (population) MLR model with one predictor (x) will be: $y = \alpha + \beta x + r$. The mean y score in the control population will be $\alpha + \beta = \alpha + \beta \times 1 + 0$, while the mean in the treatment population will be $\alpha = \alpha + \beta \times 0 + 0$. Our combined sample data can be viewed as a random sample drawn from the population shown in Figure 3a (subject to the constraint of equal control and treatment sample sizes). This distribution is hardly normal in shape, and so we would hardly expect our sample to be near normal. Yet, *all* the conditions for MLR statistical tests (or the equivalent two-sample t test) are nonetheless met in this example. Figure 3b shows a similar example, except that the control group now forms 70% of our combined sample (rather than 50%); the combined sample data would here show positive skew of the type that often accompanies the incorrect interpretation that use of MLR here would be inappropriate because "the data are not normal."

Let us look at an example of this distinction based on actual data, between the unconditional y distribution and the conditional r distributions. In the Sydney Older Persons' Study (Waite et al., 2001), data were collected relating to depression among community-dwelling older folk (aged 75 or more) –y: the CESD (Centre for Epidemiologic Studies Depression) scale (Radloff, 1977; scores 20 to 80

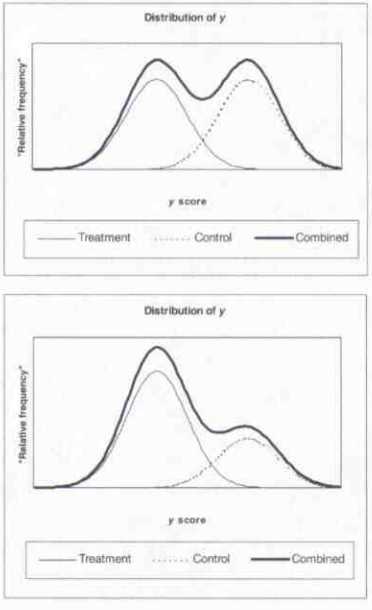

FIGURE 3 Populations in which MLR (multiple linear regression) assumptions are met yet unconditional y data are nonnormal.

with high scores reflecting more depression). Data were also collected on two disability variables – x_1: activities of daily living (ADL; bathing oneself, feeding oneself, etc); and x_2: instrumental activities of daily living (IADL; shopping, home maintenance, etc). Higher scores on the ADL and IADL scales reflected more disability of the relevant sort, and both these scales were scored such that a zero reflected a total absence of the relevant disability. Both the ADL and the IADL measures in this sample were highly positively skewed. The data from 506 subjects were submitted to an MLR predicting CESD from ADL and IADL scores, with R^2 of only 12.4%. The estimated MLR model was:

$$\text{CESD} = (26.63 - 3.13 \times \text{ADL} + 4.02 \times \text{IADL}) + \hat{r}.$$

We can compute for each subject their predicted value:

$$\text{PREDICTED VALUE} = 26.63 - 3.13 \times \text{ADL} + 4.02 \times \text{IADL},$$

and their corresponding estimated residual:

$$\text{Residual} = \hat{r} = \text{CESD} - (\text{Predicted Value}).$$

The top panel of Figure 4 shows the unconditional distribution of the CESD y scores, and we can see that there is substantial positive skew in these data (the superimposed normal curve in all panels has the same mean and *SD* as the actual data plotted in that histogram). However, in large part *this skew will be attributable to the skewed predictors*, which thus yield a highly skewed Predicted Value portion of the CESD scores, as is apparent in the middle panel of Figure 4. When we plot the Residual scores, (pooling estimates of conditional residuals), we see that they have nowhere near the same amount of skew, albeit some remains.

The point to be made here is that the distributional assumptions (b) and (c) made in MLR apply to the r numbers estimated by the data graphed in the third panel, *not* to those in the first panel—graphing the unconditional y data tells us little about the truth or falsity of (b) and (c). Further discussion of this point can be found in Cohen, Cohen, West, & Aiken (2003, chapter 4).

TRANSFORMATIONS TO ACHIEVE NORMALITY

In cases where the conditional residuals may well be (positively) skewed, common statistical advice exists to transform the y data to y^* data in such a manner that larger scores become relatively compressed while smaller ones are relatively expanded, for example, $y^* = \log(1 + y)$. The aim is to make conditions (b) and (c) more closely met, thus making our inferences based on p values more valid. However, such a

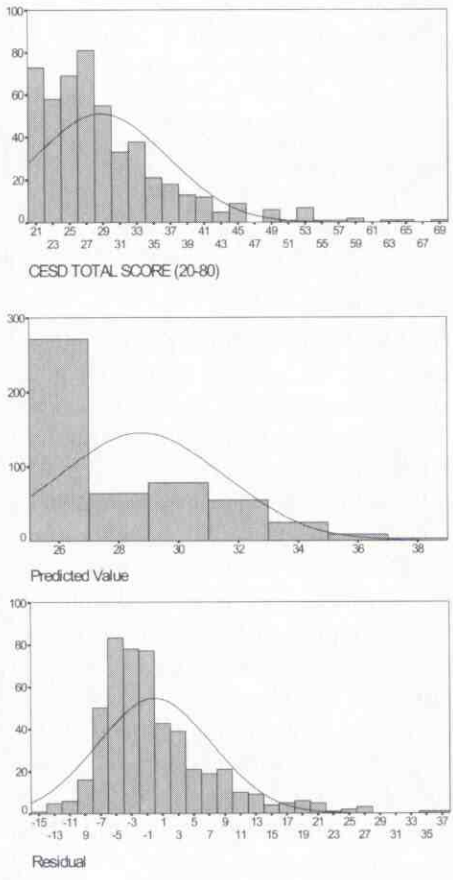

FIGURE 4 MLR (multiple linear regression) on skewed data ($n = 506$).

process is not free of empirical or scientific implications, implications that may well outweigh any statistical gain. Poor though psychological measurement is, it usually does have *some* empirical content, particularly with well-used scales on which a larger body of research already literature exists.

Consider an example of an ability test, y, scored 0 to 100, higher scores indicative of higher abilities. Suppose our research is aimed at understanding why differences exist on the scores from this test among various subjects. Suppose, as in Table 3, we have 4 participants with the given y scores. When we contemplate the original y-scores, we note that these data could be modeled with an effect of age, increasing scores by 9 points as we move from young to old; an effect of sex, increasing scores by 9 points as we move from male to female; and finally a *dominating* interaction effect of 90 points that is invoked only when participants are both

TABLE 3
Original and Transformed Ability Scores

	Tom	*Fred*	*Sue*	*Mary*
Original y score	0	9	9	99
New y^* score: $y^* = \log(1 + y)$	0	1	1	2
Possible predictors				
Sex	Male	Male	Female	Female
Age	Young	Old	Young	Old

female and old (i.e., only for Mary). When we transform to y^* data, the effects of age and sex, as with y, remain equal (at just 1 y^* unit), but interaction is totally absent. In transforming our scores (for statistical purposes) we have surrendered whatever empirical meaning the original y scores had, and we appear now to be talking about a different empirical construct y^*. This issue is particularly dangerous when such statistical transformations are done without the accompanying realization that we are now addressing an empirical construct *different* from that we set out originally to measure. To give this caricature example more substance, imagine that y is a deliberate measure of "geniusness," on which Mary is outstanding, while Fred and Sue are pretty typical, and Tom is a "bit thick." In such a case we would be interested in analyzing the variance between Mary on the one hand, and the rest, almost as a group, on the other; we would be seeking determinants of Mary's exceptionality. When we transform to y^*, and analyze variance on this scale, determinants that predict Tom's divergence from Fred and Sue might well seem to play just as much of a role as those separating Mary's score from those of Fred and Sue; we would be analyzing an empirical construct that we might refer to as "dimness–brightness." This point may have more relevance in clinical contexts where symptom scales on general populations would be expected to be highly positively skewed; in transforming we may well be moving from the desired construct of, say, "depression," to one of "euphoria–sadness" in which we simply are not empirically interested.

In a conflict between desired empirical meaning and valid p values, there will be many situations where having valid p values on a transformed scale (in which we are *not* empirically interested) is not of empirical interest. This is a good time to reiterate that the property of OLS unbiasedness does *not* depend on the distributional assumptions (b) and (c), and at least we have the former without the need for any nonlinear transformation. There will be, of course, many other situations where a good scientific rational exists for transforming data, and the previous comments should not be misread as general advice to never transform one's data, but rather as an acknowledgment that such transformations can sometimes have empirical or scientific implications, which often may outweigh any purely statistical benefit. Transformations made only for statistical reasons

can have the unfortunate consequence of "disconnecting" observed effects from what empirical meaning our raw units have, little though that often may be in social sciences.

THE "FRAGILITY" OF POWER COMPUTATIONS

The computation of power can be very useful in the few situations where we know enough to explicitly compute it. However such situations are very rare, and most actual computations rely heavily on assumptions that usually are tenuous. Let us examine some situations where we might contemplate computing power, and point out some problems with such a computation.

Suppose we are planning an experiment where we will perform a one-sample, two-tailed t test on a set of n sampled y scores. Then the following five quantities are bound together in the sense that if we know all but one of them we can compute it from the others: the power $(1 - \beta)$, the Type I Error Rate (α), the sample size (n), the putative size of effect (μ, in the same units as y), and the population standard deviation of the y-scores (σ^2). (Note that "effect size" μ is here defined in terms of unstandardized raw scores, and not in terms of standardized raw scores, as with, e.g., "Cohen's d"; in this way the influence of σ^2 is clearly isolated in the following discussion .) The power of our planned statistical test can be ascertained when we have known values of the other four quantities:

$$1 - \beta = fn(n, \mu, \alpha, \sigma^2).$$

The α rate, the size of effect and the sample size we can fix in advance. However, to ascertain power, we then still need to know the y-variance σ^2. Typically we do not know this, so we must use a *guess*. We might use the estimate obtained from some other study (perhaps a pilot study for this project, perhaps a "similar" study done previously), based on m subjects. Unfortunately, such an *estimate*, s^2, is not the population quantity σ^2 that our power computation needs. Indeed, as an estimate of σ^2, s^2 is quite variable, becoming less variable as m increases. Under (the usual) assumptions of data normality and random sampling, the quantity:

$$\frac{(m-1)s^2}{\sigma^2},$$

is known to have a chi-square distribution with $m - 1$ degrees of freedom, with known tabulated values. We can thus use m and s^2 from our pilot study to obtain a confidence interval (CI) for the inclusion of σ^2. For example, if we wish to find the upper and lower 5% values for σ^2 (i.e., a 90% CI), we solve:

$$\frac{(m-1)s^2}{\hat{\sigma}^2_{\text{lower}}} = \chi^2_{m-1}(0.95) \text{ and } \frac{(m-1)s^2}{\hat{\sigma}^2_{\text{upper}}} = \chi^2_{m-1}(0.05), \text{ obtaining}$$

$$\hat{\sigma}^2_{\text{Lower}} = \frac{(m-1)s^2}{\chi^2_{m-1}(0.95)} \text{ and } \hat{\sigma}^2_{\text{upper}} = \frac{(m-1)s^2}{\chi^2_{m-1}(0.05)}.$$

(Note that as $\chi^2_{m-1}(0.95) > \chi^2_{m-1}(0.05), \hat{\sigma}^2_{\text{lower}} < \hat{\sigma}^2_{\text{upper}}$.) In addition, now we compute the corresponding, 90% *bounds on power*, taking into account this fact that s^2 itself is just a noisy estimate:

$$\text{Estimated power}_{\text{upper}} = 1 - \hat{\beta}_{\text{Lower}} = fn(n, \mu, \alpha, \hat{\sigma}^2_{\text{lower}}).$$
$$\text{Estimated power}_{\text{lower}} = 1 - \hat{\beta}_{\text{upper}} = fn(n, \mu, \alpha, \hat{\sigma}^2_{\text{upper}}).$$

TABLE 4
Variability in Two-Tailed Power due to Estimation of Error Variance

	s = 1		*s* = 2		*s* = 4	
n	*Low Power*	*High Power*	*Low Power*	*High Power*	*Low Power*	*High Power*
α = .05, 80% CI for power						
10	62	91	21	38	8	13
20	93	100	40	70	14	24
30	99	100	57	88	19	34
40	100	100	70	95	24	44
50	100	100	80	98	29	53
α = .05, 95% CI for power						
10	53	95	17	44	8	15
20	86	100	33	78	12	27
30	97	100	48	92	16	40
40	99	100	60	98	20	51
50	100	100	70	99	24	60
α = .01, 80% CI for power						
10	32	67	7	15	2	4
20	76	99	18	43	4	9
30	94	100	31	68	6	15
40	99	100	45	84	9	21
50	100	100	57	93	12	29
α = .01, 95% CI for power						
10	24	76	5	18	2	4
20	64	100	14	51	3	10
30	87	100	24	77	5	18
40	96	100	34	90	7	26
50	99	100	45	96	9	35

Note. Given in percentages. Low, high pairs represent the specified CI for power at specified α. All computations are based on an effect size of μ = 1 *y* unit and an estimated standard deviation of *s* from a (pilot study) sample size: *m* = 25. CI = confidence interval.

As a concrete example, suppose we ran a pilot study with $m = 25$ participants, obtaining an estimated *SD* of s. Table 4 presents some computations of power CIs for various configurations. For instance, we might be interested in detecting an effect size of $\mu = 1$ y unit, using $\alpha = 0.05$ two tailed, and we might wish to compute power for a study of n subjects. Then if $n = 10$ (only) and the estimate of s we obtained from the pilot study with $m = 25$ subjects was $s = 1$, the 80% CI for power would be (62%, 91%). If our pilot study estimate of variability was larger, $s = 4$, say, then even for a "large" planned study size of $n = 50$, the true power would have 80% CI of (29%, 53%); and the 95% CI for power under these conditions would be (24%, 60%).

If, rather than perform a pilot study (under putatively "identical conditions"), we sought to use a value of s^2 from some other researcher's large study, variation of unknown amount would be expected between the unknown σ^2 for our study and that of the other researcher's when that other study and our planned study are not identical. For example, if our planned study was a supposed improvement on the other study, in terms of better control of extraneous factors, we would expect the s^2 from the other study to be estimating that other study's σ^2, itself an overestimate of the σ^2 needed for the calculations of the power of our study.

Finally, if we plan (sensibly, from a scientific point of view) to measure and adjust for potential covariates, we would have little idea of the effect of such control on our error variance until *after* we have collected our data (with their as yet unknown intercorrelations), making any preexperimental power calculation rather more speculative than frequently seems to be discussed. See Wilkinson and The APA Task Force on Statistical Inference (1999) for recent views on this issue, and see Macdonald (2002) for critical discussion of the same issues.

SIMULTANEOUS CONTROL OF TYPE I ERRORS (α_{Fam} CONTROL)

To apprehend just what is achieved with α_{Fam} control, we first need to discuss what is achieved with the use of α control in a single decision context. One common perspective on the use of p values (there are others; see, e.g., Macdonald, 2002) is the following: When we perform a *single* statistical test, we conventionally choose a significance level of $\alpha = 0.05$ (or perhaps 0.01). This choice is purely conventional, despite the fact of its overwhelming institutionalization in the social sciences. We assume the null hypothesis is true (as well as all the other conditions underlying the use of the statistical model involved; see Macdonald, 2002), and compute a p value associated with our observed data, which, if smaller than the preset α, we interpret as evidence against that null. This p value is exact if the null hypothesis to which it is addressed is true (as well as the other assumed conditions). If this null hypothesis is false, this p value is in fact not literally accurate. We never know whether the null—on which our p value computation is based—is true or false. Indeed, there are those who argue that no null hypothesis is ever true (e.g., Cohen, 1994).

This whole process of statistical inference is error-prone—it minimizes errors of one kind, for every choice of an error rate for errors of the other kind. In particular, if the null is false, and we compute a p value under the (erroneous) assumption that it is true such that the observed p value $> \alpha$, then we will make the mistake of not correctly rejecting the null (a Type II Error, with rate β for given nonzero effect size). The error rates α and β "trade off" against each other: We can preset a smaller α rate, ensuring a smaller probability of Type I Error—but only at the expense of simultaneously increasing β, the probability of errors of false acceptance of the null.

This is pretty circuitous or even "turgid" reasoning, and there are many who question its use and even its validity (for some history, see Grayson, 1998; Nickerson, 2000). To the thinking social scientist, it is not a process of "iron-clad" rigor capable of unerringly distilling scientific "truth" from noisy data.

When we turn to multiple testing (or "simultaneous inference") the reasoning becomes even more convoluted. Consider a one-way ANOVA layout with k groups of data, n observations in each group, and suppose we plan to test a large number of as yet unspecified contrasts. We might plan to use the Scheffé method. Here the observed F statistic for each given 1 DF contrast, $F_{1,k(n-1)}(\text{Obs})$, is compared for significance to $(k-1)F_{k-1,k(n-1)}(\alpha_{\text{Fam}})$, where α_{Fam} is the "family-wise" Type I Error rate, and each contrast null is rejected when:

$$F_{1,k(n-1)}(\text{Obs}) > F_{\text{Crit}} = (k-1)F_{k-1,k(n-1)}(\alpha_{\text{Fam}})$$

This procedure guarantees the following: No matter how many contrasts we test against this criterion, the probability of *at least one* Type I Error is kept limited to a maximum of the preset α_{Fam}. So, if we had good reason in a particular application to be obsessed *only* with avoiding wrong rejections of contrast-nulls, this might be a good procedure to use. Another procedure would be to focus on Type I Error control at the level of each contrast test. Here we would reject a given null only when:

$$F_{1,k(n-1)}(\text{Obs}) > F_{\text{Crit}} = F_{1,k(n-1)}(\alpha_{\text{Dec}})$$

where a decision-wise Type I Error rate of α_{Dec} is used, ensuring that when this contrast-null is true, the probability of Type I Error in just this one decision is limited to α_{Dec}.

Not surprisingly, the greater Type I control afforded by the Scheffé procedure occurs by making it (substantially!) harder to gain null rejection on any contrast. For instance, if $k = 7$ groups, each with $n = 20$ observations, the procedures each with 0.05 control of their own sort lead to critical values:

$$\text{Scheffé: } F_{\text{Crit}} = 6.F_{6,133}(\alpha_{\text{Fam}} = 0.05) = 6 \times 2.1674 = 13.0045,$$
$$\text{Decision wise: } F_{\text{Crit}} = F_{6,133}(\alpha_{\text{Dec}} = 0.05) = 3.9123.$$

Clearly, a given $F_{1,k(n-1)}(\text{Obs})$ from our data on a given contrast will more readily exceed 3.9123 than it will exceed 13.0045, leading more readily to null rejection of that contrast null. Indeed, we can ask: if we used a α_{Fam}-critical value of 13.0045 as our contrast-rejection criterion, what would the corresponding α_{Dec} be? It transpires that:

$$F_{1,133}(\alpha_{\text{Dec}} = 0.00044) = 13.0045.$$

That is, using the Scheffé procedure with $\alpha_{\text{Fam}} = 0.05$ is *exactly equivalent* to using a decision-wise procedure with $\alpha_{\text{Dec}} = 0.00044$. Similarly, with 3.9123 as the critical value, the decision-wise procedure with $\alpha_{\text{Dec}} = 0.05$ is equivalent to using the Scheffé procedure with $\alpha_{\text{Fam}} = 0.668$, as:

$$6.F_{6,133}(\alpha_{\text{Fam}} = 0.668) = 6 \times 0.65205 = 3.9123.$$

So, in summary, both procedures are identical, in that over the set of contrasts of concern the $F_{1,k(n-1)}(\text{Obs})$ datum for each contrast is compared to a fixed critical value, F_{Crit}. For the Scheffé procedure that fixed value generally is *much* larger than for the decision-wise procedure with the same nominal α-rate; alternatively, the Scheffé procedure at α_{Fam} is equivalent to a decision-wise procedure at some ascertainable value of α_{Dec} that is (much) smaller than α_{Fam}. In this way the Scheffé procedure achieves its far greater "family-wise" Type I Error control.

The rationale for the use of methods like the Scheffé method is based entirely on control of (all) Type I Errors. However, what happens when not all the contrast-nulls are true, which surely must be the usual situation (particularly if, like Cohen, 1994, we believe no single "nil" hypothesis is *ever* true!)? Such a strict criterion for rejection must lead to many Type II Errors when contrast-nulls are false. In fact, the price we pay for such severe Type I Error control can be dramatic.

To illustrate this, suppose we consider a $k = 7$ one-way layout with $n = 20$ observations in each cell; suppose the true population means are $\mu_1 = \mu_2 = 1, \mu_3 = \mu_4 = \ldots = \mu_7 = 5$. Consider the nulls for the set of contrasts consisting of all pair comparisons:

$$H_{i,j}: \mu_i - \mu_j = 0, 1 \leq i < j \leq 7.$$

Of these 21 possible comparisons, 11 nulls will be true (where only Type I Errors can occur), 10 where $3 \leq i < j \leq 7$ and one where $i = 1, j = 2$; and 10 will have false nulls (where only Type II Errors can occur), where $1 \leq i \leq 2$ and $3 \leq j \leq 7$.

Once we know the population standard deviation of the observations, we can work out the power for any testing procedure, given that procedure's F_{Crit}. Table 5 presents the results associated with different procedures. Let us examine the upper panel in some detail. The standard deviation is set equal to 5 here, and so the contrast effect size when the null is false is, for instance, $\mu_6 - \mu_1 = 5 - 1 = 4$, or 0.8 stan-

dard deviations; the power of the overall or "omnibus" F-test for this configuration is 91%. In row 1 we use $F_{Crit} = 3.9123$, based on a decision-wise procedure with $\alpha_{Dec} = 0.05$ (this is equivalent to a Scheffé procedure with $\alpha_{Fam} = 0.688$, as we saw previously); the Type II Error rate is $\beta = 0.2836$, for the 10 comparisons where the nulls are false. From these rates we can ascertain that the expected number of Type I Errors among the 11 true-null comparisons is 0.55 (= 11 × 0.05), whereas the expected number of Type II Errors among the 10 false-null comparisons is 2.836 (= 10 × 0.2836), yielding a total expected number of either type of error: 0.55 + 2.836 = 3.836 out of 21 statistical decisions. Row 2 shows the same quantities for the Scheffé procedure, where $F_{Crit} = 13.0045$ is used. Although the Scheffé procedure clearly makes a lot fewer Type I Errors (0.005 out of 11), this is at the expense of a much higher yield of Type II Errors (8.596 out of 10). The fourth row shows the same parameters for an optimal procedure, based on minimizing the total number

TABLE 5
Decisional Consequences of Scheffé and Tukey HSD Simultaneous Inference Procedures

					Decisional Errors		
Procedure	F_{Crit}	α_{Dec}	α_{Fam}[a]	β	*Type I*[b]	*Type II*[c]	*Total*[d]
Effect = 0.8 *SD* (Omnibus $F_{6,133}(0.05)$ – test power = 91%)							
Decision wise	3.912	0.050	0.688	0.284	0.550	2.836	3.386
Scheffé	13.005	0.0004	0.050	0.860	0.005	8.596	8.601
Tukey HSD	8.694	0.004	0.050	0.675	0.041	6.754	6.796
Optimal	2.349	0.128	0.883	0.135	1.406	1.348	2.754
Effect = 1 *SD* (Omnibus $F_{6,133}(0.05)$ – test power = 99%)							
Decision wise	3.912	0.050	0.688	0.100	0.550	1.003	1.553
Scheffé	13.005	0.0004	0.050	0.680	0.005	6.803	6.807
Tukey HSD	8.694	0.004	0.050	0.424	0.041	4.239	4.280
Optimal	3.317	0.071	0.767	0.070	0.779	0.705	1.483
Effect = 1.5 *SD* (Omnibus $F_{6,133}(0.05)$ – test power = 100%)							
Decision wise	3.912	0.050	0.688	0.001	0.550	0.009	0.559
Scheffé	13.005	0.0004	0.050	0.124	0.005	1.240	1.245
Tukey HSD	8.694	0.004	0.050	0.029	0.041	0.285	0.327
Optimal	6.712	0.011	0.355	0.010	0.117	0.097	0.214
Effect = 2 *SD* (Omnibus $F_{6,133}(0.05)$ – test power = 100%)							
Decision wise	3.912	0.050	0.688	0.0000002	0.550	0.000002	0.550
Scheffé	13.005	0.0004	0.050	0.002	0.005	0.020	0.025
Tukey HSD	8.694	0.004	0.050	0.0001	0.041	0.001	0.043
Optimal	11.451	0.001	0.084	0.001	0.010	0.008	0.018

Note. One-way analysis of variance with $k = 7$ cells, $n = 20$, subjects per cell, with population cell means (1, 1, 5, 5, 5, 5, 5); all pair comparisons two-tail tested (21 in total, with 11 nulls true and 10 nulls false).

[a]Entries are Scheffé family rates for all but Tukey HSD rows.[b]Expected number of Type I errors among 11 comparisons where null hypothesis is true. [c]Expected number of Type II errors among 10 comparisons where null hypothesis is false. [d]Expected number of errors among all 21 comparisons.

of errors of either kind. (To plan to use such an optimal procedure, of course, one would need to know in advance the pattern of cell means, obviating the need for any research at all!) The remaining three panels show similar results for progressively smaller standard deviations, or larger effect sizes.

The results in Table 5 are dramatic and sobering. So conservative is the Scheffé procedure that, in terms of total number of errors, it only outperforms the decision-wise procedure with the same nominal α for the most extreme example considered, an effect size of 2 *SD*s. For smaller effect sizes, optimal procedures are equivalent to decision-wise procedures with α_{Dec} in excess of 0.05. Indeed, the Scheffé procedure makes 2 to 4 times as many errors (mostly Type II) as the decision-wise procedure for all but the largest effect size. Remember, when we say smallest effect size, we are still discussing a configuration with omnibus *F*-test power of 91%; indeed, the "largest" effect size in Table 5 (2 standard deviations) is so large that statistical analysis of any sort seems hardly required—the β-rates are effectively zero irrespective of α_{Dec}; and the means of 1 and 5 are separated by over 8 standard-errors-of-the-mean ($(\sqrt{20} \times 2SD)/SD$).

Although these observations are specific to the particular configuration of cell means used in Table 5, and although the Scheffé method is not always recommended when we are interested only in pair comparisons (as opposed to, say, "Tukey's HSD"; also shown in Table 5, in the third row of each panel), the general principle will still hold: In many or most situations, setting a large F_{Crit} (via Scheffé or any other procedure) will indeed inhibit Type I Errors, but only by proliferating Type II Errors, and a "golden rule" about simultaneous α_{Fam} control pertains only when Type II Errors are of little concern—a situation one would expect or hope to be extremely rare in a creative social science.

MANOVA

The technique of multivariate analysis of variance (MANOVA)—a mnemonic representing a statistical analysis *design*, not to be confused with statistical subprograms with the keyword "MANOVA"—frequently is recommended for use in designs involving several dependent *y* variables.

What is MANOVA, and what does it achieve? Consider a situation with a simple design: $x = 0$, control and $x = 1$, treatment, where we measure a set of p dependent variables: $y_1, \ldots, y_p$. For each *y*-variable, we could conduct a univariate *F* test, examining whether control and treatment groups differ on that *y*-variable. If we use α_{Dec} for each of these p tests, we would have a family-wise Type I Error rate well in excess of the decision-wise rate individually used, accumulating over multiple tests in much the same manner as discussed in the preceding section. One common reason why MANOVA is recommended is that it provides family-wise control in this situation. It achieves this as follows: Using all p *y* scores for each

subject, we can imagine constructing a new univariate score as a linear combination of the original p y scores:

$$y_c = c_1 y_1 + \ldots + c_p y_p,$$

and then submitting this new univariate y_c score to an F test, obtaining an observed value F_c. If we let the linear combination c range over all possibilities, we can contemplate that particular linear combination that gives the *maximum* F_c value for the design or effect in question. This is precisely what MANOVA does, finds the maximum F_c value, providing also a p value based on the (null) assumption that there really is no difference in any of the y scores associated with the design variable X. Because individual y scores themselves are linear combinations, for example,

$$y_1 = 1.y_1 + 0.y_2 + \ldots + 0.y_p,$$

and the p value is testing the *maximal* linear combination, family-wise control is achieved. In addition, just as in the previous section, the price we would expect to pay for family-wise control of Type I Errors is Type II Errors. As a by-product of the analysis, we obtain the maximal linear combination, and when significant, are frequently invited to interpret it empirically.

Broadly speaking, research can be classified as exploratory or confirmatory. When it is exploratory, and we have no theory of the interplay among our y variables, let alone possible differential effects of x on them, and when we have good reason for conservative family-wise control, MANOVA may well be useful. Even here, a significant maximal linear combination may be difficult to interpret empirically (we will see why soon), and researchers typically return to looking at the univariate F values and effects of x on individual y variates, rather obviating the need for MANOVA in the first place.

In cases where the research is confirmatory, however, it is very difficult to understand empirically just what the maximal linear combination may mean, and MANOVA can be a very "blunt weapon" indeed. A sensible alternate strategy in some situations may be to combine the y-measures oneself into a meaningful, single variable, based on theoretical, *a priori* grounds, and proceed with a simple univariate analysis. Some examples of structured y, x data will help to illustrate this point. Consider a case where we have the simplest possible design ($x = 1$-0 for Treatment versus Control, equal group sizes; and MANOVA specializes to Hotelling's T^2), and we have $p = 6$ y variables, the last 5 of which relate to a single common factor θ:

$$\begin{aligned} y_2 &= \theta + u_2, Var(u_2) = 1, \\ &\vdots \\ y_6 &= \theta + u_6, Var(u_6) = 1, \end{aligned}$$

where all the latent residuals are uncorrelated. Suppose also that these y variables are influenced by x via only their common factor:

$$\theta = \beta x + r, \; Var(r) = 1,$$

where r is uncorrelated with the u residuals above. Finally, suppose the first y variable is influenced by both θ and directly by x:

$$y_1 = \theta + x + u_1, \; Var(u_1) = 1,$$

where u_1 is uncorrelated with r and the other u residuals.

A (caricature) concrete example might involve a treatment for depression (x), which we know leaves subjects physically debilitated, but we hope with improved mood. With all higher scores reflecting greater morbidity levels, then θ might represent a single common factor of depression-mood, of which y_2 to y_6 are pure symptom measures (such as "I feel blue," "I am unhappy," ...); while y_1 is biased by physical aspects as well as indexing depression mood (e.g., "Everything is an effort"). The effect, β, of the treatment on depression mood (θ) would be negative or positive according to whether the treatment, respectively, worked successfully or unsuccessfully to lower θ.

Then under these conditions:

$$y_1 = (\beta + 1)x + r + u_1, \text{ and } y_i = \beta x + r + u_i, \; 2 \leq i \leq 6.$$

Table 6 presents examples of such a configuration for three different values of β, the effect carried to all y-variables via the common factor. In example (a), $\beta = -1$, ensuring that the univariate R^2 value on y_1 is 0%, being 11% on the five remaining y-variables; yet the maximal linear combination (yielding $R^2 = 26\%$) is:

$$0.75 \times y_1 - 0.3(y_2 + \ldots + y_6),$$

emphasizing y_1 more than any other y-variable. In example (b), $\beta = 1$, ensuring that all the y-variables are influenced by x via their shared common factor; yet the maximal linear combination consists of y_1 alone. Finally, in example (c), $\beta = 0$, ensuring the only effect of x is on y_1 alone, yet all the other y-variables (with R^2s of 0%) nonetheless play a role in the maximal linear combination.

These examples show that if even a reasonable, simple structure exists among the y variables, cases can exist where the univariate F effects and the maximal linear combination are far from easy to interpret. This issue, together with questioning of the universal benefit of overall Type I Error control, indicate that MANOVA need not be an automatic choice for analysis of multivariate y data. In

TABLE 6
Some MANOVA Examples

	y Variables						
	1	*2*	*3*	*4*	*5*	*6*	*Max*
Example (a): $\beta = -1$							
Linear combination[a]	0.75	–0.30	–0.30	–0.30	–0.30	–0.30	
$R^{2,b}$	0%	11%	11%	11%	11%	11%	26%[c]
Example (b): $\beta = 1$							
Linear combination	1	0	0	0	0	0	
R^2	33%	11%	11%	11%	11%	11%	33%
Example (c): $\beta = 0$							
Linear combination	0.94	–0.16	–0.16	–0.16	–0.16	–0.16	
R^2	11%	0%	0%	0%	0%	0%	18%

Note. $p = 6$, dependent variables, two-group design. MANOVA = multivariate analysis of variance.
[a]The linear combination of $y_1 - y_6$ that yields the maximum F value in the two-group design (i.e., the MANOVA maximal discriminant). [b]The R^2 value when the corresponding y_j value is analyzed. [c]The R^2 value when the maximal linear combination is analyzed (also the maximum R^2 for any linear combination).

Example C, particularly, direct interpretation of (multiple) univariate F ratios might seem better. The Appendix describes in detail the mathematical theory behind these observations.

"VARIABLE IMPORTANCE" IN MLR: VENN DIAGRAMS AND THE FLAWED CONCEPT OF "SHARED VARIANCE"

When the x predictors of y in an MLR are uncorrelated, many textbooks discuss evaluating the "importance" of each predictor by apportioning y's total variability among the different x predictors. In what follows, we are concerned with population parameters only; we use unstandardized partial regression coefficients only ("b" is used to denote such a coefficient when only one x predictor is used, and "β" is used when more x predictors than one are in the MLR equation); and suppose here we have just two predictors, with MLR thus given by: $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + r$ (although the points to be made next generalize to situations with more than two predictors). Let:

$$Var(y) = \sigma_y^2, Var(r) = \sigma_r^2, \text{ and } Var(x_1) = \sigma_1^2, Var(x_2) = \sigma_2^2.$$

If the two x predictors are uncorrelated with each other (and, by definition, with r), then:

$$\sigma_y^2 = \beta_1^2\sigma_1^2 + \beta_2^2\sigma_2^2 + \sigma_r^2,$$

and the y variance can meaningfully be partitioned into components related separately to each x predictor and to r. Because x_1, x_2, r are all uncorrelated, were we to fit just x_1 to y, we would obtain:

$$y = a + b_1 x_1 + r^*, \text{ where } b_1 = \beta_1 \text{ and } r^* = \beta_2 x_2 + r;$$

and so the portion of y variance associated with x_1 *alone* would also be $\beta_1^2\sigma_1^2$ (and ditto for x_2 alone). In terms of *explained* (by an MLR model) y variance, we would consider just those parts of y predicted by the given MLR, and when x_1 and x_2 are uncorrelated:

$$\tilde{y}_1 = \alpha + \beta_1 x_1 + \beta_2 x_2, \text{ and}$$
$$\hat{y}_1 = \alpha_1 + b_1 x_1 = \alpha_1 + \beta_1 x_1, \hat{y}_2 = \alpha_2 + b_2 x_2 = \alpha_2 + \beta_2 x_2;$$

and:

$$V_{12} = Var(\tilde{y}) = \beta_1^2\sigma_1^2 + \beta_2^2\sigma_2^2, \text{ and}$$
$$V_1 = Var(\hat{y}_1) = \beta_1^2\sigma_1^2, V_2 + Var(\hat{y}_2) = \beta_2^2\sigma_2^2;$$
$$\text{that is } V_{12} = V_1 + V_2.$$

For such reasons, when the x predictors are uncorrelated, frequently it is claimed that the relative sizes of these V_1 and V_2 portions can be used as an index of the "importance" to y of each x predictor. This would seem to be fairly sensible in purely *predictive* applications—where primarily we are not addressing a empirical or scientific quantitative relationship—for example, using historical data simply to predict next year's wheat yield in Australia from any useful x predictors.

However, if we are addressing an empirical or scientific quantitative relationship among the attributes (of participants) measured by y and the xs, then the concept of "importance" seems rather peculiar. In this case, presumably we are addressing a scientific issue not too unlike that of physicists when they investigate the relationship, say, among the attributes (of objects-in-motion) momentum (M), mass (m) and velocity (v): $M = mv$. Which variable here is more important—mass or velocity? The question does not make much sense, and if our investigation happens to be on a particular local sample of objects-in-motion, where mass and velocity just happen to be uncorrelated, then their variances are only local to this particular study; that is, different such studies apparently would conclude *importance* differently.

Whether one accepts the above analogy with physics, and whether one wishes to use such y variance contributions as measures of importance, the situation becomes more complicated when the predictors x_1 and x_2 are correlated (ρ_{12}), and here in general:

$$b_1 = \beta_1 + \beta_2\left(\frac{\sigma_{12}}{\sigma_1^2}\right) \neq \beta_1, \text{ where } \sigma_{12} = Cov(x_1, x_2) = \rho_{12}\sigma_1\sigma_2. \qquad (1)$$

(Recall that b_1 is the unstandardized regression coefficient obtained when fitting an MLR to y with just x_1 as the only predictor. Similar formula and comment apply for b_2.) In this case, the explained y-variances from an MLR with both predictors will have form:

$$V_{12} = \beta_1^2\sigma_1^2 + 2\beta_1\beta_2\sigma_{12} + \beta_2^2\sigma_2^2, \tag{2}$$

and, from an MLR equation with single predictor x_1, the variance accounted for by x_1 will be:

$$\begin{aligned} V_1 &= b_1^2\sigma_1^2 \\ &= \left[\beta_1 + \beta_2\left(\frac{\sigma_{12}}{\sigma_1^2}\right)\right]^2 \sigma_1^2, \text{ from (1)} \\ &= \beta_1^2\sigma_1^2 + 2\beta_1\beta_2\sigma_{12} + \beta_1^2\left(\frac{(\sigma_{12})^2}{\sigma_1^2}\right). \end{aligned} \tag{3}$$

(A similar expression with subscripts "1" and "2" swapped holds for V_2, the explained variance when only x_2 is used as a predictor in an MLR equation.)

Using Equations 2 and 3, we can exhibit the familiar observation—that adding a second predictor (say, x_2) to an MLR model always increases, or at least cannot decrease, the explained variance:

$$V_{12} - V_1 = \beta_1^2\sigma_1^2 - \beta_2^2\left(\frac{(\sigma_{12})^2}{\sigma_1^2}\right) = \beta_2^2\sigma_2^2(1-\rho_{12}^2) \geq 0.$$

Moreover, again using Equations 2 and 3, we see that, unlike the case where the x-predictors are uncorrelated, the explained variance from both predictors (V_{12}) does not equal the sum of explained variances from each alone ($V_1 + V_2$):

$$\begin{aligned} \Delta = (V_1 + V_2) - V_{12} &= 2\beta_1\beta_2\sigma_{12} + \beta_1^2\left(\frac{(\sigma_{12})^2}{\sigma_2^2}\right) + \beta_2^2\left(\frac{(\sigma_{12})^2}{\sigma_1^2}\right) \\ &= 2\beta_1\beta_2\sigma_{12} + \beta_1^2\sigma_1^2\left(\frac{(\sigma_{12})^2}{\sigma_1^2\sigma_2^2}\right) + \beta_2^2\sigma_2^2\left(\frac{(\sigma_{12})^2}{\sigma_2^2\sigma_1^2}\right) \\ &= 2\beta_1\beta_2\sigma_{12} + \rho_{12}^2(\beta_1^2\sigma_1^2 + \beta_2^2\sigma_2^2). \end{aligned} \tag{4}$$

Note that whenever the term $2\beta_1\beta_2\sigma_{12}$ is negative and of magnitude greater than or equal to the (always positive) term $\rho_{12}^2(\beta_1^2\sigma_1^2 + \beta_2^2\sigma_2^2)$, then Δ will be negative or zero.

There are some textbooks (e.g., Cohen et al., 2003; Howell, 2002; Tabachnick & Fidell, 2001) that discuss attempting to evaluate "variable importance" in this case of correlated predictors by perusing the proportions of the explained y variance from our full model (i.e., the proportions of V_{12}) "attributable" to different sources—explained y variance due to:

$$x_1\text{-alone: } P_1 = \frac{V_1}{V_{12}},$$

$$x_2\text{-alone: } P_2 = \frac{V_2}{V_{12}},$$

$$\text{"shared variance": } P_\Delta = \frac{\Delta}{V_{12}}.$$

By virtue of Equation 4—and noting that P_1 and P_2 must always lie between 0 and 1, as $V_{12} \geq V_1$, $V_2 \geq 0$—we have:

$$1 = P_1 + P_2 - P_\Delta,$$

and this perspective *on first glance* appears quite appealing—it is *as if* P_Δ is counted twice, once in P_1 and once in P_2, and thus we must subtract it, just once, to apparently reconstitute the complete explained variance from the full model. This perspective is reinforced by the common presentation of such concepts in Venn diagrams, such as Figure 5.

However, in my opinion, this perspective or approach is confusing and is seriously flawed, making it a *particularly poor* teaching aid. The reason for this opinion is that the quantity Δ easily can be *negative*, and thus the quantity P_Δ also would be negative, and so in *no way* could correspond to any portion of component or explained y variance, shared or otherwise; nor could it be represented on a Venn diagram, where area is meant to reflect (portions of) variance, and both "area" and

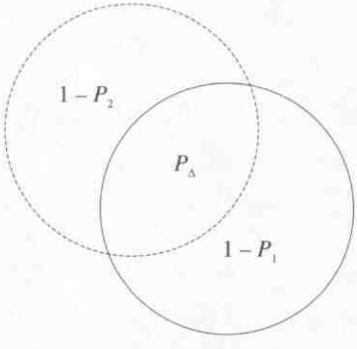

FIGURE 5 Venn diagrams of "shared variance." The total area, the union of both circles, represents 100% of y variance explained by full MLR (multiple linear regression) model, with both predictors. Area in dashed circle is P_1 $\{= (1 - P_2) + P_\Delta\}$; area in solid circle is P_2 $\{= \{1 - P_1) + P_\Delta\}$.

"variance" are positive concepts. For this reason, even when P_Δ is positive, it cannot be shared variance in any sense whatever, yet such a perspective invites us to use this misleading interpretation. Some textbooks that use these Venn diagrams, unfortunately do not draw their readers' attention to this limitation (e.g., Howell, 2002; Tabachnick & Fidell, 2001); although others do (Cohen et al., 2003). In attempting to address this phenomenon, the concept of a "suppressor" variable has been invoked, often without satisfactory distinction between predictive and explanatory applications. I agree with Pedhazur (1997):

> By and large, conceptions of suppressor variables were formulated from the perspective of prediction, rather than explanation. Accordingly, most, if not all, discussions of suppressor effects appeared in the psychometric literature in the context of validation .... It is noteworthy that the notion of suppression is hardly alluded to in the literature of some disciplines (e.g., sociology, political science). ... it [the notion of suppression] also increased the potential for ignoring the paramount role of theory in interpreting results of multiple regression analysis. (pp. 186–188)

Let us make these concepts more concrete with a simple empirical or scientific or explanatory or theory-based example, rather than a purely predictive example. Suppose a researcher is interested in investigating the possibility of "gender bias" (against females) in the corporate workplace. They decide to see if, for employees of equal educational level, the female employees receive less income than the males. They gather employees from the corporations of interest, noting their income (y, scored in \$1000s per annum), gender (x_1, scored males = 1, females = 0) and their educational status (x_2, scored high = 1, low = 0). Equal numbers of each gender, and of each educational status happen to be gathered, so the variances of both predictors happen to be 0.25. The most straightforward way of investigating this issue would be to inspect the mean incomes for each gender, at each level of educational status, using perhaps a 2×2 ANOVA. However, if we assume (for the purpose of this example) that no interaction exists on income, we could fit an MLR model. The parameters in this MLR model:

$$\text{INCOME} = \alpha + \beta_1\text{GENDER} + \beta_2\text{ES} + r,$$

would have the following informative interpretations:

- β_1: among employees of the same educational status (high or low), β_1 is the additional annual Income (in \$1000s per annum) associated with being a male;
- β_2: among employees of the same gender, β_2 is the additional annual Income associated with being more highly educated.

Figure 6 describes this scenario. Our researcher would be interested in whether β_1 is zero (no gender bias) or positive (gender bias), and if so, the extent of such bias (the

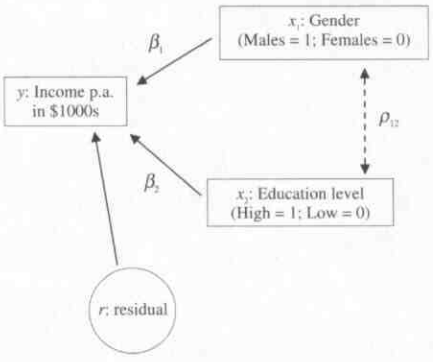

Figure 6. A (simple) theoretical framework for investigating gender bias in the corporate workplace.

size of such a positive β_1). The extent to which gender and educational status intercorrelate, their variances, and which of them contributes more to income variance—i.e., is of more *apparent* importance—actually are issues of no real interest (although it may be useful to compare β_1 with β_2, taking into account the known empirical scaling of our predictors x_1 and x_2).

Table 7 presents a series of scenarios that highlight the issues raised previously. In Scenarios A, B, and C, $\beta_1 = 2$ and $\beta_2 = 5$. That is, the same empirical or scientific relationship exists in all three scenarios, with educational status "controlled" (or held constant), the males in each population on average earn $2000 per annum more than the females. However, in each of these three populations, the correlation between the predictors varies.

In Scenario A, gender and educational status are uncorrelated, and if we compare the income levels of the two genders, not attending to Educational Status, we are comparing equally educated groups, and we should reach the same conclusion about gender bias. So we do, the b_1 coefficient obtained from fitting gender alone (i.e., not attending to educational status) equals that when gender is controlled. Finally, because $\rho_{12} = 0$, the apparent shared variance is $\Delta = 0$, $P_\Delta = 0$ and the Venn diagram of this scenario would be like Figure 5, except that the dashed and solid circles would not intersect.

In Scenario B, gender and educational status have a positive correlation of 0.5—there is a tendency for the males to be better educated; thus, when we compare males to females, with educational status ignored (by fitting $y = a + b_1x_1 + r^*$), males average $4,500 per annum more in salary than females ($b_1 = 4.5$), not just the $2000 that is due to their gender per se ($\beta_1 = 2$), but a further $2500 related to the fact that they also are better educated (in accordance with Equation 1). This situation corresponds exactly to Figure 5, where the overlap in circles represents P_Δ = 44%, and is the sort of scenario that probably motivated this poor way of attempting to represent shared variance.

TABLE 7
Some MLR Models Illustrating the Flawed Concept of Shared Variance

	Scenario						
	A	*B*	*C*	*D*	*E*	*F*	*G*
Model with both predictors: INCOME = $\alpha + \beta_1$GENDER + β_2ES + r							
β_1	2	2	2	–2	–2	–1	0
β_2	5	5	5	5	5	8	5
Predictors' correlation							
β_{12}	0	0.5	–0.5	0.5	–0.5	0.25	0.5
Models with a single predictor: INCOME = $a + b_1$GENDER + r*							
b_1	2	4.5	–0.5	0.5	–4.5	–1	2.5
INCOME = $a + b_2$ES + r*							
b_2	5	6	4	4	6	7.75	5
Proportions of y variance explained from full model							
P_1	14%	52%	1%	1%	52%	2%	25%
P_2	86%	92%	84%	84%	92%	98%	100%
P_Δ	0%	44%	–14%	–14%	44%	0%	25%

Note. MLR = multiple linear regression.

In Scenario C, we have a population where females happen to be better educated ($\rho_{12} = -0.5$), and thus when gender alone is interpreted they actually earn \$500 more than the males ($b_1 = -0.5$); but this is because they are better educated, not because they are female. Here $P_\Delta = -14\%$, and no analogue of Figure 5 exists.

In Scenarios D and E the effect of gender, with educational status controlled, is reversed ($\beta_2 = -2$): the corporations actually discriminate *against* males (of equal education). Scenario D presents what is typically called a suppressor effect, with the effect of educational status being suppressed by gender: that is, when educational status alone is fit, $b_2 = 4$, but because ρ_{12} is positive (the better educated tending to be male) and β_1 is negative (males suffer discrimination), the true effect of educational status with gender controlled is suppressed (from $\beta_2 = 5$ to $b_2 = 4$). In such cases, P_Δ can be negative, as in this example (although it does not always have to be). Had we scored gender in Scenario D opposite to that in Scenario C (i.e., had we scored male = 0, female = 1), then the two scenarios describe identical empirical or scientific situations: gender bias against females in the corporate workplace; yet Scenario D exhibits what is called a suppressor effect, whereas Scenario C does not, simply because our scoring convention in Scenario D led to a negative β_1 and positive ρ_{12}. Scenario E, likewise, is equivalent to Scenario B with Gender reverse scored.

Scenario F is interesting in that $V_{12} = V_1 + V_2$ exactly, $P_\Delta = 0$, just as in Scenario A. In addition, a Venn diagram with nonintersecting circles would have to be used in Scenario F, despite the fact $\rho 12 = 0.25 \neq 0$. Finally, Scenario G shows a situation

where there is no gender bias ($\beta_1 = 0$), yet $P_1 = 25\%$, indicating that gender has some apparent importance.

The βs (the *un*standardized partial regression coefficients, in *y* units per *x* unit) are the quantities of main empirical or scientific interest in MLR; and apportioning variance using various derived, variance-based indexes (some reasonable, some flawed) does not seem of great relevance in an empirical or scientific context. This, at least, is one opinion.

DISCUSSION

So none of the previous common techniques is "fool-proof," and some common beliefs are far from universal. The same is true of *every* quantitative method we use. Their proper use demands thoughtful justification in each application. Of course, the same is also true of every other aspect of a piece of psychological research: equipment, subject selection, treatment administration, etc. It seems to be unfortunately common with quantitative methods that social scientists often fail to confront genuine statistical uncertainties, seeking and citing "golden rules," which can never be universal and which must be inappropriate in some contexts.

Indeed, in my opinion, we find ourselves historically in an era where unthinking golden rules proliferate. The recent contretemps about significance testing is an example of the confusion such a poor research milieu can create (for fuller discussion of these issues, see Grayson, 1998, Nickerson, 2000, and other references therein, particularly Gigerenzer et al., 1989; Oakes, 1986). Part of the reason for this, one suspects, relates to the parlous state of quantitative measurement in the social sciences. Failure to confront such issues, by hiding them behind a mathematical–technical veneer, is *not* the way forward. One classic example of such unfortunate behavior is the longstanding and widespread belief or advice that standardizing variables, and dealing with correlations, is somehow a better way scientifically to proceed because it is scale free. Ignoring one's units of measurement (poor though they may be) is a worrying way for any science to proceed (see Michell, 2000). In addition to the advice at the end of the last section, to attend to the unstandardized βs as the main empirical or scientific parameters of interest, even though the measurement of our *y* and *x* constructs is poor, we might simultaneously proceed by attempting to better measure our constructs, rather than to turn away from the whole issue.

However again, as a final plea, do not accept my perspective on these matters—make up your own mind, but do think about such issues. They are never as clear-cut as usually they are presented, or as one might wish and hope for them to be!

REFERENCES

Abramowitz, M., & Stegun, I. (1965). *Handbook of mathematical functions* (National Bureau of Standards, Applied Mathematics Series 55). Washington, DC: U.S. Department of Commerce and National Bureau of Standards.

Bock, R. D. (1985) *Multivariate statistical methods in behavioral research*. Lincolnwood, IL: Scientific Software.

Bollen, K. A. (1989) *Structural equations with latent variables*. New York: Wiley.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 12,* 997–1003.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Krüger, L. (1989). *The empire of chance*. Cambridge, England: Cambridge University Press.

Graybill, F. A. (1969). *Introduction to matrices with applications in statistics*. Belmont, CA: Wadsworth.

Grayson, D. A. (1998). The frequentist facade and the flight from evidential inference. *British Journal of Psychology, 89,* 325–345.

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

Howell, D. C. (2002). *Statistical methods for psychology* (5th ed.). Pacific Grove, CA: Duxbury.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1995) *Continuous univariate distributions* (Vol. 2). New York: Wiley.

Macdonald, R. R. (2002). The incompleteness of probability models and the resultant implications for theories of statistical inference. *Understanding Statistics, 1,* 167–189.

Michell, J. (2000) Normal science, pathological science and psychometrics. *Theory and Psychology, 10,* 639–667.

Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods, 5,* 241–301.

Oakes, M. (1986) *Statistical inference: A commentary for the social and behavioural sciences*. Chichester, England: Wiley.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research* (3rd ed.). Fort Worth, TX: Harcourt Brace.

Radloff, L. S. (1977). The CES–D scale: A self report depression scale for research in the general population. *Applied Psychological Measurement, 1,* 385–401.

Tabachnick, B. G., & Fidell, L.S. (2001). *Using multivariate statistics* (4th ed.). New York: Harper and Row.

Thompson, W. D., & Walter, S. D. (1988). A reappraisal of the kappa coefficient. *Journal of Clinical Epidemiology, 41,* 949–958.

Waite, L. M., Broe, G. A., Creasey, H., Grayson, D. A., Edelbrock, D., Cullen, J., et al. (2001). Clinical diagnoses and disability among community dwellers aged 75 or over. *Australasian Journal of Ageing, 20,* 67–72.

Wilkinson, L., & The APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals. *American Psychologist, 54,* 594–604.

APPENDIX

Power

Power was computed using the approximation to the noncentral t CDF (see Abramowitz & Stegun, 1965, p. 949):

$$\begin{aligned}
&\text{CDF}(t_v') = \Psi(z), \text{ with } \Psi \text{ being the CDF of a SNRV, } z \sim N(0,1), \text{ and} \\
&z = \frac{t_v'\left(1-\frac{1}{4v}\right)-\delta}{\sqrt{1+\frac{t_v'^2}{2v}}}; \text{ where} \\
&v = \text{degrees of freedom } (n-1), \\
&t_v' = \text{noncentral } t \text{ value, and} \\
&\delta = \frac{\sqrt{n}\mu}{\sigma}, \text{ the noncentrality parameter.}
\end{aligned} \tag{A1}$$

For the chosen alternative underlying the results in Table 4 ($\mu = 1$), the two-tailed power based on α was computed by using the Equation A1 twice to compute, with "$t_v(\alpha)$" denoting the central t distribution critical value at v degrees of freedom that isolates α in the upper, right tail:

$$\begin{aligned}
\text{Power}_{\text{right tail}} &= 1 - \text{CDF}\left(t_v\left(\frac{\alpha}{2}\right)\right), \\
\text{Power}_{\text{left tail}} &= \text{CDF}\left(-t_v\left(\frac{\alpha}{2}\right)\right); \text{ yielding} \\
\text{Power} &= \text{Power}_{\text{right tail}} + \text{Power}_{\text{left tail}}.
\end{aligned}$$

Family-Wise Control

In this section, the key approximation is that for the noncentral F statistic by a SNRV, z (see Abramowitz & Stegun, 1965, p. 948; Johnson, Kotz, & Balakrishnan, 1995, p. 492):

$$\begin{aligned}
&\text{CDF}(F') = \Psi(z), \text{ with} \\
&z = \frac{\left[\frac{v_1 F'}{(v_1+\lambda)}\right]^{1/3}\left[1-\frac{2}{9v_2}\right]-\left[1-\frac{2(v_1+2\lambda)}{9(v_1+\lambda)^2}\right]}{\sqrt{\frac{2(v_1+2\lambda)}{9(v_1+\lambda)^2}+\frac{2}{9v_2}\left[\frac{v_1 F'}{(v_1+\lambda)}\right]^{2/3}}}; \text{ where} \\
&v_1, v_2 = \text{degrees of freedom, top and bottom} \\
&F' = \text{noncentral } F \text{ value, and} \\
&\lambda = \frac{\sum_j n_j(\mu_j-\mu)^2}{\sigma^2}, \text{ the noncentrality parameter.}
\end{aligned} \tag{A2}$$

To ascertain omnibus F-test power for the data in Table 5, we set $v_1 = 6$, $v_2 = 133$, calculate λ for a particular configuration of the μs and σ^2, and using (A2), obtain, for α:

$$\text{Power} = 1 - \text{CDF}(F_{v_1,v_2}(\alpha)),$$

where $F_{v_1,v_2}(\alpha)$ is the central F critical value.

For the (two-tailed) individual contrast tests, the observed F-value is always distributed as an F_{1,v_2} statistic, central when the contrast null is true, and noncentral when false, with:

$$\lambda = \frac{20[(5-3)^2 + (1-3)^2]}{\sigma^2} = \frac{160}{\sigma^2}, \text{ for the different effect sizes.}$$

For the 10 true contrasts, the α-rate was ascertained by taking F_{Crit} (whatever the strategy: decisional, Scheffé, or Tukey HSD) and finding the corresponding central F_{1,ν_2} tail area. For the 11 individual false contrast decisions (whatever the critical value F_{Crit}, decisional, Scheffé or Tukey HSD), Equation A2 was used—with $\nu_1 = 1$, $\nu_2 = 133$ and λ as immediately above—to ascertain the common β rate for the 11 contrasts where the nulls were false:

$$\beta = \text{CDF}(F_{Crit}),$$

and the expected numbers of Type I, Type II and Total decisional errors in the 21 contrast decisions were then obtained as: Type I – 10α, Type II – 11β, Total – $10\alpha + 11\beta$.

MANOVA

We suppose that all random variables are mean-corrected. The following issue is discussed at the level of population parameters, consistently estimated by random-sample based corresponding quantities. Let y denote a random vector of dependent variables. Let x denote the two-group (equal numbers of subjects) design variable:

$$\begin{aligned} x &= -0.5, \text{ control} \\ &= 0.5, \text{ treatment; with} \\ \sigma_x^2 &= 0.25. \end{aligned}$$

Our model in the text addressing the effects of x on y assumes:

$$\begin{aligned} &y_{p\times 1} = \gamma_{p\times 1}x + \lambda_{p\times 1}\theta + u_{p\times 1}; E(r^2) = \sigma_r^2, E(uu') = D = diag(d_i), E(ux) = 0_{p\times 1}. \\ &\theta = \beta x + r, E(r^2) = \sigma_r^2; E(rx) = 0. \end{aligned}$$

Thus:

$$y = (\gamma + \beta\gamma)x + (\gamma r + u), \text{ where } E((\gamma r + u)x) = 0_{p\times 1}.$$

We can partition the total y-variability into components associated with the "hypothesis" and with "error" (residual or "within cell", i.e., common to fixed x-profiles):

$$\begin{aligned} E(yy') &= (\gamma + \beta\lambda)E(x^2)(\gamma + \beta\lambda)' + E(\lambda r + u)(\lambda r + u)' \\ &= [(\gamma + \beta\lambda)\sigma_x^2(\gamma + \beta\lambda)'] + [\lambda\sigma_r^2\lambda' + D] \\ &= H + E, \text{ say.} \end{aligned} \tag{A3}$$

[Note that the matrix H is rank 1.]

Any linear combination c of the y-variates will have variance:

$$\begin{aligned} &E(c'y)^2 = c'Hc + c'Ec, \text{ with} \\ &\frac{R^2_{c'y}}{1 - R^2_{c'y}} = \frac{c'Hc}{c'Ec} = \tau_{c'y}; \text{ and, in particular:} \\ &\frac{R^2_{y_i}}{1 - R^2_{y_i}} = \frac{h_{ii}}{e_{ii}}, \text{ or } R^2_{y_i} = \frac{h_{ii}}{h_{ii} + e_{ii}}. \end{aligned} \tag{A4}$$

We seek the linear combination c_{Max} (subject to some scaling constraint like $c'_{max}c_{max} = 1$) such that:

$$\frac{c'_{max}Hc_{max}}{c'_{max}Ec_{max}} = \tau_{max} = \tau_{c_{max}y},$$

is maximized, as thus is the corresponding univariate F ratio.

It transpires (see e.g., Bock, 1985) that this maximization problem reduces to finding the largest eigenvalue of τ_{Max} of $|H - \tau E| = 0$, or of:

$$|E^{-1}H - \tau I| = 0.$$

The corresponding eigenvector will give us the maximal discriminant c_{Max}.

The matrix $E^{-1}H$ is not (generally) symmetric, and in the discussion below we will need certain lemmas dealing with eigenvalues and determinants of nonsymmetric matrixes.

Lemma 1. The inverse of the matrix:

$$C = D_{p\times p} + ka_{p\times 1}a'_{1\times p};\ \text{scalar } k \neq -(a'D^{-1}a)^{-1};D = diag(d_i),\ \text{full rank},$$

is:

$$C^{-1} = D^{-1} - \left(\frac{k}{(1+ka'D^{-1}a)}\right)D^{-1}aa'D^{-1}.$$

Proof: See for example, Graybill (1969), Theorem 8.3.3. ❑

Lemma 2. The determinant of the matrix:

$$C = D_{p\times p} + a_{p\times 1}b'_{1\times p};D = diag(d_i);$$
$$\text{where } a = \begin{pmatrix} a_1 \\ \vdots \\ a_p \end{pmatrix}, b = \begin{pmatrix} b_1 \\ \vdots \\ b_p \end{pmatrix},$$

is, in general:

$$|C| = \prod_{i=1}^{p} d_i + \sum_{i=1}^{p}\left(a_i b_i \prod_{\substack{j=1 \\ j\neq i}}^{p} d_j\right),$$

and is, when C,D are full rank:

$$|C| = \prod_{i=1}^{p} d_i(1 + a'D^{-1}b).$$

Proof: See Graybill (1969), Theorem 8.4.3. ❑

Lemma 3. For any rank 1 $p \times p$ matrix ab', such that $a'b \neq 0$, the only nonzero eigenvalue is $a'b$, with corresponding eigenvector ka, any scalar $k \neq 0$.

Proof: Consider the nonzero roots of:

$$\begin{aligned} 0 &= |ab' - \tau I| \\ &= (-\tau)^p \left|I - \left(\frac{1}{\tau}\right)ab'\right| \\ &= (-\tau)^p \left(1 - \left(\frac{1}{\tau}\right)a'I^{-1}b\right)\prod 1,\ \text{from Lemma 2,} \\ &= -(-\tau)^{p-1}(\tau - a'b), \end{aligned}$$

where the only nonzero root must be $\tau = a'b$. To find the corresponding eigenvector, we need to find c such that:

$$[ab' - \tau I]c = [ab' - (a'b)I]c = 0, \text{ whence } c = ka, \text{ any scalar } k \neq 0. \qquad \square$$

We can now use Lemma 3 to obtain expressions for τ_{max} and c_{max} in closed form.

Theorem. If, as above in (A3):

$$H = (\gamma + \beta\lambda)\sigma_x^2(\gamma + \beta\lambda)' \text{ and } E = \lambda\sigma_F^2\lambda' + D,$$

where:

$$1_n \ -\sigma_F^2\lambda' D^{-1}\lambda,$$

then the maximal (and only nonzero) eigenvalue of $E^{-1}H$ is:

$$\tau_{max} = \sigma_x^2(\gamma + \beta\lambda)'(\lambda\sigma_F^2\lambda' + D)^{-1}(\gamma + \beta\lambda), \tag{A5}$$

with corresponding eigenvector:

$$c_{max} = \quad (\lambda\sigma_F^2\lambda' + D)^{-1}(\gamma + \beta\lambda). \tag{A6}$$

Proof: Under these conditions, we can apply Lemma 1, identifying C with:

$$D + \sigma_F^2\lambda\lambda',$$

and Lemma 3, identifying:

$$a \text{ with } (\lambda\sigma_F^2\lambda' + D)^{-1}(\gamma + \beta\lambda) \text{ and } b \text{ with } \sigma_x^2(\gamma + \beta\lambda),$$

and the result follows. $\square$

Using (A4), we have the maximal $-R^2$:

$$R_{max}^2 = \frac{\tau_{max}}{1 + \tau_{max}}. \tag{A7}$$

The examples in the text and Table 6 use Equations A4 through A7 with the following fixed values:

$$p = 6; D = I_6; \lambda = 1_{6\times 1}; \gamma = \begin{pmatrix} 1 \\ 0_{5\times 1} \end{pmatrix}; \sigma_F^2 = 1; \sigma_x^2 = 0.25;$$

and with various values of β.