Statistical power for the two-factor repeated measures ANOVA

PATRICK J. POTVIN and ROBERT W. SCHUTZ
University of British Columbia, Vancouver, British Columbia, Canada

Determining a priori power for univariate repeated measures (RM) ANOVA designs with two or more within-subjects factors that have different correlational patterns between the factors is currently difficult due to the unavailability of accurate methods to estimate the error variances used in power calculations. The main objective of this study was to determine the effect of the correlation between the levels in one RM factor on the power of the other RM factor. Monte Carlo simulation procedures were used to estimate power for the A, B, and AB tests of a 2×3, a 2×6, a 2×9, a 3×3, a 3×6, and a 3×9 design under varying experimental conditions of effect size (small, medium, and large), average correlation (.4 and .8), alpha (.01 and .05), and sample size ($n = 5, 10, 15, 20, 25$, and 30). Results indicated that the greater the magnitude of the differences between the average correlation among the levels of Factor A and the average correlation in the AB matrix, the lower the power for Factor B (and vice versa). Equations for estimating the error variance of each test of the two-way model were constructed by examining power and mean square error trends across different correlation matrices. Support for the accuracy of these formulae is given, thus allowing for direct analytic power calculations in future studies.

Determining a study's power a priori is an important part in the planning and formulation of experimental research. It allows researchers the opportunity to evaluate whether their experimental design has a sufficient probability of showing a statistically significant effect, thus warranting the time, money, and effort necessary to conduct the research. In light of its benefits in experimental planning, many analytical methods and computer programs have become available for determining the power of a number of statistical tests (e.g., Borenstein & Cohen, 1988; Bradley, 1988; Cohen, 1988; Erdfelder, Faul, & Buchner, 1996; Friedman, 1982; Gorman, Primavera, & Allison, 1995; Kraemer & Thiemann, 1987; Lipsey, 1990). There are some statistical tests, however, for which power analysis procedures have been less extensively described. This is particularly the case for complex repeated measures analysis of variance (RM ANOVA) designs. Many of the programs noted above do not include options for computing power for the *F* test in an RM ANOVA, and those that do, do so only for designs with a single RM factor (or for two or more RM factors under very restrictive conditions of equal correlational patterns for all factors).

In an RM ANOVA the power of the *F* test of the RM factor (and all its interactions) is a function of the magnitude of the correlations between the levels of the RM factor. These correlation coefficients can range from near-zero to near-one, depending on the nature of the

study and the reliability of the dependent variable. Consequently, the power of the test can vary considerably from study to study, even with constant sample sizes, magnitude of effect, and within-cell variances. Within a single study, if there are two RM factors, it could be possible that the power to detect a difference of say *d* units could be high for one RM factor and very low for the other RM factor. Obviously, researchers would like to be able to estimate, a priori, the power for each main effect and interaction in any RM design. Unfortunately, this is not currently possible, and most textbooks and journal articles do not even acknowledge the difficulty with respect to this aspect of power and sample size determination. For example, a recent and very comprehensive paper (Levin, 1997) on power in ANOVA designs includes split-plot, RM, and ANCOVA designs, but never with more than one RM factor. Given that designs with two or more factors are quite common, especially in studies on learning, the absence of appropriate methods to compute power in these situations is a serious deficiency in a researcher's tool chest of statistical methods.

Although a considerable amount of research has been conducted on power analysis of RM ANOVA, much of this work has focused on comparing power values between univariate and multivariate RM tests under varying conditions of nonsphericity (Davidson, 1972; Grima & Weinberg, 1987; Mendoza, Toothaker, & Nicewander, 1974). More importantly, most power and/or sample size methods described have been restricted to those designs involving a single within-subjects variable, particularly the one-way RM designs (Davidson, 1972; Marcucci, 1986; Robey & Barcikowski, 1984; Vonesh & Schork, 1986) and two-way mixed or "group by trials" models

(Bradley & Russell, 1998; Marcucci, 1986; Muller & Barton, 1989; Muller, LaVange, Ramey, & Ramey, 1992; Mulvenon & Betz, 1993; Rochon, 1991). It is acknowledged that some methodologists are aware of the need to compute power in designs with two or more RM factors and that they have included such design options in their computer programs. For example, Bradley's (1988) very useful program DATASIM can now "accommodate designs having any mixture of up to three independent and/or correlated-groups factors" (Bradley, Russell, & Reeve, 1996). However, researchers can make use of such options to compute an a priori power only for designs with a single RM factor, or, for two or more RM factors, they must be willing to assume that the magnitudes of the correlations between factor levels are the same for each factor. To use programs such as DATASIM to compute power requires the computation of the expected error variance for each of the RM effects—but analytical methods to compute these error variances exist only for single-factor RM designs (unless one is willing to assume that the error variances are the same for each RM factor—a most tenuous assumption).

To determine power for a univariate RM ANOVA, a noncentrality parameter must be computed on the basis of the experimental conditions of the design. The noncentrality parameter, λ, a function of effect size and sample size, is a measure representing the factor by which the F ratio departs from the central F distribution when a difference between treatment means actually exists (Winer, Brown, & Michels, 1991). Although three other statistics are also used as noncentrality or effect size measures—δ^2 (Davidson, 1972; Robey & Barcikowski, 1984), ϕ (Pearson & Hartley, 1951; Tang, 1938; Winer et al., 1991), and f (Cohen, 1988; Kirk, 1995)—all are closely related to λ, and like λ, essentially represent the magnitude of difference between population means in relation to the population error variance of the dependent variable involved. For within-subjects RM ANOVA designs with q levels on the one and only RM factor (which we label a "trials" effect here), the noncentrality parameters can be computed as follows:
(a) for the trials effect in a one-way RM design, $j = 1, \ldots, q$;

$$\lambda = \frac{n \sum (\mu_j - \mu)^2}{\sigma_e^2}; \quad (1)$$

(b) for the group main effect in a two-way $(A_p \times B_q)$ mixed design, $i = 1, \ldots, p$;

$$\lambda_A = \frac{nq \sum (\mu_i - \mu)^2}{\sigma_e^2}; \quad (2)$$

(c) for the trials main effect in a two-way $(A_p \times B_q)$ mixed design, $j = 1, \ldots, q$;

$$\lambda_B = \frac{np \sum (\mu_j - \mu)^2}{\sigma_e^2}; \text{ and} \quad (3)$$

(d) for the groups $\times$ trials interaction in a two-way $(A_p \times B_q)$ mixed design, $i = 1, \ldots, p, j = 1, \ldots, q$;

$$\lambda_{AB} = \frac{n \sum \sum (\mu_{ij} - \mu_i - \mu_j + \mu)^2}{\sigma_e^2}, \quad (4)$$

where μ_{ij} is the cell mean, μ_i and μ_j are the marginal means for the levels of the randomized group (RG) and RM factors, respectively, μ is the grand mean, n is the sample size per group, p and q are the number of levels of the RG and RM factors, respectively, and σ_e^2 represents the error variance for the specific effect (Bradley, 1988; Winer et al., 1991).

As shown in Equations 1–4, the numerator of λ for each test can be computed simply by inputting p, q, the desired n, and an estimate of the means. Approximating the denominator of λ, the error variance for these tests, is often much more complex. The error variance (MS_{error}) is a function of σ^2, the common within-cell variance (i.e., the estimated average variance among scores of the dependent variable within a group), and $\bar{\rho}$, the estimated average of the $q(q-1)/2$ correlation coefficients among repeated trials (Winer et al., 1991). For one-way RM tests, and the trials main effect and interaction test of the two-way mixed model, the expected error variance is as follows:

$$E(MS_{\text{error}}) = \sigma_e^2 = \sigma^2 (1 - \bar{\rho}). \quad (5)$$

For the group main effect test of the same design, the relationship between σ^2 and $\bar{\rho}$ is somewhat different and involves a third variable, q, the number of trials of the RM factor. In this case, the error term is expressed as

$$E(MS_{\text{error}}) = \sigma_e^2 = \sigma^2 [1 + (q-1)\bar{\rho}]. \quad (6)$$

Once λ has been determined (or any of its equivalents), power can then be estimated for a specific RM ANOVA test by referring to appropriate power tables[1] (Kirk, 1995; Pearson & Hartley, 1951; Rotton & Schonemann, 1978; Tang, 1938) or inputting λ values directly into software programs that include algorithms for computing the cumulative distribution function of the noncentral F distribution (e.g., DATASIM; Bradley, 1988). For designs with one RM factor, an alternative method offered by some computer power programs (e.g., PASS; NCSS, 1991) is to input the conditions of the experimental design (p, q, n, means, σ^2, $\bar{\rho}$, etc.) directly and have the computer calculate values of λ and power automatically.

When two or more RM factors are present in a design, however, a problem exists in determining λ, and, more specifically, in computing its denominator term or error variance. In the case of the two-way RM ANOVA, three average correlation values are possible, one for each independent matrix of the design (within-factor A, within-factor B, and the AB submatrices), and it is not clear how these values affect the error variance and thus the power of a particular test. We expected that, similar to what happens to the power of the group factor in the two-factor mixed model design (see Equation 6), the greater the magnitude of the correlations among the levels of Fac-

tor A, the larger the error variance for Factor B (and thus the lower the power). Pilot work conducted by the researchers of this study suggested that the relationship between error variance and correlations for designs with two RM factors is more complicated than that expressed in the power formulae of designs with just one correlation matrix. That is, the average correlation coefficient of a given matrix in a multiple RM design does not adequately account for the change that occurs in the error variance of its respective test. It appears that few attempts have been made to resolve this issue computationally. Although Winer et al. (1991) and Dodd and Schultz (1973) have provided a post hoc procedure for determining power for these designs using omega-squared, ω^2, a measure of the magnitude of the experimental effect, such a method is not very practical for a priori power analysis since it requires the researcher to know ahead of time the mean square error of the test involved.

As a result of the problems that exist in computing an appropriate error term, a priori power estimation for RM designs with multiple within-subjects variables remains difficult or unavailable. Currently, investigators whose experimental designs involve two or more RM factors are faced with either collecting values of residual error from previous research, an often difficult and impractical alternative, or avoiding the power issue altogether. Since such designs are often encountered in the health and behavioral sciences, it is important that the power of these statistical tests under varying conditions be determined and their values be made available.

Since analytical solutions to this problem do not appear to exist, Monte Carlo (MC) procedures can be used to provide approximations of error variances and power for specific RM designs under varying conditions. In the past, MC simulation has proven useful for approximating power in the two-way mixed model (Grima & Weinberg, 1987; Mendoza et al., 1974; Muller & Barton, 1989), but this method has not been extended to those designs involving two or more RM factors. Thus the purposes of this study were to conduct MC simulations of ANOVA for the two-factor repeated measures design to (1) determine the effect of the magnitude of the correlations among the levels of one RM factor on the magnitude of the error variance, and therefore on the power, of the other RM factor; (2) generate a series of user-friendly power tables for RM ANOVA designs with two within-subjects variables under a variety of sample size, effect size, and correlation matrix conditions; and (3) attempt to develop equations to compute approximations of the error variance (and thus λ) for tests of the two-way RM design by examining mean square error (MS_{error}) values from simulation results under different correlation matrices and identifying the relationships involved. If we were successful with respect to the third purpose, it would be unnecessary to present the findings of the second. The availability of accurate estimations of the error variance for each of the three effects in the two-way RM ANOVA would permit

direct calculation of power values from the noncentral F distribution, and thus tables of approximate power values would not be necessary.

METHOD

Selecting RM Designs and Experimental Conditions

Empirical data collected from studies and dissertations from a variety of disciplines and involving RM ANOVA tests were used to establish an ecologically valid range of design conditions (i.e., effect sizes, average correlation values, and sample sizes) for which power would be generated in this study. On the basis of these data, a total of three different effect sizes—.2 (small), .5 (medium), and .8 (large)—two average correlation values, .4 and .8, and six different sample sizes—5, 10, 15, 20, 25, and 30—were chosen. In addition, several RM levels of the two-way (A × B) ANOVA design were selected on the basis of their frequent occurrence in the field. For the A factor, this included two and three trials, and for the B factor, three, six, and nine trials. Two levels of alpha were chosen (.01 and .05) to provide a range of power values over different Type I error rates for the A, B, and AB tests. Thus, the total number of experimental conditions for which power was estimated in this study was 2,592.

It must be noted that the "effect size" used here differs from the effect size f, defined by Cohen (1988). First, Cohen did not suggest an effect size for the two-factor repeated measures design, and second, since a purpose of this study was to provide easy-to-use power tables for the practitioner, we wished to utilize an effect size that would be easy to calculate. Consequently, the effect size for each main effect is calculated as

$$d^* = \frac{\mu_{max} - \mu_{min}}{\sigma}, \qquad (7)$$

where μ_{max} and μ_{min} are the expected largest and smallest marginal means within a factor and σ is the expected average within-cell standard deviation. This is similar to Cohen's d used in a t test, and referred to as "the range of the standardized means" by Cohen in the ANOVA context. For the interaction, we used

$$d^* = \frac{\left|(\mu_1 - \mu_2)_{max} - (\mu_1 - \mu_2)_{min}\right|}{\sigma}, \qquad (8)$$

where $(\mu_1 - \mu_2)_{max}$ represents the maximum difference between cell means within a level of A over all levels of B, and $(\mu_1 - \mu_2)_{min}$ is the minimum difference of this same effect.[2] This results in a d^* for the interaction, which is on a slightly different scale than the d^* for the main effect. For example, in a 2 × 2 RG design, our "medium" effect size of $d^* = .5$ for the main effects approximately equates to $f = .25$ (which is Cohen's value for a medium effect size), but our $d^* = .5$ for the interaction equates to an f of approximately .18. This is not an issue in using the power tables, but one cannot directly compare the power for a main effect and an interaction, although both have an effect size of .5. We feel our d^* procedures are justified because of their ease of calculation—and the power values are read directly from these d^* estimates. Subsequent power calculations are computed under the expectation that the k means of a factor are equally spaced along the minimum-to-maximum continuum. If all means are clustered at either end of the continuum, our power calculations will slightly underestimate the true power, and if all except the two extreme means are clustered at the median, our values will slightly overestimate the true power.
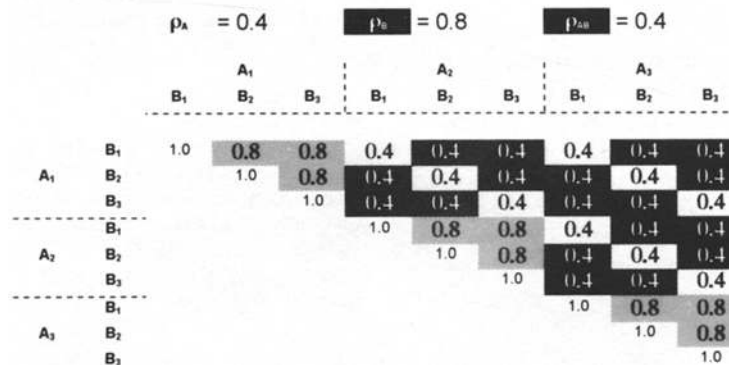
Power Determination

A MC simulation program developed by Eom (1993) and modified by the researchers of this study was used to determine power es-

Generic Matrix

Example

$\rho_A = 0.4$ $\rho_B = 0.8$ $\rho_{AB} = 0.4$
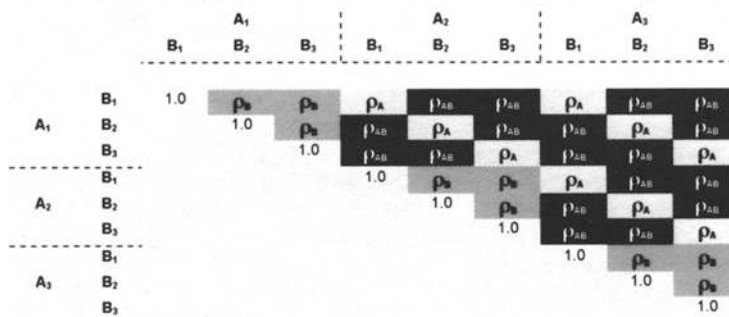
Figure 1. Representation of a correlation matrix for a 3 (A) × 3 (B) RM ANOVA: General form and numeric example. ρ_A and ρ_B represent the average correlation among the A and B (pooled) trials, respectively, and ρ_{AB} represents the average correlation among the AB coefficients having dissimilar levels.

timates. This program, written in FORTRAN 77, utilized a number of the subroutines of the International Mathematical and Statistical Library (IMSL, 1991). All computations were done in double precision. For each design condition, a population variance–covariance matrix and a mean vector were used as input to generate a population ($N = 120,000$) of continuous data from a k-variate normal distribution using the IMSL subroutine DRNMVN. Random samples of data were repeatedly drawn and subjected to the appropriate RM ANOVA test (using the IMSL subroutine ANOVA). Actual values of power were computed by totaling the number of F tests found significant for a given α and dividing the result by the overall number of tests performed for the simulation. Values of means inputted ranged from 8.8 to 10.4, and within-cell variances equaled 1.0 across the different levels of RM. These values were selected to provide the specific ESs (d^*) desired.

For each design, four different variance–covariance matrices (equivalent to correlation matrices, since all variances were 1) were constructed to represent different levels of dependencies between the levels within each factor. With two RM factors, the variance–covariance matrix consists of three submatrices, one representing the correlations among the levels of Factor A, one representing the correlations among the levels of Factor B, and one consisting of the correlation coefficients representing the relationships between the levels across the factors (the "interaction" matrix). Figure 1 provides both a generic matrix of this type for a 3 × 3 design and one of the

specific variance–covariance matrices used in this study. The average of all the ρ_A correlations, $\bar{\rho}_A$, is the value corresponding to $\bar{\rho}$ in Equation 5, which is known to be positively related to the power of the A main effect. Similarly, $\bar{\rho}_B$, the average of all the ρ_Bs, is related to the B main effect. The ρ_{AB} values are involved, in some way, in the error variances for the interaction effect. The nature of this relation, and of how $\bar{\rho}_A$ affects the estimate of σ_B^2, and $\bar{\rho}_B$ affects the estimate of σ_A^2, are investigated as the secondary purpose of this study.

In all cases, RM designs were analyzed using a univariate rather than multivariate approach since anecdotally the former is by far the more common one utilized by health and behavioral researchers, despite its more restrictive assumptions. Although the multivariate test is considered a better choice when the assumption of sphericity is violated (Rouanet & Lepine, 1970; Schutz & Gessaroli, 1987), the univariate model offers a valid and at times better approach (Mendoza et al., 1974; Muller & Barton, 1989) when it is not. Thus we restricted our power estimates to data meeting the conditions of local sphericity—that is, $\varepsilon = 1.0$ for each of the three error variance–covariance matrices. Although equal correlations are not a necessary condition for sphericity, for ease of generating the matrices we used constant values for the correlations within a factor (see Figure 1 for such an example).

To ensure reasonable accuracy of power estimates, the number of tests (replications) performed per simulation by the MC program

was set at 3,000, resulting in a standard error of proportion for true power values of .50 and .99 of .009 and .002, respectively. With 95% confidence, therefore, the accuracy of power values derived from the MC simulations of this study was expected to be about ±.018 for tests with moderate power (.50) and ±.004 for those exhibiting power at the extremes (.99 or .01).

Accuracy and Reliability of Power Estimates

Several measures were undertaken to verify the accuracy and reliability of the MC program at approximating power. First, in order to ensure that the MC routines were computing correctly, ANOVA test results generated from this program were compared with those computed from a well-known statistical analysis package (BMDP, 1988). This was done by extracting test data generated by the simulation program and then subjecting them to the appropriate statistical procedure(s) using BMDP. For all conditions examined, ANOVA statistics (mean sum of square values, F ratios) for the two-factor RM design were identical across the two programs.

A second method for assessing the accuracy of the MC program involved comparing power values from the simulation program for a one-way RM design with direct analytic calculations of power. This was done by calculating λ and using the associated probability from the noncentral F distribution (using DATASIM, Bradley, 1988). A total of 162 conditions were involved in which absolute differences in power between MC and DATASIM were determined for designs with 3, 6, and 9 RM. According to Muller and Barton (1989), absolute differences equal to or below .04 (4.0%) are sufficient for power purposes. Of the 162 differences, none were above this value and only 4 were above .025 (2.5%). It was not possible to provide an exact test for the two-way RM design because of the unavailability of the formula to compute λ; however, Bradley (personal communication, 1998) used DATASIM to run two-way simulations for the 3 × 3 design, $d^* = 5$, with all four correlation structures. Of 96 power values calculated, the largest difference between his results and ours was .027, and that was the only difference greater than .02. It was therefore concluded that both programs produced similar power estimates.

Finally, the reliability of the MC programs was assessed by replicating simulation runs under identical experimental conditions. Simulations were repeated for several conditions of the two-way RM design in which only the random number generator seed was changed. The largest range among power values from repeated simulations of a given condition was .025. On the basis of this and results from other procedures performed, the MC program was considered reliable and capable of providing accurate estimates of true power.

RESULTS AND DISCUSSION

Power Tables

Table 1 provides an example of the type of data (power estimates) generated for each of the six designs. The table gives the power of a specific test (A and B main effects, AB interaction) in the 3 × 6 RM ANOVA for the different levels of alpha (α), effect size (ES), average correlation ($\bar{\rho}$), and sample size (n). In addition, each column within a given ES and α represents power values for one of the four correlation matrices examined under the two-way RM design (e.g., the 484 matrix where $\bar{\rho}$ for the A, B, and AB effects equals .4, .8, and .4, respectively). Power can be determined by selecting the test and level of significance and crossing the column matching the estimated ES and correlation matrix with the desired sample size. As an example, power for the A main effect test of a 3 × 6 RM ANOVA (Table 1) having an ES (d^*) of .2, a $\bar{\rho}$ among the

levels of Factors A and B of .8 (i.e., the 888 matrix), and $n = 20$ would be .65 at $\alpha = .01$ and .87 at $\alpha = .05$.

As noted in the introduction to this paper, one of the purposes of this study was to develop tables of approximate power estimates for a number of specific conditions and designs in order to assist researchers in determining power and/or sample size in planning their studies. However, we also noted that should we be fortunate enough to ascertain the equations for the error terms for each repeated measures factor, it would not be necessary to present a selection of design-specific power tables. With such equations one could analytically compute a noncentrality parameter (e.g., λ) and then derive an exact estimate of power with a program such as DATASIM. As we show below, we have developed equations that provide very accurate estimates of the error variances, and thus the presentation of a set of power tables for a specific set of conditions of α, $\bar{\rho}$, n, d^*, p, and q is not necessary. We have presented one power table, for the 3 × 6 RM design, in order to examine the results germane to the first purpose of this study, that is, the effects of the correlational structure of one factor on the power of the other factor. The values given in Table 1 are the analytic values (computed using Equations 9, 10, and 11) rather than the simulated values, because the former are more exact. The simulated power estimates exhibited virtually no bias and a high degree of accuracy. For example, of the 432 calculated power values presented in Table 1, the mean difference between the simulated and calculated values was −0.002, and the largest difference was 0.024.

The Effect of ρ on Power

Of special interest are the power trends observed across the different correlation matrices within a given test and design. As expected, the larger the correlation among the levels of a factor, the greater the power of the F test for the main effect of that factor. The extent of this effect is considerable; for example, in a 3 × 6 design, with $d^* = 0.20$, $n = 15$, and $\alpha = .05$, doubling the correlation in the A factor (from .4 to .8; i.e., the 444 and 844 matrices), results in an increase in power from 0.28 to 0.72 for the A main effect (Table 1). The magnitude of the effect is similar for the B factor, even though it has twice as many levels of the RM factor as does A. With a small effect size ($\alpha = .05$ and $n = 30$), the power is .29 for the 444 matrix and increases to .78 with the 484 matrix.

The more important finding with respect to the effect of ρ on power relates to the effect of the correlations associated with one factor on the power of the test of the main effect of the other factor. Specifically, if the correlations among the levels of B are larger than those within the AB matrix (i.e., $\bar{\rho}_B - \bar{\rho}_{AB} > 0.0$), there is a reduction in the power for the test of the A effect (and the test on B is similarly affected by the A correlations). For example, in regard to the A main effect in Table 1, with the conditions $d^* = 0.20$, $\alpha = .05$ and $n = 30$, it can be seen that for $\bar{\rho}_A = .4$, the power is .55 when $\bar{\rho}_B$ and $\bar{\rho}_{AB}$ are also low (.4), but the power for A drops markedly to .16

Table 1
Power of the A, B, and AB Effects for A 3(A) × 6(B) ANOVA
With Repeated Measures on Both Factors

		ES											
		Small (.20)				Medium (.50)				Large (.80)			
$\bar{\rho}$ for A:		.4	.4	.8	.8	.4	.4	.8	.8	.4	.4	.8	.8
$\bar{\rho}$ for B:		.4	.8	.4	.8	.4	.8	.4	.8	.4	.8	.4	.8
$\bar{\rho}$ for AB:		.4	.4	.4	.8	.4	.4	.4	.8	.4	.4	.4	.8
α	n												
		Main Effect of Factor A (3 Levels)											
.01	5	.03	.01	.07	.07	.17	.03	.62	.62	.53	.09	.98	.98
	10	.06	.02	.24	.24	.59	.09	.99	.99	.98	.31	1.00	1.00
	15	.11	.03	.45	.45	.87	.18	1.00	1.00	1.00	.57	1.00	1.00
	20	.17	.03	.65	.65	.97	.27	1.00	1.00	1.00	.77	1.00	1.00
	25	.23	.04	.79	.79	.99	.37	1.00	1.00	1.00	.89	1.00	1.00
	30	.30	.05	.89	.89	1.00	.47	1.00	1.00	1.00	.96	1.00	1.00
.05	5	.11	.06	.23	.23	.43	.13	.90	.90	.85	.27	1.00	1.00
	10	.19	.08	.50	.50	.84	.26	1.00	1.00	1.00	.60	1.00	1.00
	15	.28	.10	.72	.72	.97	.39	1.00	1.00	1.00	.82	1.00	1.00
	20	.38	.12	.87	.87	.99	.52	1.00	1.00	1.00	.93	1.00	1.00
	25	.47	.14	.94	.94	1.00	.64	1.00	1.00	1.00	.98	1.00	1.00
	30	.55	.16	.98	.98	1.00	.73	1.00	1.00	1.00	.99	1.00	1.00
		Main Effect of Factor B (6 Levels)											
.01	5	.02	.04	.01	.04	.08	.39	.03	.39	.31	.92	.10	.92
	10	.03	.10	.02	.10	.29	.91	.09	.91	.83	1.00	.33	1.00
	15	.05	.20	.02	.20	.53	1.00	.17	1.00	.98	1.00	.59	1.00
	20	.07	.31	.03	.31	.74	1.00	.26	1.00	1.00	1.00	.80	1.00
	25	.09	.43	.04	.43	.88	1.00	.37	1.00	1.00	1.00	.92	1.00
	30	.12	.54	.04	.54	.95	1.00	.47	1.00	1.00	1.00	.97	1.00
.05	5	.08	.13	.06	.13	.25	.68	.12	.68	.60	.99	.27	.99
	10	.11	.27	.07	.27	.54	.98	.24	.98	.96	1.00	.59	1.00
	15	.15	.41	.09	.41	.78	1.00	.37	1.00	1.00	1.00	.82	1.00
	20	.20	.55	.11	.55	.91	1.00	.50	1.00	1.00	1.00	.94	1.00
	25	.24	.68	.12	.68	.97	1.00	.62	1.00	1.00	1.00	.98	1.00
	30	.29	.78	.14	.78	.99	1.00	.72	1.00	1.00	1.00	1.00	1.00
		A × B Interaction											
.01	5	.01	.01	.01	.01	.02	.03	.03	.03	.03	.09	.09	.09
	10	.01	.02	.02	.02	.02	.07	.07	.07	.06	.30	.30	.30
	15	.01	.02	.02	.02	.03	.14	.14	.14	.11	.56	.56	.56
	20	.01	.02	.02	.02	.05	.22	.22	.22	.18	.77	.77	.77
	25	.02	.03	.03	.03	.06	.32	.32	.32	.25	.90	.90	.90
	30	.02	.03	.03	.03	.08	.42	.42	.42	.34	.96	.96	.96
.05	5	.05	.06	.06	.06	.07	.11	.11	.11	.10	.24	.24	.24
	10	.06	.07	.07	.07	.09	.21	.21	.21	.18	.55	.55	.55
	15	.06	.08	.08	.08	.12	.33	.33	.33	.28	.79	.79	.79
	20	.06	.09	.09	.09	.15	.45	.45	.45	.38	.92	.92	.92
	25	.07	.11	.11	.11	.19	.57	.57	.57	.48	.97	.97	.97
	30	.07	.12	.12	.12	.23	.67	.67	.67	.58	.99	.99	.99

$\bar{\rho}$, average of A, B, or AB correlation coefficients in a given matrix; ES, effect size (d^*); n = sample size.

when $\bar{\rho}_B$ is high (.8) and $\bar{\rho}_{AB}$ is low (.4). At first glance this suggests that it is the large $\bar{\rho}_B$ that causes the power reduction, but this is not so. In regard to the next two columns of Table 1 ($d^* = 0.20$, $\alpha = .05$, $n = 30$), with $\bar{\rho}_A = .8$, it can be seen that the power is .98 for $\bar{\rho}_B = \bar{\rho}_{AB} = .4$, and this does not change with an increase in $\bar{\rho}_B$ ($\bar{\rho}_B = \bar{\rho}_{AB} = .8$). This is because the ($\bar{\rho}_B - \bar{\rho}_{AB}$) difference is constant for these two conditions. This relationship held true for all correlation conditions simulated, and is clearly shown in the analytic expressions presented later in this paper.

Figure 2 (top panel) depicts this common power trend for different n of the A main effect of a 3 × 6 design with small ES and $\alpha = .05$. It shows that the average corre-

tion among the levels of Factor B has no effect on the power of the A main effect when the $\bar{\rho}_B$ and $\bar{\rho}_{AB}$ values are equal. On the other hand, when $\bar{\rho}_B \geq \bar{\rho}_{AB}$ (the situation in the 484 matrix), this causes a considerable decrease in the power of the A effect. A similar pattern of results occurs with the test of the B main effect (middle panel of Figure 2). These results parallel the relationship between the magnitude of the correlation of the RM factor and the power of the groups effect (the nonrepeated measures factor) in a mixed model design where the MS_{error} for the groups effect is increased by a factor of 1 + $(q - 1)\bar{\rho}$ (see Equation 6). Apparently, the same type of effect occurs in the two-way RM ANOVA; the exact

Main Effect of A

Main Effect of B

AB Interaction
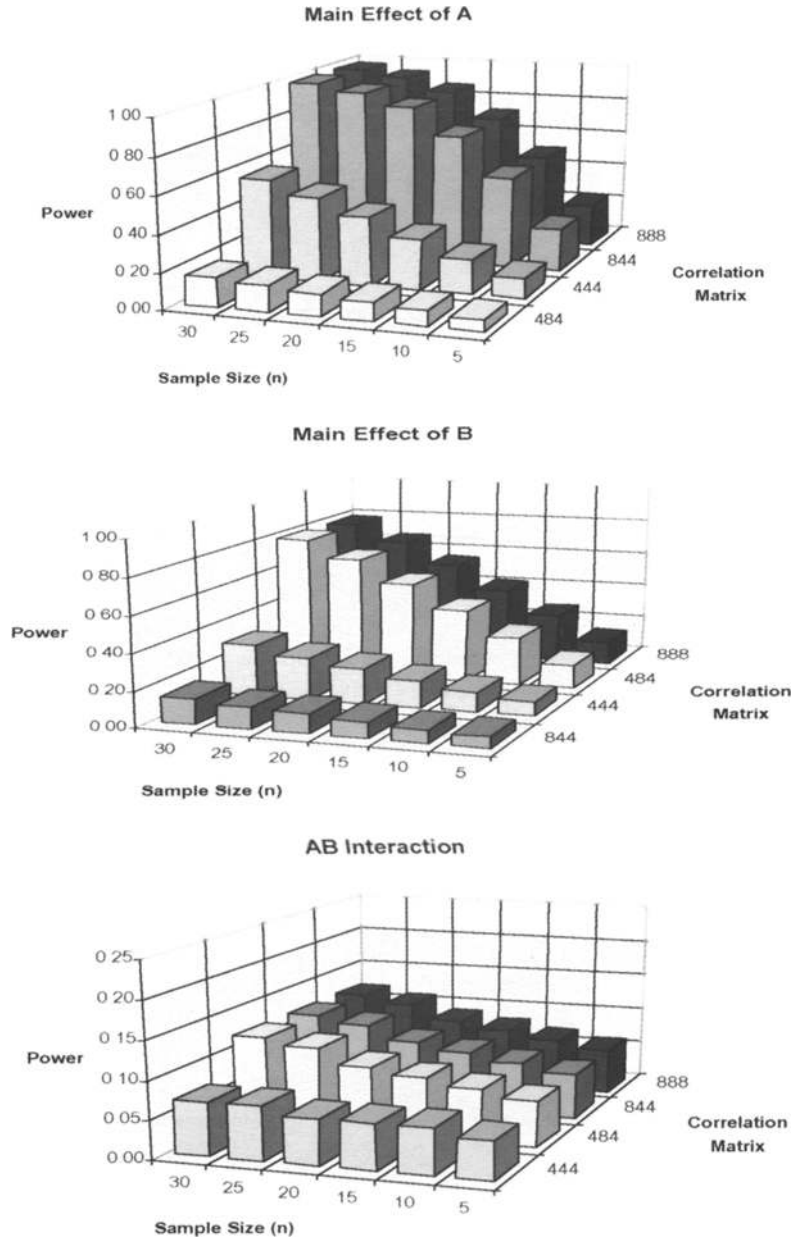
Figure 2. Comparison of power across different correlation matrices for tests of a 3 × 6 RM ANOVA design, based on small effect size and $\alpha = .05$. (The first, second, and third numbers of a matrix represent average correlation values of the A, B, and AB effects, respectively).

nature of this relationship is shown later in Equations 9 and 10.

A different power trend emerges for the tests of the AB interaction relative to those of the main effect tests. Referring to the Table 1 again, we see that matrices in which at least one factor (A, B, or both) has a $\bar{\rho} = .8$ result in the highest power for the AB test. As illustrated in the bottom panel of Figure 2, three of the four matrices (888, 844, and 484) produce equal power for the AB test

while the 444 matrix exhibits inferior power values for all n. These results suggest that power for the interaction test in the two-way RM model remains the same among different correlation structures as long as all matrices involved have at least one RM factor with a $\bar{\rho}$ among its trials equal in magnitude to the highest overall $\bar{\rho}$ observed among the AB matrices (in this case, .8). This appears to be true at least for all conditions where $\bar{\rho}_{AB}$ equals the lower $\bar{\rho}$ among the two factors.

Table 2
Mean Statistics Generated From Monte Carlo Simulation
for Different Conditions of the Two-Way RM ANOVA*

		Correlation Matrix											
		3 × 3				3 × 6				3 × 9			
Statistic	Test	444	484	844	888	444	484	844	888	444	484	844	888
MS_{effect}	A	2.47	3.30	2.09	2.09	4.34	6.34	3.93	3.92	6.12	9.45	5.79	5.82
	B	2.48	2.07	3.35	2.10	1.65	1.25	2.43	1.25	1.49	1.08	2.24	1.07
	AB	0.76	0.35	0.36	0.36	0.69	0.28	0.29	0.29	0.68	0.27	0.27	0.28
MS_{error}	A	0.60	1.40	0.20	0.20	0.61	2.60	0.20	0.20	0.60	3.81	0.20	0.20
	B	0.60	0.20	1.39	0.20	0.59	0.20	1.39	0.20	0.60	0.20	1.39	0.20
	AB	0.60	0.20	0.20	0.20	0.60	0.20	0.20	0.20	0.60	0.20	0.20	0.20
F Ratio	A	4.61	2.68	11.83	11.62	8.09	2.76	21.96	22.30	11.49	2.81	32.40	32.74
	B	4.65	11.64	2.72	11.90	2.92	6.52	1.83	6.47	2.56	5.54	1.65	5.52
	AB	1.34	1.87	1.94	1.91	1.17	1.45	1.47	1.47	1.15	1.39	1.39	1.40
Power	A	0.53	0.26	0.95	0.96	0.83	0.27	1.00	1.00	0.96	0.28	1.00	1.00
	B	0.52	0.96	0.24	0.96	0.54	0.98	0.24	0.98	0.61	0.99	0.27	0.99
	AB	0.10	0.22	0.24	0.23	0.09	0.20	0.20	0.21	0.11	0.23	0.24	0.24

*Based on medium effect size, $n = 10$ and $\alpha = .05$.

The Effect of ρ on the Mean Square Terms

Examination of the relative magnitudes of the numerator and denominator of the F ratios helps to explain these findings. Table 2 gives the average MS_{error}, mean square effect (MS_{effect}), F, and power values for different designs and tests generated from the MC simulation under medium effect size ($n = 10$ and $\alpha = .05$). Since a change in the magnitude and structure of a correlation matrix affects only σ^2_{error}, differences in power across different correlation matrices should correspond to alterations in MS_{error} when all other conditions are held constant. That is, matrices producing the greatest power for a test should have the lowest MS_{error} values among structures involved and vice versa. As Table 2 shows, this is the case among all three tests (A, B, and AB) of the two-way design. As an example, conditions for the A test of a 3 × 3 design involving an 844 or 888 matrix produce the lowest MS_{error} (.20) and highest F (11.83 and 11.62, respectively) and power values (.95 and .96), while those having a 484 matrix produce the largest MS_{error} (1.40) and concomitantly the lowest F (2.68) and lowest power (.26). Meanwhile, tests with a 444 matrix result in values in between those of the other matrices; $MS_{error} = .60$, $F = 4.61$, and power = .53. The B test shows similar results, with the exception that the rank orders of the 844 and 484 matrices are reversed. These findings indicate that the MS_{error} for a given main effect (e.g, Factor A) test is dependent not only on $\bar\rho_A$, but also on both $\bar\rho_B$ and $\bar\rho_{AB}$, and that the in-

Table 3
Mean Square Errors and F Values: A Comparison of Monte Carlo (MC) Results and Analytical Calculations (ACs)

Test	ANOVA Design	ρ Matrix	Effect Size (d^*)	σ^2	MS_{error} MC	AC	Diff.	%Diff.	F Ratio MC	AC	Diff.	% Diff.
A	2 × 3	464	.8	4.0	3.9855	4.0000	−0.0145	−0.36	8.8945	8.8074	0.0871	1.0
	2 × 6	463	.2	4.0	8.3794	8.4000	−0.0206	−0.25	1.4559	1.5340	−0.0781	−5.4
	3 × 3	843	.5	1.0	0.4018	0.4000	0.0018	0.45	15.5002	15.6004	−0.1002	−0.6
	3 × 3	922	.8	81.0	8.0821	8.1000	−0.0179	−0.22	2.8454	2.8770	−0.0316	−1.1
	3 × 4	422	.8	4.0	4.7840	4.8000	−0.0160	−0.33	8.2193	8.1778	0.0415	0.5
				Absolute Mean:	0.0142	0.32			0.0677	1.7		
B	2 × 3	464	.8	4.0	1.6103	1.6000	0.0103	0.64	7.0454	7.2500	−0.2046	−2.9
	2 × 6	463	.2	4.0	1.9986	2.0000	−0.0014	−0.07	1.1880	1.1843	0.0037	0.3
	3 × 3	843	.5	1.0	1.6064	1.6000	0.0064	0.40	4.7439	4.6769	0.0670	1.4
	3 × 3	922	.8	81.0	178.0750	178.2000	−0.1250	−0.07	1.1217	1.1194	0.0023	0.2
	3 × 4	422	.8	4.0	4.0196	4.0000	0.0196	0.49	7.7456	7.8044	−0.0588	−0.8
				Absolute Mean:	0.0325	0.33			0.0673	1.1		
AB	2 × 3	464	.8	4.0	1.5928	1.6000	−0.0072	−0.45	2.6763	2.5893	0.0870	3.3
	2 × 6	463	.2	4.0	1.2036	1.2000	0.0036	0.30	1.0982	1.0850	0.0132	1.2
	3 × 3	843	.5	1.0	0.1002	0.1000	0.0002	0.20	5.7846	5.7873	−0.0027	−0.0
	3 × 3	922	.8	81.0	8.0898	8.1000	−0.0102	−0.13	1.1613	1.1683	−0.0070	−0.6
	3 × 4	422	.8	4.0	1.6023	1.6000	0.0023	0.14	2.6929	2.7188	−0.0259	−1.0
				Absolute Mean:	0.0047	0.24			0.0272	1.2		

Note—For simulation results, F of a given test was determined by taking the average of 3,000 values generated. Analytic values were obtained by computing λ and then converting to F as follows: $E(F) = [df_2/(df_2 - 2)][1 + \lambda/df_1]$ (see Winer et al., 1991). All F values based on $n = 30$.

fluence of $\bar{\rho}_B$ becomes evident only when its value becomes greater than $\bar{\rho}_{AB}$. Although the MS_{effect} also changes across correlation matrices (Table 2), it is not responsible for the differences in power observed since, according to the expected mean squares model for a two-factor RM ANOVA, the MS_{effect} is a component of the treatment effect and error term, and therefore a change in MS_{error} will be reflected in MS_{effect} as well.

For the interaction (AB) test, the equal power and F values among those matrices with at least one factor having $\bar{\rho} = .8$ (484, 844, and 888) are due to equal MS_{error} values. In this case, when the $\bar{\rho}$ of either the A or B pooled matrix is equal to .8, the error variance remains the same as when the $\bar{\rho}$ of the overall AB matrix is equal to .8. Only when the $\bar{\rho}$ among trials of both factors decreases (444 matrix) do we see an increase in the error term, resulting in a decrease in F and power. Unlike main effect tests, the number of repeated measures does not affect the MS_{error}. Thus, the error variance of AB tests is affected solely by the magnitude of the $\bar{\rho}$ of the pooled matrix (A or B) having the highest average correlation among its trials, at least when the lower $\bar{\rho}$ of the pooled matrices equals $\bar{\rho}_{AB}$. Evidently, the overall $\bar{\rho}$ of the AB matrix does not seem to be a determining factor.

Approximations of Equations to Estimate MS_e

The results of this simulation study provided useful information for identifying the analytical expressions of error variance, and thus λ, for tests of the two-way RM ANOVA. By examining the effects of different $\bar{\rho}$ values and number of trials on residual variances and using a logical iterative process, we developed approximations that account for the relationship among these parameters. We started with the assumption that the error variance for a main effect would decrease as a function of the magnitude of the average correlation for that effect and increase as a function of the average correlation for the other effect (as in the mixed model design). That is, the error variance for the A main effect would include some form of the terms $\sigma^2(1 - \bar{\rho}_A)$ and $\sigma^2(q - 1)\bar{\rho}_B$, where q is the number of levels of the B main effect. The task was to determine the exact nature of these terms and to incorporate the $\bar{\rho}_{AB}$ term into the equation. The following expressions for the mean square errors were subsequently constructed:

$$\sigma_e^2 = \sigma^2(1 - \bar{\rho}_A) + \sigma^2(q - 1)(\bar{\rho}_B - \bar{\rho}_{AB})$$

for the main effect test of A,   (9)

$$\sigma_e^2 = \sigma^2(1 - \bar{\rho}_B) + \sigma^2(p - 1)(\bar{\rho}_A - \bar{\rho}_{AB})$$

for the main effect test of B,   (10)

$$\sigma_e^2 = \sigma^2(1 - \rho_{max}) - \sigma^2(\bar{\rho}_{min} - \bar{\rho}_{AB})$$

for the AB interaction,   (11)

where p and q are the number of levels of the A and B factors, respectively; $\bar{\rho}_A$ and $\bar{\rho}_B$ are the averages of the off-diagonal correlation coefficients of the A and B matrices, respectively; $\bar{\rho}_{AB}$ is the average of the correlation coeffi-

cients of all AB pairs of the overall (AB) matrix with dissimilar levels; $\bar{\rho}_{max}$ and $\bar{\rho}_{min}$ are the highest and lowest average correlation values in the two matrices; and σ^2 is the within-cell variance of the dependent variable involved.

As can be seen, all three equations are composed of two components. The first part of each equation [e.g., $\sigma^2(1 - \bar{\rho}_A)$] is similar to the error variance formula for the trials and interaction tests of the two-factor mixed design (see Equation 5) and reflects how σ_e^2 is altered as $\bar{\rho}_A$, $\bar{\rho}_B$, or $\bar{\rho}_{max}$ increases or decreases. The second component in the main effect approximations [e.g., $\sigma^2(q - 1)(\bar{\rho}_B - \bar{\rho}_{AB})$] resembles that of Equation 6 for the group test of the mixed model. Here, an increase in either the number of pooled trials or $\bar{\rho}$ of the pooled matrix will increase σ_e^2 when all other variables are held constant. Also evident in Equations 9, 10, and 11 is how the second component cancels out when $\bar{\rho}_{AB}$ equals $\bar{\rho}_A$, $\bar{\rho}_B$, or $\bar{\rho}_{min}$, resulting in an effect on σ_e^2 that is due entirely to the main effects or maximum average correlation. Furthermore, note that if $\bar{\rho}_A = \bar{\rho}_B = \bar{\rho}_{AB} = \bar{\rho}$, then all three equations reduce to $\sigma_e^2 = \sigma^2(1 - \rho)$, the error term for designs with only one RM factor. Finally, if $\bar{\rho}_A$ and $\bar{\rho}_{AB}$ are set to zero, as they would be if A was an independent groups factor, Equation 9 reduces to Equation 6 and Equations 10 and 11 reduce to Equation 5, where Equations 5 and 6 give the expected error terms for a mixed model design.

These approximations of σ_e^2 for the two-factor RM ANOVA were subsequently validated under other design conditions with different correlation matrices and RM levels. The correlation matrices used in this validation employed values that were very different (e.g., $\bar{\rho}_A = .9$, $\bar{\rho}_B = .2$, $\bar{\rho}_{AB} = .2$) from those used in the simulations (.4 and .8 for all matrices) that generated the power values used to "derive" Equations 9–11. Table 3 provides a comparison of MS_{error} and F values generated from MC simulations and those derived using Equations 9–11. These derivations were performed by first calculating σ_e^2 using Equations 9, 10, and 11; then computing λ using Equations 2, 3, and 4; and then using standard formulae for converting λ to F. For all conditions examined, MS_{error} values computed using the approximations were identical to those generated through simulation to two significant figures, while absolute differences of F between the two methods never surpassed a value of .21. Hence, these equations provide a reasonably accurate means of approximating the residual variance of tests in the two-way RM design. When Equations 9–11 are substituted for the respective denominator terms of Equations 2–4, λ values can be computed directly for univariate ANOVA designs with two within-subjects variables when estimates of the population means, σ^2, and $\bar{\rho}$ values are available. Given this calculated value for λ, power can then be calculated from the noncentral F distribution, and it is a relatively easy step for programmers to include the two-way RM ANOVA in their computer programs.

Summary and Conclusions

In light of the absence of a method for approximating a priori power for univariate ANOVA tests with two within-subjects variables, power estimates for the two-

way RM design were developed using MC simulation. Power values were generated for each test (A, B, AB) of a 2 × 3, 2 × 6, 2 × 9, 3 × 3, 3 × 6, and 3 × 9 design under a variety of experimental conditions including different effect sizes (small, medium, and large), different average correlation values (.4 and .8), different levels of significance (.01 and .05), and different sample sizes (5, 10 ,15, 20, 25, and 30). It was shown that the magnitude of the difference $(\bar{\rho}_B - \bar{\rho}_{AB})$ is strongly related (inversely) to the power of the F test on Factor A (and similarly, $\bar{\rho}_A - \bar{\rho}_{AB}$ affects the power of Factor B). In addition, formulae to compute approximations of error variances for the A, B, and AB tests of the two-way RM ANOVA were developed by evaluating power and MS_{error} trends across different correlation structures. Evidence given for the accuracy of these approximations supports their use in computing λ, thus providing the first analytical method for estimating the power of these designs without requiring previous knowledge of the mean square error variance of a given test. As a result of this work, researchers may now determine the power for their design by computing λ directly and then referring to conventional power tables or computer programs. It is hoped that the approximations of error variance presented here can be incorporated into the algorithms of existing or future computer power programs, thus providing an efficient means of estimating power for these RM ANOVA designs.

REFERENCES

BMDP [Computer software] (1988). Los Angeles: University of California Press.

BORENSTEIN, M., & COHEN, J. (1988). *Statistical Power Analysis: A Computer Program* [Computer program]. Hillsdale, NJ: Erlbaum.

BRADLEY, D. R. (1988). *DATASIM* [Computer program]. Lewiston, ME: Desktop Press.

BRADLEY, D. R., & RUSSELL, R. L. (1998). Some cautions regarding statistical power in split-plot designs. *Behavior Research Methods, Instruments, & Computers*, **30**, 462-477.

BRADLEY, D. R., RUSSELL, R. L., & REEVE, C. P. (1996). Statistical power in complex experimental designs. *Behavior Research Methods, Instruments, & Computers*, **28**, 319-326.

COHEN, J. (1988). *Statistical power analysis for the behavioral sciences* (3rd ed.). Hillsdale, NJ: Erlbaum.

DAVIDSON, M. L. (1972). Univariate versus multivariate tests in repeated-measures experiments. *Psychological Bulletin*, **77**, 446-452.

DODD, D. H., & SCHULTZ, R. F., JR. (1973). Computational procedures for estimating magnitude of effect for some analysis of variance designs. *Psychological Bulletin*, **79**, 391-395.

EOM, H. J. (1993). *The interaction effects of data categorization and noncircularity of the sampling distribution of generalizability coefficients in analysis of variance models*. Unpublished doctoral dissertation, University of British Columbia.

ERDFELDER, E., FAUL, F., & BUCHNER, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, & Computers*, **28**, 1-11.

FRIEDMAN, H. (1982). Simplified determinations of statistical power, magnitude of effect, and research sample sizes. *Educational & Psychological Measurement*, **42**, 521-526.

GORMAN, B. S., PRIMAVERA, L. H., & ALLISON, D. B. (1995). POWPAL: A program for estimating effect sizes, statistical power, and sample sizes. *Educational & Psychological Measurement*, **55**, 773-776.

GRIMA, A. M., & WEINBERG, S. (1987). *An analysis of repeated measures data: An exploration of alternatives (MANOVA)*. Unpublished doctoral dissertation, New York University.

INTERNATIONAL MATHEMATICAL AND STATISTICAL LIBRARIES (1991). (10th ed.). Houston, TX: Visual Numbers Inc.

KIRK, R. E. (1995). *Experimental design: Procedures for the behavioral sciences*. Pacific Grove, CA: Brooks/Cole.

KRAEMER, H. C., & THIEMANN, S. (1987). *How many subjects?* Beverly Hills, CA: Sage.

LEVIN, J. R. (1997). Overcoming feelings of powerlessness in "aging" researchers: A primer on statistical power in analysis of variance designs. *Psychology & Aging*, **12**, 84-106.

LIPSEY, M. W. (1990). *Design sensitivity*. Newbury Park, CA: Sage.

MARCUCCI, M. (1986). A comparison of the power of some tests for repeated measurements. *Journal of Statistical Computation & Simulation*, **26**, 37-53.

MENDOZA, J. L., TOOTHAKER, L. E., & NICEWANDER, W. A. (1974). A Monte Carlo comparison of the univariate and multivariate methods for the groups by trials repeated-measures design. *Multivariate Behavioral Research*, **9**, 165-177.

MULLER, K. E., & BARTON, C. N. (1989). Approximate power for repeated-measures ANOVA lacking sphericity. *Journal of the American Statistical Association*, **84**, 549-555.

MULLER, K. E., LAVANGE, L. M., RAMEY, S. L., & RAMEY, C. (1992). Power calculations for general linear multivariate models including repeated measures applications. *Journal of the American Statistical Association*, **87**, 1209-1224.

MULVENON, S. W., & BETZ, M. A. (1993). *Analytic formulae for power analysis in repeated measures designs*. Unpublished doctoral dissertation, Arizona State University.

NUMBER CRUNCHER STATISTICAL SYSTEM. (1991). PASS (Power Analysis and Sample Size), Version 1.0. Kaysville, UT: Jerry L. Hintze.

PEARSON, E. S., & HARTLEY, H. O. (1951). Charts of the power function of the analysis of variance tests, derived from the noncentral F-distribution. *Biometrika*, **38**, 112-130.

ROBEY, R. R., & BARCIKOWSKI, R. S. (1984). Calculating the statistical power of the univariate and the multivariate repeated measures analyses of variance for the single group case under various conditions. *Educational & Psychological Measurement*, **44**, 137-143.

ROCHON, J. (1991). Sample size calculations for two-group repeated-measures experiments. *Biometrics*, **47**, 1383-1398.

ROTTON, J., & SCHONEMANN, P. H. (1978). Power tables for analysis of variance. *Educational & Psychological Measurement*, **38**, 213-229.

ROUANET, H., & LEPINE, D. (1970). Comparison between treatments in a repeated-measurement design: ANOVA and multivariate methods. *British Journal of Mathematical & Statistical Psychology*, **23**, 17-163.

SCHUTZ, R. W., & GESSAROLI, M. E. (1987). The analysis of repeated measures designs involving multiple dependent variables. *Research Quarterly for Exercise & Sport*, **58**, 132-149.

TANG, P. C. (1938). The power function of the analysis of variance tests with tables and illustrations of their use. *Statistical Research Memoirs*, **2**, 126-149.

VONESH, E. F., & SCHORK, M. A. (1986). Sample sizes in the multivariate analysis of repeated measurements. *Biometrics*, **42**, 601-610.

WINER, B. J., BROWN, D. R., & MICHELS, K. M. (1991). *Statistical principles in experimental design* (3rd ed.). New York: McGraw-Hill.

NOTES

1. Some of these tables require λ to be converted to ϕ or f in order to determine power.

2. For two-way RM designs having three or more levels on both factors, d^* was calculated as in Equation 8, but since many values for the numerator term of this function are possible (e.g., one for differences between the A1 and A2 cell means, another for differences between A1 and A3, another for the differences between A2 and A3, etc.), only the numerator yielding the largest absolute difference was used to compute effect size.