Behavior Research Methods
2010, 42 (2), 366-372
doi:10.3758/BRM.42.2.366

Articles From the SCiP Conference

Applying the permutation test to factorial designs

D. J. K. Mewhort, Brendan T. Johns, and Matthew Kelly
Queen's University, Kingston, Ontario, Canada

The permutation test follows directly from the procedure in a comparative experiment, does not depend on a known distribution for error, and is sometimes more sensitive to real effects than are the corresponding parametric tests. Despite its advantages, the permutation test is seldom (if ever) applied to factorial designs because of the computational load that they impose. We propose two methods to limit the computation load. We show, first, that orthogonal contrasts limit the computational load and, second, that when combined with Gill's (2007) algorithm, the factorial permutation test is both practical and efficient. For within-subjects designs, the factorial permutation test is equivalent to an ANOVA when the latter's assumptions have been met. For between-subjects designs, the factorial test is conservative. Code to execute the routines described in this article may be downloaded from http://brm.psychonomic-journals.org/content/supplemental.

Because computation is becoming increasingly inexpensive, statistical algorithms that would have required a supercomputer a decade ago are now practical on a desktop personal computer. One example is the permutation test, a computationally intensive alternative to the *F* or *t* test. Unlike the *F* test, the permutation test (also called a randomization test) does not assume a particular distribution for error, and, unlike nonparametric tests that transform the data into ordinal values, it uses all information in the data (see Bradley, 1968).

The Permutation Test

Because the permutation test may be unfamiliar, a review of its logic is in order. Perhaps the most straightforward comparative experiment is the two-group completely randomized experimental design. The unit of analysis in such an experiment (i.e., the subjects) might be plots of land in a study of fertilizers, pigs in a study of animal growth, or undergraduates in a study of recall. At the outset of the study, a pool of subjects is selected to be as uniform as possible, but the individual subjects within the pool are inherently variable. The experimenter divides the subjects into conditions: an experimental cell and a control cell. To assess the possibility that variability among the subjects is large enough to mask the treatment, we turn to statistical reasoning.

Statistical reasoning enlists probability theory to deal with the possibility that variability among subjects has masked the treatment of interest. The justification for applying probability concepts rests on how the subjects were assigned to the conditions in the experiment before the treatment was administered. Assignment *at random* implies that each subject has an equal chance to be assigned to either condition; random assignment introduces chance to the experiment and simultaneously defines a direct model for how chance may affect the outcome. If the treatment has no effect, differences between the groups must reflect subject-to-subject variability. Before the groups were formed, however, each subject had an equal chance of being assigned to either group. When he or she ran the experiment, the experimenter picked one assignment at random from the set of possible assignments. By using a random assignment, the experimenter introduced a chance component to the experiment. The chance component justifies the use of probability theory to assess the outcomes of the possible assignments. To do so, he or she must first select an appropriate measure with which to compare the cells (e.g., the difference in two means), then calculate that measure for each of the ways in which the subjects could have been assigned to the cells. Armed with the chance calculation, the experimenter need only count the number of outcomes that would have yielded a difference in the measure as extreme as, or more extreme than, the difference actually obtained. If the number of extreme outcomes is small relative to the number of possible outcomes, the experimenter has to accept one of two propositions: (1) The difference reflects noise, and chance has played an unlikely trick, or (2) the effects of the treatment are real. It is conventional to reject the first proposition (known as the null hypothesis, H_0) in favor of the second if the number of extreme outcomes represents less than 5% of the possible outcomes. The test leading to the choice is called the *test for significance*, and a test conducted by considering all of the possible outcomes is called a *per-*

D. J. K. Mewhort, mewhortd@queensu.ca

mutation, or *randomization*, test (Box & Andersen, 1955; Pitman, 1937).

A permutation test is tedious to compute. For example, if two cells each contain 10 subjects, there are 184,756 combinations to consider.[1] Hence, even though the chance model maps directly onto the procedure of the experiment—a procedure common in many areas of science, including agriculture, biology, medicine, and psychology—the permutation test is not widely used. Instead, researchers use a test based on an indirect model for chance, first proposed by C. F. Gauss (1777–1855) as part of an analysis of measurement error (see Stigler, 1986).

Gauss's analysis of measurement error treats each score as the sum of two parts: a constant defined by the conditions common to the subjects (e.g., the apparatus and measurement procedures) and an error component defined by the individual subjects when the measurement was taken. Because the subjects were assigned to cells at random, the error values can be treated as random samples from a single population defined by the constant and by measurement error. The test for significance estimates the probability that the scores came from the same distribution of error, assuming, of course, that the treatment(s) had no real effect.

A test for significance based on sampling theory requires calculation of only the mean and variance for each cell, and, for that reason, the arithmetic is easier to compute than the arithmetic for a permutation test. The fly in the ointment is that the indirect chance model requires the experimenter to know (or assume) the distribution of error. Gauss derived the normal distribution and discovered that it often describes measurement error; experimenters routinely rely on his discovery. Whether measurement error is well described by the normal distribution remains an open question (e.g., Mecceri, 1989). According to Gabriel Lippmann (1908 Nobel Laureate in Physics), empiricists accept it because they think it is a mathematical theorem, and mathematicians accept it because they think it is an experimental fact (cited by Thompson, 1942).

Because the permutation test's chance model is based on the assignment of experimental units to conditions, the test does not assume a particular distribution of error. Nevertheless, it works best when error has the properties that Gauss built into his account of measurement error: Small errors are more frequent than large ones, and the distribution is symmetrical, so the mean of the measurements is not biased systematically (Hoeffding, 1952). Because the chance model is based on the assignment of subjects to conditions, each test is local to a particular experiment, and, for that reason, how the underlying distribution of error affects the test's performance remains largely unexplored (but see Baker & Collier, 1966; Keller-McNulty & Higgins, 1987; Kempthorne & Doerfler, 1969; Romano, 1990). Finally, because the permutation test closely parallels the procedure of a comparative experiment, there is good reason to prefer it over parametric tests (e.g., Ludbrook & Dudley, 1998), and because it does not depend on a particular error distribution, it can be more sensitive to true differences among cells than the corresponding F test can (e.g., Mewhort, 2005).

Perhaps the best-known example of a permutation test is Fisher's (1966) reanalysis of Charles Darwin's data from a study of plant fertilization. In Darwin's experiment, 15 pairs of plants were grown in pots, one pair per pot. One randomly selected plant from each pair was self-fertilized, the other was cross-fertilized, and Darwin measured the plants' growth as a function of fertilization. There are 2^{15} possible outcomes for the experiment; the set of outcomes can be computed by systematically swapping the sign of the difference in growth for the two plants in each pot (see Mewhort, 2005, for a 25-line f90 subroutine to compute the set of outcomes). The normal-error chance model yielded the same result as the permutation test, and Fisher used that fact to argue for the validity of tests based on normal error. In effect, Fisher treated the permutation test as the benchmark against which to judge the normal-error model. In his words, because permutation tests "assume less knowledge, or more ignorance, of the experimental material than do the standard [normal-distribution] tests" (p. 47), they provide the means with which to assess an experiment when there is reason to believe that "simpler tests may have been injured by departures from normality" (p. 48).

APPLYING THE RANDOMIZATION TEST TO FACTORIAL EXPERIMENTS

Despite its several advantages, discussions of the permutation test are usually focused on two-cell designs largely because the randomization test becomes unmanageable in larger designs. Recall the two-cell completely randomized example. With 10 subjects in each cell, the number of combinations of possible outcomes was $C(20,10) = 184{,}756$. Three or four times that number is manageable on a modern PC. Suppose, however, that the experimental question calls for a 2×2 factorial design of four cells with 10 subjects in each cell. Now the number of combinations is huge: $C(40,10) \times C(30,10) \times C(20,10) = 847{,}660{,}528 \times 30{,}045{,}015 \times 184{,}756 = 4{,}705{,}360{,}871{,}073{,}570{,}227{,}520$—a value large enough to render the randomization test impractical for routine work.

Is there a practical way to use the permutation test in a factorial experiment? In this article, we suggest that there is a practical technique, at least for within-subjects experimental designs. In particular, we show that the computational load can be reduced by using orthogonal contrasts to exploit Gill's (2007) new computational algorithm to count cases.

Reducing the Computational Labor

Approximate tests based on sampling. One way to reduce the load is to exploit an approximate method based on sampling a subset of the possible outcomes. Hayes (2000), for example, studied the problem of heterogeneity of variance for cells with unequal numbers of observations by sampling 5,000 of the possible outcomes instead of computing the full number of possible outcomes (see also Edgington, 1995; Manly, 1997; Mewhort, Kelly, & Johns, 2009).

The approximate permutation test has the advantage of limiting the computational load, but it introduces a potential

problem. As Pagano and Tritchler (1983) noted, an approximate test is unsatisfying, because it raises "the possibility of different investigators obtaining different results with the same data" (p. 435).[2] In the 27 years since Pagano and Tritchler noted the problem, computing costs have dropped dramatically. As a result, one can minimize the problem by increasing the sample size. Nevertheless, what is needed is a method that will limit the computational load without introducing the possibility that different investigators might obtain different conclusions from the same data.

Gill's (2007) algorithm. Gill (2007) invented an extremely clever algorithm that brings the computing cost for a two-cell permutation test into manageable proportions. His method involves a Heaviside impulse function and a Fourier expansion to count extreme cases. Briefly, under H_0, all combinations of the data in a permutation test are equally likely. The idea is to compute the proportion of cases that is as extreme as, or more extreme than, the data observed. Gill defined a statistic T with an observed value t. Hence, the one-tailed probability of interest can be defined as $p(T > t) + p(T = t)/2$.

To compute the probability, Gill (2007) exploited the Heaviside function, H,

$$H(x) = \begin{cases} 1, & x > 0 \\ \frac{1}{2}, & x = 0 \\ 0, & x < 0 \end{cases}. \tag{1}$$

Using the Heaviside function, the one-tailed alpha can be defined as

$$a = \frac{1}{N}\sum_{r=1}^{N} H\left(t_r - t\right), \tag{2}$$

where t_r is the value on the rth combination. To evaluate alpha, Gill used the Fourier expansion; that is,

$$H(x) = \frac{1}{N} + \frac{2}{\pi}\Im\sum_{k'=1}^{\infty}\frac{\exp(ikx)}{k}, \tag{3}$$

where $k = 2 \times k' - 1$ (k' is an iteration parameter with a theoretical limit at infinity), and $\Im(a)$ is the imaginary part of a. For practical purposes, 50 terms of the series yields a satisfactory answer; hence, we set the upper limit for k' to 50. To ensure the validity of the expansion (i.e., to ensure that $a < \pi$), Gill scales the data so that

$$\max\left(t - t_r\right) = 9 \times \frac{\pi}{10}. \tag{4}$$

The value of $\max(t - t_r)$ is easy to compute by first ranking the data to obtain the most extreme combination.

With Gill's (2007) algorithm, the computational cost of computing a two-cell permutation test can be brought to a practical level on a recent PC; it is no more costly than computing an F or t, and it is vastly faster than computing the full enumeration of all combinations (an f90 program to compute Gill's algorithm is available at http://brm.psychonomic-journals.org/content/supplemental).

Approximate Tests Without Sampling

Gill's (2007) algorithm reduces the computational load magnificently in two-cell designs. If a factorial design could be recast into a set of two-cell comparisons, Gill's algorithm could be applied to factorial designs. The idea in recasting factorial designs into two-cell comparisons is to reduce the computational load while escaping the potential problem, identified by Pagano and Tritchler (1983), of different investigators obtaining different results from the same data.

Table 1
Orthogonal Contrasts for a 2 × 2 Factorial Design

	A_1		A_2	
Factor	B_1	B_2	B_1	B_2
A	−1	−1	1	1
B	−1	1	−1	1
A×B	1	−1	−1	1

Orthogonal contrasts provide a possible method with which to decompose factorial designs (see Doncaster & Davey, 2007). Orthogonal contrasts are usually used in an ANOVA to partition variance into nonoverlapping components. To illustrate, consider the sum of squares (SS) for any arbitrary four values. Using M to denote the mean of the four values and m_i to denote each of the separate values, the SS_{values} can be defined as

$$SS_{values} = \sum_{i=1}^{4}\left(M - m_i\right)^2. \tag{5}$$

Table 1 shows orthogonal contrasts for 2 × 2 factorial designs. The table shows the two levels of each factor (labeled A_1, A_2, B_1, and B_2) associated with a weight indicating how the cell contributes to each factor. For Factor A, for example, there are two cases labeled A_1, and, to compute the A effect, their combined values are compared with the combined values for the two cases of A_2. The weights for Factors A and B appear in the first rows of the table, and the weights for their interaction—calculated by multiplying the weights for the main effects—appear in the bottom row.

With the weights presented in Table 1, it is easy to compute the SS_{values} into three components, one for each factor: A, B, and A×B. For Factor A, we use WT_{Ai} to refer to the four weights that define the A effect; for Factor B, we use WT_{Bi} to refer to the four weights that define the B effect; and so forth. For Factor A, the calculation is

$$SS_A = \frac{\sum_{i=1}^{i=4}\left(WT_{Ai} \times m_i\right)^2}{\sum_{i=1}^{i=4}\left(WT_{Ai}^2\right)}. \tag{6}$$

The SSs for the other components are defined analogously.[3]

Because the comparisons are orthogonal, the $SS_{means} = SS_A + SS_B + SS_{A\times B}$; that is, all of the information (i.e., variability) among the scores has been taken into account in nonoverlapping parts. Nonorthogonal comparisons are not independent. Because nonorthogonal contrasts use overlapping information, the corresponding sum of nonorthogonal contrasts will not equal the SS for the scores.

Using orthogonal contrasts to recast a factorial design into two-cell comparisons, the computational load is greatly reduced. If each cell has 10 subjects in a com-

pletely randomized experimental design, the total number of ways in which they can be assigned to the cells is

$$C(40,10) \times C(30,10) \times C(20,10) = 4{,}705{,}360{,}871{,}073{,}570{,}227{,}520\ (\sim 4 \times 10^{21}).$$

Using orthogonal contrasts, however, each factor (A, B, and A×B) combines the data into two 20-subject cells. Hence, each factor requires only $C(40,20) = 137{,}846{,}528{,}820$ combinations. Combining the three comparisons, the factorial analysis would require $3 \times C(40,20) = 413{,}539{,}586{,}460$ combinations—a reduction in the computational cost by more than 10 orders of magnitude (specifically, by a factor of 11,378,259,845). In summary, using orthogonal contrasts to define the comparisons, the number of combinations can be reduced to a fraction of the total, but, unlike sampling-based approximate tests, there is no danger that different investigators might obtain different results using the same data.

Reducing the computational requirements by several orders of magnitude is a step forward, however, only if the resulting comparisons are sound. Although approximate tests based on sampling reduce the computational load, their validity is based on the idea that a large sample of possible comparisons is representative of the complete set. Approximate tests are open to sampling error, as Pagano and Tritchler (1983) noted. The question remains, then, whether orthogonal contrasts cheat the logic of the permutation test. Recall that the permutation test requires us to assess the proportion of cases that yield results as extreme as or more extreme than the results obtained, not a subset of those cases. Because the contrasts organize the analyses into components that examine fewer than the total number of the cases, a skeptic might ask whether permutation tests based on orthogonal contrasts are fair to the logic of the test.

One way to address the skeptic's question is to compare results using permutation tests based on contrasts against the corresponding results using an ANOVA. The logic, here, is to apply Fisher's (1966) validation strategy in reverse. Fisher used the permutation test to validate tests based on normal error. The following analyses reverse the logic by comparing results from an ANOVA against results from randomization tests. In both cases, the tests will be based on orthogonal contrasts under conditions that meet the assumptions of an ANOVA. If the exact and distribution-based tests yield substantially the same result, it follows that permutation tests based on orthogonal contrasts do not cheat the logic of the randomization test.

Correlated observations. The first Monte Carlo analysis is based on a repeated- or correlated-measures design of the sort illustrated by Fisher's (1966) reanalysis of Darwin's plant fertilization experiment. In a repeated measures design, each condition is administered to the same subjects, and the order of the conditions is randomized across time. The chance model is based on the number of ways in which each condition can be ordered across subjects. For a two-cell example with N subjects, there are 2^N permutations to consider; for a four-cell example, there are $(4!)^N$ permutations to consider. Using orthogonal comparisons, we can apply Gill's (2007) algorithm to three contrasts, each of which requires 2^N permutations.

Table 2
Orthogonal Contrasts for a 2 × 2 × 2 Factorial Design

	A_1				A_2			
	B_1		B_2		B_1		B_2	
Factor	C_1	C_2	C_1	C_2	C_1	C_2	C_1	C_2
A	1	1	1	1	−1	−1	−1	−1
B	1	1	−1	−1	1	1	−1	−1
C	1	−1	1	−1	1	−1	1	−1
A×B	1	1	−1	−1	−1	−1	1	1
A×C	1	−1	1	−1	−1	1	−1	1
B×C	1	−1	−1	1	1	−1	−1	1
A×B×C	1	−1	−1	1	−1	1	1	−1

In the first Monte Carlo study, we generated data for each of 20 subjects for the eight cells of a three-factor experiment (Factors A, B, and C; see Table 2). We used a random number generator adapted from software by Press, Teukolsky, Vetterling, and Flannery (1992; called *ran3*) combined with a normal-deviate filter (called *gasdev*; also adapted from Press et al., 1992). We selected the parameters so that the sensitivity of the tests for each effect (A, B, C, A×B, A×C, B×C, and A×B×C) would cover a full range—that is, so that the probability of rejecting the null hypothesis at the 5% level of significance would range from .05 with a null treatment to approximately 1 with a large separation between the means. Specifically, we set the mean, μ, of the normal generator at 0 and the standard deviation, σ, at 1. We increased the separation among the means for each factor and interaction from 0 to 5.5 in 11 equal steps.

For each set of artificial data, we computed seven permutation tests defined by contrasts for the comparisons shown in Table 2. In addition, we computed the corresponding ANOVA using the same data; each F test used an error term based on the interaction of the contrast with subjects (i.e., $C \times S$)—the standard procedure for a factorial design.

To facilitate the calculation, we defined a vector $\boldsymbol{\beta}_C$ of difference scores for each contrast, C, calculated by weighting each subject's data by the appropriate contrast weights; that is,

$$\boldsymbol{\beta}_{iC} = \sum_{j=1}^{8} \text{Score}_{ij} \times WT_{jC}, \tag{7}$$

where i indexes the subject, and j indexes each of the eight cells in turn.

Armed with $\boldsymbol{\beta}_C$, the necessary SS can be computed:

$$\text{SS}_C = \frac{\left(\sum_{i=1}^{N} \boldsymbol{\beta}_{iC}\right)^2}{N \times \sum_{j=1}^{8} WT_{jC}^2} \tag{8}$$

and

$$SS_{C\times S} = \frac{\sum_{i=1}^{N} \beta_{iC}^2}{N \times \sum_{j=1}^{8} WT_{jC}^2} - SS_C. \quad (9)$$

Hence, the $F_C = SS_C / SS_{C\times S}/(N-1)$. For the ANOVA, we calculated the probability using routines recommended by Press et al. (1992).

To compute the permutation test, we calculated the differences between the means for each possible factor (i.e., the **β** vectors) and tallied the number of outcomes as extreme as or more extreme than the obtained (initial) outcome using Gill's (2007) algorithm. The resulting probability was defined by the ratio of the number of extreme cases to the number of possible assignments. Finally, we repeated the exercise 10,000 times to obtain stable estimates.

Note that the ANOVA and the permutation test use the same data from the experiment, namely the **β** vectors. Nevertheless, the probability calculations are based on very different principles: sampling from the same Gaussian distribution and permutation of possible differences, respectively.

Figure 1 summarizes the results. The top panel shows a family of sensitivity curves obtained by calculating the proportion of times that the *F* test allowed the null hypothesis to be rejected at the 5% level of significance as a function of the separation between means. The bottom panel shows the corresponding data using the permutation test instead of the *F* test.

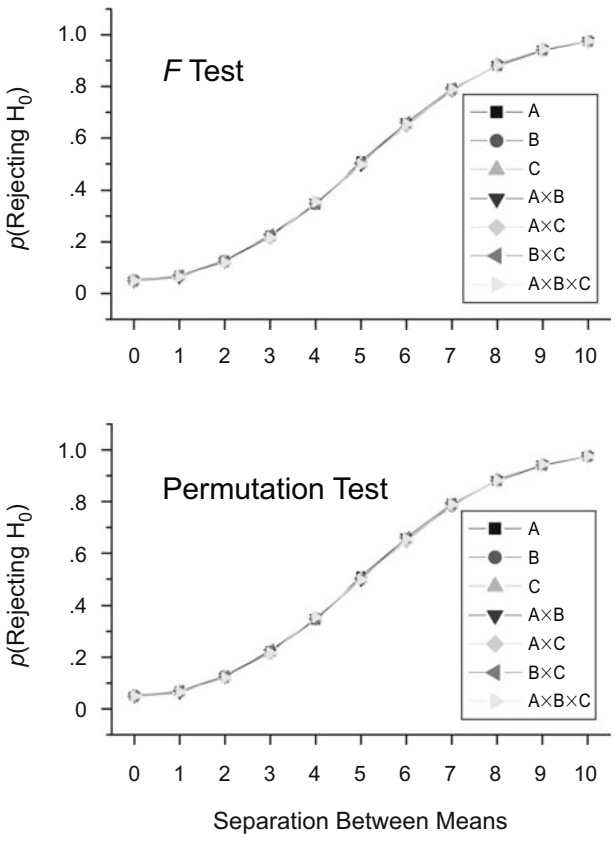

Figure 1. Comparison of the *F* and permutation tests using orthogonal contrasts. The curves labeled A to A×B×C refer to the factorial effects described in Table 2.

As is shown in Figure 1, when the null hypothesis was true, both the permutation test and the *F* test allowed the null hypothesis to be rejected about 5% of the time. The probability of detecting the effect increased systematically as the separation between the means was increased. There was no systematic advantage for the permutation test over the *F* test or vice versa, and there was no difference between main effects and interactions. Performance of all of the tests was identical. As Fisher (1966) anticipated, when the underlying error is normal, the *F* and permutation tests provided essentially the same results. We conclude, therefore, that by using a subset of the total number of outcomes, the contrast-based calculation does not cheat the permutation test. Rather, by organizing the comparison among the cells into nonoverlapping components, orthogonal contrasts also provide an efficient approximate permutation test that escapes the problems associated with approximate tests based on sampling.

Figure 1 shows results for 20 subjects. We have run the comparison using several numbers of subjects and several different repeated measures designs. With one qualification, the *F* and permutation tests are equivalent when the assumptions of the *F* test have been met. The qualification concerns the number of subjects: When *N* is small (e.g., 4), the number of possible outcomes for each factor is so small (e.g., $2^4 = 16$) that it is not possible for the permutation test to find a significant result, regardless of the difference between the means. The *F* test—based on normal distribution theory—can show a significant difference if the difference between means is large enough. The difference can be attributed to the gift of information provided by the Gaussian assumptions in the *F* test's model for error. Hence, the randomization test should not be used when *N* is less than about 10.

Independent observations. In the next Monte Carlo study, we generated data for eight cells using the same software tools as before. Each cell included data for 10 independent subjects. As before, we selected parameters so that the sensitivity of the tests for each effect (A, B, C, A×B, A×C, B×C, and A×B×C) would cover a full range—that is, so that the probability of rejecting the null hypothesis at the 5% level of significance would range from .05 with a null treatment to approximately 1 with a large separation between the means. Specifically, we set the mean, μ, of the normal generator at 0 and the standard deviation, σ, at 1. We increased the separation among the means for each factor and interaction from 0 to 5.5 in 11 equal steps.

When the null hypothesis was true, both the permutation test and the *F* test allowed the null hypothesis to be rejected about 5% of the time. The probability of detecting a true effect increased systematically as the separation between the means was increased. Provided that all of the variability among cells could be described by exactly one comparison, the *F* and randomization tests were

equivalent. When there were two or more effects (e.g., two main effects, two interactions, a main effect and an interaction, or some combination) the randomization test and the *F* test diverged. Specifically, the randomization test became increasingly too conservative as more true effects were added.

To illustrate the increasingly conservative behavior of the randomization test, we compared three situations: the case in which all systematic variability among the means fell on one effect (Factor A), on two contrasts (A and B×C), and on three contrasts (A, C, and B×C). Figure 2 summarizes the probability of rejecting the H_0 for Factor A as a function of the difference between means. The three increasing curves show, in the first case, that the randomization test behaved as expected: The probability of detecting a true effect (Factor A) increased as the separation of the means increased; although not shown in the figure, the probability of rejecting H_0 for other effects was stable at the nominal 5%. In the second and third cases, by contrast, the sensitivity curve was lower than that in the first case, and the probability of rejecting a true null hypothesis fell below the nominal 5%. As is shown in the figure, adding additional systematic variance to additional contrasts (A×C and C) systematically reduced the test's sensitivity to the true effect of Factor A. The three decreasing curves presented in Figure 2 show the probability of rejecting a true null hypothesis (a null A×C interaction) when systematic variance was added to Factors A, B×C, and C. In the demonstration illustrated in Figure 2, all of the effects were of the same magnitude; the conservative behavior of the test would be exaggerated further if the magnitude of the additional factor(s) were increased.

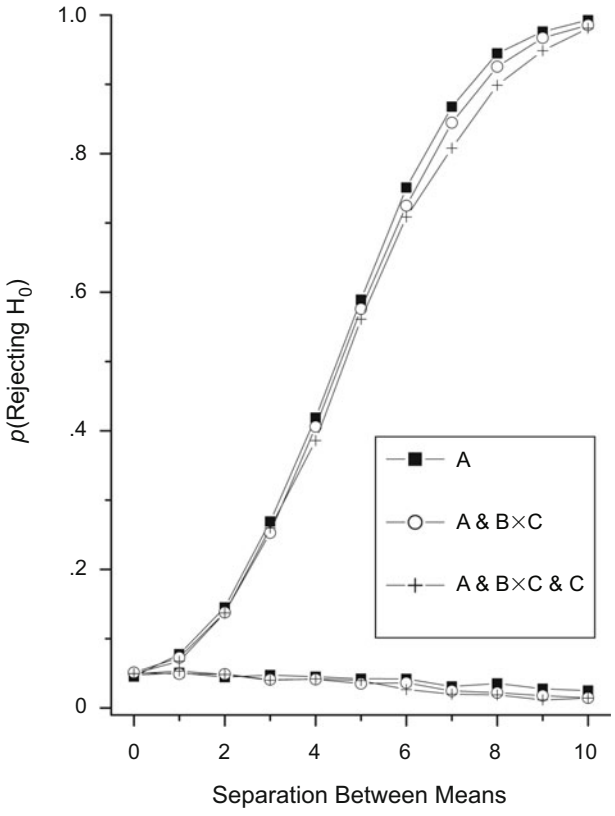

Figure 2. The independent-observation randomization test when systematic variance lies on one, two, or three contrasts. Contrasts A, B×C, and C are as described in Table 2. The increasing curves show the probability of detecting a true effect for Contrast A. The decreasing curves show the probability of rejecting a null effect. As additional effects are added, the test becomes more conservative.

The reason for the increasingly conservative behavior of the randomization test is easy to understand. For each comparison, the data are split into two equal cells. The randomization test compares the difference between the means of the cells with the variability within the two cells defined by the contrast. When only one of the factors or interactions is non-null, all of the variability within the two cells constitutes noise. When more than one factor or interaction is non-null, the variability in the two cells is inflated by the variability associated with the non-null factor(s). As a result, the randomization test becomes too conservative.

Summary and Conclusions

The randomization/permutation test does not depend on a known distribution for error and can be more sensitive to real effects than is the corresponding parametric test. In spite of its advantages, the randomization/permutation test is usually thought to be impractical for factorial designs, because the computational costs become unmanageable in larger designs. Gill's (2007) algorithm brings the computational costs of a permutation test to a practical level, but it was designed for a two-cell problem. A factorial design can be recast as a series of two-cell comparisons using orthogonal contrasts, but because orthogonal contrasts examine only a subset of the full number of permutations in a full permutation test, they might cheat the logic of the test. We examined the issue by comparing the orthogonal-contrast approach against a standard ANOVA under conditions that met the ANOVA's assumptions.

For repeated measures designs, the results showed that orthogonal contrasts work well; they are equivalent to normal-distribution tests when applied to data that meet the assumptions of the latter tests. For within-subjects (correlated-observation) designs, then, one can apply the randomization test with confidence.

For completely randomized (between-subjects) designs, by contrast, orthogonal contrasts produce a flawed test for significance. Depending on the number of non-null factors, it can become too conservative. Specifically, it becomes too conservative if systematic variance lies on more than one factor (i.e., contrast). In practice, the test's conservative nature implies that one can trust that a contrast shown to be a significant result is reliable. That said, because the between-subjects test is insensitive to true effects if systematic variance lies on more than one contrast, there is a serious risk that it might miss small but true effects.

In summary, the randomization test is a practical option for repeated measures factorial designs: Recasting the analysis into orthogonal contrasts does not cheat the logic of the permutation test. The permutation test follows

directly from the procedure in a comparative experiment (Ludbrook & Dudley, 1998), does not depend on a known distribution for error, and is sometimes more sensitive to real effects than are the corresponding parametric tests (e.g., Mewhort, 2005). Our results suggest that it can be used instead of an ANOVA for all of these reasons.

AUTHOR NOTE

The research was supported by a grant from the Natural Science and Engineering Research Council (NSERC) of Canada to the first author. The junior authors were supported by an NSERC summer research scholarship. We thank R. K. Jamieson for comments on an earlier draft. Computation time to confirm our implementation of Gill's (2007) algorithm was made available through an IBM Academic Initiative. We thank IBM Canada for their generosity in making computing resources available. Correspondence concerning this article should be addressed to D. J. K. Mewhort, Department of Psychology, Queen's University, Kingston, ON, K7L 3N6 Canada (e-mail: mewhortd@queensu.ca).

REFERENCES

BAKER, F. B., & COLLIER, R. O., JR. (1966). Some empirical results on variance ratios under permutation in the completely randomized design. *Journal of the American Statistical Association*, **61**, 813-820.

BOX, G. E. P., & ANDERSEN, S. L. (1955). Permutation theory in the derivation of robust criteria and the study of departures from assumption. *Journal of the Royal Statistical Society: Series B*, **17**, 1-34.

BRADLEY, J. V. (1968). *Distribution-free statistical tests.* Englewood Cliffs, NJ: Prentice Hall.

DONCASTER, C. P., & DAVEY, A. J. H. (2007). *Analysis of variance and covariance: How to choose and construct models for the life sciences.* Cambridge: Cambridge University Press.

EDGINGTON, E. S. (1995). *Randomization tests* (3rd ed.). New York: Dekker.

FISHER, R. A. (1966). *Design of experiments* (8th ed.). Edinburgh: Oliver & Boyd.

GILL, P. M. W. (2007). Efficient calculation of *p*-values in linear-statistic permutation significance tests. *Journal of Statistical Computation & Simulation*, **77**, 55-61. doi:10.1080/10629360500108053

HAYES, A. F. (2000). Randomization tests and the equality of variance assumption when comparing group means. *Animal Behaviour*, **59**, 653-656. doi:10.1006/anbe.1999.1366

HOEFFDING, W. (1952). The large-sample power of tests based on permutations of observations. *Annals of Mathematical Statistics*, **23**, 169-192.

KELLER-MCNULTY, S., & HIGGINS, J. J. (1987). Effect of tail weight and outliers on power and Type-I error of robust permutation tests for location. *Communications in Statistics: Simulation & Computation*, **16**, 17-35.

KEMPTHORNE, O., & DOERFLER, T. E. (1969). The behaviour of some significance tests under experimental randomization. *Biometrika*, **56**, 231-248.

LUDBROOK, J., & DUDLEY, H. (1998). Why permutation tests are superior to *t* and *F* tests in biomedical research. *American Statistician*, **52**, 127-132.

MANLY, B. F. J. (1997). *Randomization, bootstrap, and Monte Carlo methods in biology* (2nd ed.). London: Chapman Hall.

MECCERI, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, **105**, 156-166.

MEWHORT, D. J. K. (2005). A comparison of the randomization test with the *F* test when error is skewed. *Behavior Research Methods*, **37**, 425-435.

MEWHORT, D. J. K., KELLY, M., & JOHNS, B. T. (2009). Randomization tests and the unequal-*N*/unequal-variance problem. *Behavior Research Methods*, **41**, 664-667. doi:10.3758/BRM.41.3.664

PAGANO, M., & TRITCHLER, D. (1983). On obtaining permutation distributions in polynomial time. *Journal of the American Statistical Association*, **78**, 435-440.

PITMAN, E. J. G. (1937). Significance tests which may be applied to samples from any population. *Journal of the Royal Statistical Society Supplement*, **4**, 119-130.

PRESS, W. H., TEUKOLSKY, S. A., VETTERLING, W. T., & FLANNERY, B. P. (1992). *Numerical recipes in FORTRAN: The art of scientific computing* (2nd ed.). New York: Cambridge University Press.

ROMANO, J. P. (1990). On the behavior of randomization tests without a group invariance assumption. *Journal of the American Statistical Association*, **85**, 686-692.

STIGLER, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Cambridge, MA: Harvard University Press.

THOMPSON, D. W. (1942). *On growth and form* (2nd ed.). Cambridge: Cambridge University Press.

NOTES

1. In general, the number of combinations of N objects taken r objects at a time, $C(N, r)$, is $C(N, r) = N! / [(N-r)! \times r!]$.

2. Mewhort et al. (2009) confirmed Hayes's (2000) results using full-enumeration permutation tests.

3. In an ANOVA, of course, the scores labeled m_i are means of cells, and calculation of the sum of squares includes weights based on the number of observations summarized by each mean.

SUPPLEMENTAL MATERIALS

Code to execute the routines described here, to be used with the f90 compiler, and a short demo to illustrate the code's use may be downloaded from http://brm.psychonomic-journals.org/content/supplemental.

(Manuscript received October 13, 2009;
revision accepted for publication December 6, 2009.)