*Journal of Statistical Computation and Simulation*,
2003, Vol. 73(2), pp. 85–113

# PERMUTATION TESTS FOR MULTI-FACTORIAL ANALYSIS OF VARIANCE

MARTI J. ANDERSON[a,*] and CAJO J. F. TER BRAAK[b,†]

[a]*Centre for Research on Ecological Impacts of Coastal Cities, Marine Ecology Laboratories, A11, University of Sydney, NSW, 2006, Australia;* [b]*Biometris, Wageningen University and Research Centre, Box 100, 6700 AC, Wageningen, The Netherlands*

*(Received 1 February 2001; In final form 13 August 2002)*

Several permutation strategies are often possible for tests of individual terms in analysis-of-variance (ANOVA) designs. These include restricted permutations, permutation of whole groups of units, permutation of some form of residuals or some combination of these. It is unclear, especially for complex designs involving random factors, mixed models or nested hierarchies, just which permutation strategy should be used for any particular test. The purpose of this paper is two-fold: (i) we provide a guideline for constructing an exact permutation strategy, where possible, for any individual term in any ANOVA design; and (ii) we provide results of Monte Carlo simulations to compare the level accuracy and power of different permutation strategies in two-way ANOVA, including random and mixed models, nested hierarchies and tests of interaction terms. Simulation results showed that permutation of residuals under a reduced model generally had greater power than the exact test or alternative approximate permutation methods (such as permutation of raw data). In several cases, restricted permutations, in particular, suffered more than other procedures, in terms of loss of power, challenging the conventional wisdom of using this approach. Our simulations also demonstrated that the choice of correct exchangeable units under the null hypothesis, in accordance with the guideline we provide, is essential for any permutation test, whether it be an exact test or an approximate test. For reference, we also provide appropriate permutation strategies for individual terms in any two-way or three-way ANOVA for the exact test (where possible) and for the approximate test using permutation of residuals.

*Keywords*: ANOVA; Experimental design; Fixed and random factors; Hierarchical designs; Mixed models; Non-parametric; Randomization tests; Resampling methods

## 1 INTRODUCTION

Analysis of variance (ANOVA) is a statistical tool used extensively in the biological, psychological, medical, ecological and environmental sciences. Traditional parametric ANOVA is generally quite robust to violation of its assumption of normally distributed errors (Cochran, 1947; Scheffé, 1959; Snedecor and Cochran, 1967). The assumption of normality is, however, unreasonable for many kinds of data (*e.g.*, ecological variables showing a mean–variance relationship: Taylor, 1961; Gaston and McArdle, 1994). In the analysis of univariate

* Corresponding author. Present address: Department of Statistics, University of Auckland, Private Bag 92019, Auckland, New Zealand. Tel.: 64-9-373-7599 x 5052; Fax: 64-9-373-7018; E-mail: mja@stat.auckland.ac.nz
† Tel.: 31-31747-6929; Fax: 31-31741-8094; E-mail: c.j.f.terbraak@plant.wag-ur.nl

ISSN 0094-9655 print; ISSN 1563-5163 online
DOI: 10.1080/0094965021000015558

data, one can often avoid the problem of non-normal data by finding a suitable transformation or by using generalised linear models (GLMs) or generalised linear mixed models (GLMMs), where a non-normal error structure may be specified explicitly (*e.g.*, Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989; Breslow and Clayton, 1993).

An alternative approach is to use permutation tests, where errors are not assumed to be normally distributed, yet exact tests are achieved (Fisher, 1935; Pitman, 1937; Scheffé, 1959). By "exact", we mean that the type I error of the test is exactly equal to the *a priori* chosen significance level for the test. The use of permutation tests has received renewed attention in recent years with the advent of much faster and more accessible computer power (Crowley, 1992; Edgington, 1995; Manly, 1997). In general, for an exact test by permutation, the reference distribution of a relevant test statistic under the null hypothesis is constructed by calculating its value for all possible re-orderings (permutations) of the observations (or a large random subset of such re-orderings, Hope, 1968). A *P*-value is then calculated as the proportion of the values of the statistic obtained under permutation that are equal to or more extreme than the observed value.

Permutation tests are especially useful and relevant for multivariate analysis, where distributional assumptions are even more difficult to fulfill (*e.g.*, Mardia, 1971; Olson, 1974; Johnson and Field, 1993). Also, in many situations where there are multiple responses, trying to model all of them with the same kind of error structure (*e.g.*, by using GLMs as in longitudinal analysis, see Liang and Zeger, 1986; Zeger and Liang, 1992) will generally be inappropriate. Thus, there have been many tests proposed for the comparison of *a priori* groups of multivariate data that rely on permutation of the observation vectors (*e.g.*, Mantel and Valand, 1970; Hubert and Schultz, 1976; Mielke *et al.*, 1976; Smith *et al.*, 1990; Excoffier *et al.*, 1992; Clarke, 1993; Edgington, 1995; Pillar and Orlóci, 1996; Gower and Krzanowski, 1999; McArdle and Anderson, 2001; Anderson, 2001). Permutation testing is an important feature of the computer program CANOCO (ter Braak and Šmilauer, 1998) for canonical ordination of ecological data and of the computer program NPMANOVA (Anderson, 2001) for non-parametric multivariate analysis based on a distance matrix. Motivation for this research arose from our desire to extend the facilities of these programs for permutation tests in any complex ANOVA experimental design.

There is general agreement concerning the appropriate procedure to achieve an exact permutation distribution for one-way ANOVA (*e.g.*, Edgington, 1995; Manly, 1997; Good, 2000), for either univariate or multivariate sets of observations. However, as soon as an experimental design includes more than one factor, there is a divergence of opinion concerning appropriate permutation methods that may be used for particular tests. For example, to obtain a test of an interaction term by permutation, ter Braak (1992) suggested permutation of residuals under a full model, Manly (1997) and Gonzalez and Manly (1998) suggested unrestricted permutation of the raw data, while Edgington (1995) contended that a valid test for interaction cannot be done using permutations.

Some extensive empirical and theoretical work has been done to investigate appropriate strategies for an approximate permutation test of an individual term in multiple linear regression, for which no exact test exists (Anderson and Legendre, 1999; Anderson and Robinson, 2001). It was found that permutation of residuals under a reduced model (Freedman and Lane, 1983) performed the best in the widest set of circumstances (Anderson and Legendre, 1999) and also comes the closest to the conceptually exact test (Anderson and Robinson, 2001).

Since ANOVA with fixed factors is simply a special case of multiple regression, but with categorical predictor variables, one could simply be satisfied that the results obtained for multiple regression apply equally well for individual terms in ANOVA designs. However, ANOVA designs introduce two important additions (or strategies) to the possibilities available for multiple regression: (i) restricted permutations (*i.e.*, allowing permutations to occur only within levels of other factors) and (ii) permutation of units other than the observations (*e.g.*, permuting the units induced by a nested factor in the test of a higher ranked factor). The second strategy may be pertinent in designs with random factors. In many ANOVA designs, an exact test for an individual factor can be constructed using one or other or both of these additional strategies. It is virtually entirely unknown, however, just how such tests compare with the various approximate tests in terms of their power for different kinds of designs. Thus, it is unclear, for any particular term in a complex ANOVA design, which combination of (i) choice of permutable units, (ii) use of restricted permutation and/or (iii) permutation of raw data or residuals will provide the most powerful or most appropriate strategy.

In order to construct an appropriate permutation distribution for any term in a complex ANOVA design, it is necessary to clarify several important issues, including: (a) Should raw data be permuted, or some form of residuals? (b) Which units should be permuted: individual observation units or some larger units, such as levels of a nested factor in a test of the higher ranked factor of a hierarchy? (c) When should permutations be restricted to occur within levels of other factors? (d) How may interaction terms be tested using permutations? (e) Which tests are exact and which are approximate? (f) What test-statistic should be used for the test?

The purpose of this paper is to provide precisely such a clarification of these issues. First, we give a guideline for the construction of an exact permutation test, where possible, in any given situation for any factor in a complex ANOVA design. This guideline sheds light on how the design of the experiment or survey, including whether factors are fixed or random, nested or crossed with other factors, determines how valid permutation tests are to be done. No such general guideline, including the possibility for random effects, mixed models and nested hierarchies, currently exists in such a succinct form in the literature, to our knowledge.

Second, we give results from specific sets of empirical Monte Carlo simulations done to compare the level accuracy and power of several possible permutation strategies (including the exact test, where possible) for individual terms in several models of two-way ANOVA. These were chosen so that, from them, important general statements and recommendations for multi-way ANOVA could be made. These are the most extensive simulations for permutation tests in complex ANOVA designs provided to date, to our knowledge, as well as being the only ones to include nested hierarchies and mixed or random effects models. Although we restrict our attention here to univariate ANOVA designs, the results will hold, in general, for the analogous permutation tests on multivariate response vectors.

## 2 CONSTRUCTING AN EXACT TEST

We provide a guideline for constructing an exact permutation test for individual terms in a multi-factorial ANOVA. We first need to delimit the class of designs and introduce some definitions.

We limit our discussion to equi-replicated orthogonal ANOVA designs, following Nelder (1964a; 1964b). By orthogonal designs, we mean effects are independent, whether they be crossed or nested. In considering expected mean squares for terms in any ANOVA model (and for simulations) we also rely on the usual summation restrictions for fixed effects, *i.e.*, that the sum individual treatment effects across all levels of a fixed factor equals zero. For alternative formulations of expected mean squares (*i.e.*, without summation restrictions or for unbalanced designs) consult Searle *et al.* (1992). We consider that the same general guideline below should be followed for permutation tests for unbalanced designs, but do not provide any particular details of this here for particular cases of unbalance.

The null hypothesis for a test of any particular factor, whose categorical levels we may generally call treatments, is that there is no treatment effect. More generally, the null hypothesis is that the errors associated with units in different treatments have the same distribution. In the case of more complex designs, the null hypothesis is conditional: given the other terms in the model (such as main effects in the test of an interaction), the errors associated with units in different treatments have the same distribution. We consider that exchangeability of units for permutation tests gains its validity by virtue of the statement of similar error distributions under the null hypothesis (*e.g.*, Kempthorne, 1952).

Although permutation tests avoid the assumption of normality, they still assume exchangeability of relevant units under the null hypothesis. Exchangeability can be ensured through the random allocation of treatments to units in experimental design (*e.g.*, Fisher, 1935; Kempthorne, 1955; Scheffé, 1959) or must be assumed for observational studies (*e.g.*, Kempthorne, 1966). The assumption of exchangeability is tantamount to the assumption that errors are independent and identically distributed ("i.i.d."). Note that this does not avoid the assumption of homogeneity of error variances (*e.g.*, Boik, 1987; Hayes, 1996).

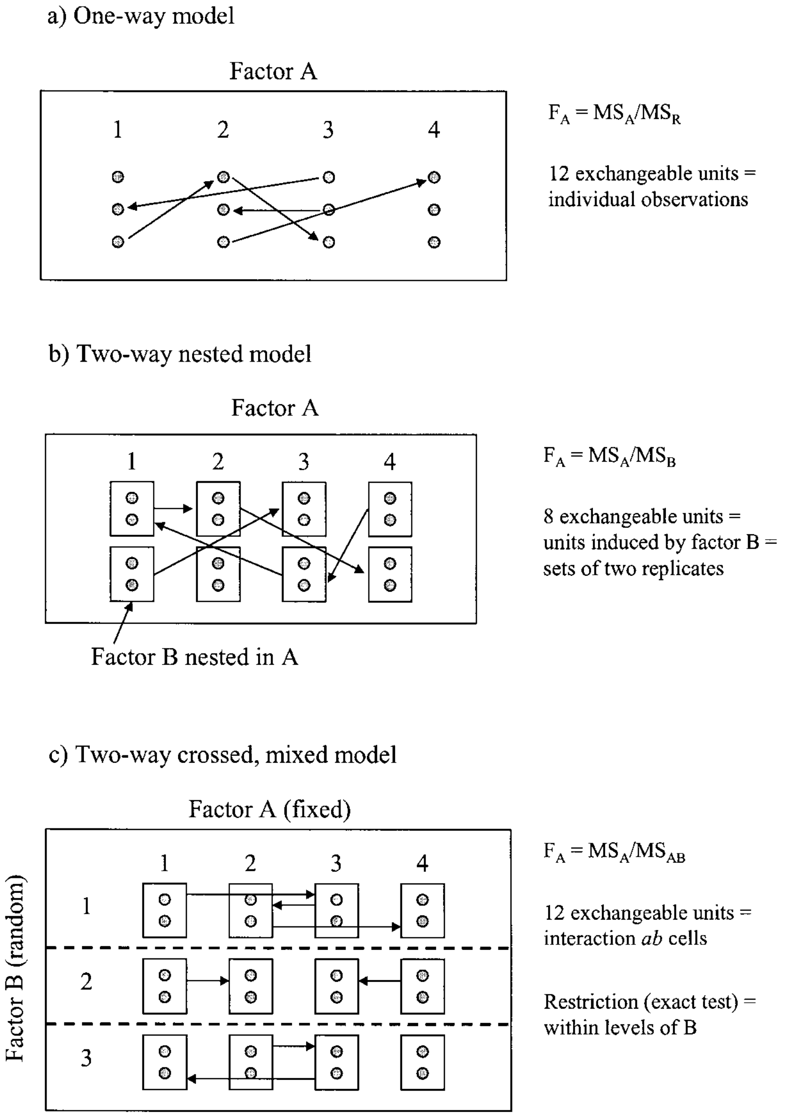
Next, we provide definitions for the "order" of a term and what is meant by the "exchangeable units" for a test. By the "order" of a term, we mean the following: main effects are of first order, a two-way interaction term is of second order, etc. A term that is nested in another term has an order one more than the order of the term within which it is nested. "Exchangeable units" are identified in the denominator mean square of an $F$-ratio for any particular term. Consider Figure 1. Where the residual mean square is the mean square in the denominator for a test, this indicates that individual units (observations) are exchangeable under the null hypothesis (Fig. 1a). Where the denominator mean square of an $F$-ratio is the mean square of, for example, a nested term, then the exchangeable units consist of the units induced by the nested term. For example, in Figure lb there are eight ($2 \times 4$) units induced by the nested term B, each unit consisting of two replicates. Where the denominator of an $F$-ratio is the mean square of, for example, an interaction term, then the units consist of the blocks of cells identified by that interaction (*e.g.*, if one factor has four levels and a second factor has three levels, then the interaction identifies $4 \times 3 = 12$ cells which are the exchangeable units, Fig. 1c). A term whose mean-square appears as the denominator in the $F$-ratio of the test of another term in the model identifies the exchangeable units for that test. The exchangeable units have the same distribution under the null hypothesis.

We provide the following general guideline:

*An exact permutation test for any term in an ANOVA model is achieved by permuting the exchangeable units identified by the denominator mean-square of the F-ratio and restricting permutations to occur within the levels of terms of either smaller order or of the same order as the term being tested.*

The guideline yields a permutation test based on restricted permutations of the exchangeable units. The guideline ensures that, under the null hypothesis, the likelihood of the data is invariant under these permutations, and consequently the test is exact. It encapsulates the idea that all unknown parameters in the model not being tested should be kept constant under permutation in order to isolate the test on the factor of interest. In the context of equi-replicated orthogonal ANOVA designs we identify two different aspects to guarantee the invariance: (i) components of variation that may contribute to variability in the term being tested, and (ii) other non-zero terms in the linear ANOVA model. The first of these is addressed by considering the appropriate exchangeable units for the permutation test. The second is addressed by restricting the permutations of those units within levels of other factors.

The construction of the $F$-ratio itself provides the necessary information concerning the exchangeable units by identifying the components of variation that contribute to variability in the term being tested. The ratio is constructed by reference to the expected mean squares



a) One-way model

$F_A = MS_A/MS_R$

12 exchangeable units = individual observations

b) Two-way nested model

$F_A = MS_A/MS_B$

8 exchangeable units = units induced by factor B = sets of two replicates

c) Two-way crossed, mixed model

$F_A = MS_A/MS_{AB}$

12 exchangeable units = interaction *ab* cells

Restriction (exact test) = within levels of B

FIGURE 1 Diagram of permutation strategies for exact tests: a) test of A in a one-way ANOVA model; b) test of factor A, the higher-ranked factor in a two-way nested design, with random factor B nested within factor A and c) test of factor A, a fixed factor, in a two-way crossed mixed model design (where factor B is random).

of individual terms. If components of variation that are not being tested are present in the numerator's expected mean square, they must appear in the expected mean square of the denominator in order to isolate the component of variation of interest for the test using the $F$-ratio. The denominator mean square thus identifies the components of variation that contribute to the variability in the term being tested and, by this, the exchangeable units to be used in the permutation test.

Some terms in the model may not contribute to variability in the term of interest but, if non-zero, would be confounded with ("mixed with") the term under test. Restricted permutation within levels of these factors avoids this by ensuring that the sizes of their effects in the model remain constant under permutation. For example, in a two-way ANOVA without interaction, permutations for a test of one factor should be restricted to occur within levels of the other factor for an exact test.

For a discussion of restricted permutation for many kinds of complex ANOVA designs in psychology, see Edgington (1995). For a discussion of restricted permutation in the context of multiple regression, where predictor variables take several fixed repeated values, see Brown and Maritz (1982).

The exact permutation test will be unique for any particular term in an ANOVA model. An exact permutation test, however, may not always exist, or may not have enough possible permutations to enable reasonable power for detecting alternatives. This leads us to consider approximate tests.

## 3 APPROXIMATE TESTS

In situations where the strategy required for an exact test is not possible (*i.e.*, where the restrictions required leave no possible permutations), then no exact test exists, but an approximate permutation test may be used. This occurs, for example, in the case of interaction terms. Except in some special cases (Welch, 1990), it is not possible to construct an exact permutation test for an interaction using the $F$-statistic, as restricting permutations within levels of the main effects leaves no alternative possible permutations: the $F$-ratio obtained with the observed data is the only possibility. (For alternative approaches using other test statistics, see Pesarin, 2001).

Another situation where an exact test is not feasible occurs when the combination of restrictions and exchangeable units results in there being too few possible permutations to obtain a reasonable test. For example, if there are only two levels of a nested factor (B) in each of two levels of a higher ranked (but lower order) factor (A), this leaves only 6 possible permutations for the test of A, 3 of which would give unique values for the test statistic and one of which is the observed value. This is clearly not sufficient to provide a reasonable test. For complex designs, there are essentially three different approaches for an approximate test: permutation of residuals under a reduced model (Still and White, 1981; Freedman and Lane, 1983), under the full model (ter Braak, 1992) or unrestricted permutation of raw data (Manly, 1997; Gonzalez and Manly, 1998).

For a comparison of these approaches in the more general context of multiple regression, see Anderson and Legendre (1999) and Anderson and Robinson (2001). Permutation under the reduced model comes the closest to a conceptually exact test (Anderson and Robinson, 2001). For this reason, we here generally restrict our attention concerning permutation of residuals to reduced-model residuals, although full-model residuals can also be used in these situations, and would be expected to give highly comparable results (Anderson and Legendre, 1999). We also add that the problems raised by Kennedy and Cade (1996) for the method of permutation of raw data, which concerned the effects of outliers in nuisance variables for multiple regression, cannot occur in the context of ANOVA, where factors are categorical and thus do not contain outliers (Anderson and Legendre, 1999).

In what follows, we describe and compare the possible permutation strategies (approximate and exact, where possible) available for tests of individual terms in two-way ANOVA for nested, crossed fixed and crossed mixed models. These demonstrate the principles involved in constructing an exact test, which we then extend to three-way models. In particular, we give results of empirical simulations that compare the various possible permutation strategies, demonstrating which of them provides the greatest empirical power for tests.

## 4 NESTED DESIGN

Consider a nested (hierarchical) ANOVA design with the following linear model:

$$y_{ijk} = \mu + A_i + B(A)_{j(i)} + \varepsilon_{ijk}$$

where $\mu$ is the unknown population mean, $B(A)_{j(i)}$ is the effect of the *j*th level of factor B within the *i*th treatment level of factor A, symbolised by $A_i$, and $\varepsilon_{ijk}$ is the unknown error associated with observation $y_{ijk}$. The number of levels in factors A and B will be designated by $a$ and $b$, respectively, and the number of replications per AB combination by $n$. We consider A as a fixed factor, but for this model the same discussion will apply whether A is fixed or random. Factor B, being nested, will, for present purposes, always be considered random. Note that although factor B has $b$ levels, it introduces a total of $a \times b$ effects, denoted by $B(A)_{j(i)}$, into the model. For the permutation test, we assume only that the $\varepsilon$'s are independent and identically distributed (i.i.d.), but not that they are (necessarily) normal. Throughout, we shall denote sums of squares, mean squares and $F$-ratios for a particular term (say for factor B) as $\mathrm{SS_B}$, $\mathrm{MS_B}$ and $F_\mathrm{B}$, respectively. Also, the subscript "R" shall indicate the residual, while the subscript "T" shall indicate the total. For the above model, a test of factor B is provided by the statistic $F_\mathrm{B} = \mathrm{MS_B}/\mathrm{MS_R}$ and a test of factor A is provided by $F_\mathrm{A} = \mathrm{MS_A}/\mathrm{MS_B}$ (*e.g.* Kempthorne, 1952; Scheffé, 1959; Winer *et al.*, 1991).

### 4.1 Test of the Nested Factor, B

The null hypothesis for an exact test of the effect of factor B can be phrased: given the presence of A, about which we make no assumption, the effect of B within A is not different from zero. Thus, the permutations of observations ($y$) are done across the levels of B, but these are restricted to occur within each of the levels of A. This means that whether or not factor A has an effect, we can test for significant variability due to factor B. This strategy is consistent with the guideline.

This provides an exact test because (i) $\mathrm{SS_T}$ stays constant across all permutations, (ii) restricting permutations within levels of A means that $\mathrm{SS_A}$ remains constant across all permutations, and (iii) $\mathrm{SS_T} = \mathrm{SS_R} + \mathrm{SS_B} + \mathrm{SS_A}$. Thus, permutation mixes (exchanges) variation only between $\mathrm{SS_B}$ and $\mathrm{SS_R}$, which is exact for the test of $F_\mathrm{B} = \mathrm{MS_B}/\mathrm{MS_R}$.

For an approximate test, one can calculate and permute the reduced-model residuals $r_{ijk} = y_{ijk} - \bar{y}_{i..}$, where $\bar{y}_{i..}$ is the mean of the observations in level $i$ of factor A. This amounts to estimating and "removing" the effect of A by subtracting the mean of the appropriate level of A from each observed value. Permuting these residuals will yield an approximate test.

An alternative approximate test of factor B is provided by simply permuting all observations without restriction. This mixes variability among all terms in the model, including variability due to factor A (*i.e.*, $\mathrm{SS_A}$), which is not held constant under permutation. However, as this "mixing" of $\mathrm{SS_A}$ under permutation impinges on each of $\mathrm{SS_B}$ and $\mathrm{SS_R}$, the ratio $F_\mathrm{B} = \mathrm{MS_B}/\mathrm{MS_R}$, on average, is unaffected, giving a reasonable approximate test (Gonzalez and Manly, 1998; Anderson and Robinson, 2001).

Simulations were done to demonstrate the type I error and relative power of the above proposed strategies of permutation to detect variability due to factor B. Data were simulated according to the above model with $a = 4, b = 5$ and the sample size within each cell ($n$) was set at $n = \{2, 5, \text{ or } 10\}$, where the effect of factor A was non-zero: $A_i = \{33.3, 33.3, 33.3, -100\}$. The additive effect of levels of the random factor B were chosen randomly from a normal distribution with mean zero and gradual increases in standard deviation. Note that $b = 5$ different values of $B(A)_{j(i)}$ were chosen within each level of factor A.

That is, a total of $ab = 20$ different effects were chosen randomly, with five assigned to each of the $a = 4$ levels of factor A. This is the meaning of a nested effect: its levels are specific and peculiar to each level of factor A. Although other error distributions could be used in modeling levels of random effects in simulations, we used the normal distribution for this, as it allowed the most fundamental issues in ANOVA designs to be investigated in the first instance.

For this and for all subsequent simulations, random errors were chosen from each of four different distributions: (i) a standard normal distribution, (ii) a uniform distribution on the interval (1, 10), (iii) a lognormal distribution, which consisted of exponentiating random values from a standard normal distribution, or (iv) an exponential (1) distribution whose values were then cubed. The latter distribution was used to simulate data that were radically non-normal (*e.g.*, Manly, 1997).

For each scenario, a total of 1000 sets of simulated data were produced and tested with each permutational approach using 999 random permutations. The test statistic used in all simulations was the $F$-ratio with a denominator appropriate for the design. For each simulated data set, a $P$-value for the normal-theory $F$-test, using the tables, was also obtained for comparison. In this and in subsequent simulations, pair-wise comparisons of the power (as the number of rejections of the null hypothesis) of different methods were done using Wilcoxon's signed-ranks tests.

Type I errors for all test procedures did not differ significantly from one another or from the nominal significance level of 0.05. These simulations also showed that all test procedures had virtually identical power when errors were normal (Fig. 2a, Tab. I), whereas the traditional $F$-test was most powerful in the case of uniform errors (Tab. I). With lognormal or exponential cubed errors, however, the exact test (permuting data within levels of A) had the greatest power, followed by permutation of residuals (Fig. 2b, Tab. I). Permutation of residuals approached the power of the exact test with increases in $n$. Permutation of raw data without restriction and the normal-theory $F$-test had significantly less power than permutation of residuals or the exact test (Tab. I).

## 4.2 Test of the Higher Ranked Factor, A

The null hypothesis for an exact test of the effect of factor A can be phrased: given the variability across levels of B, about which we make no assumption, the effect of A is not different from zero. The expected mean square for factor A contains a component of variation attributable to the nested factor B. Therefore, $F_A = MS_A/MS_B$ whose denominator indicates the exchangeable units for the test as levels of B. Rather than permuting individual replicates, replicates within each level of B are kept together as a unit and these units are permuted across levels of A for the exact test (Fig. 1b).

This provides an exact test because (i) $SS_T$ stays constant across all permutations, (ii) permuting whole levels of B as units means that $SS_R$ remains constant across all permutations, and (iii) $SS_T = SS_R + SS_B + SS_A$. Thus, permutation mixes (exchanges) variation only between $SS_A$ and $SS_B$, which is exact for the test of $F_A = MS_A/MS_B$.

Note that if it had been determined from a previous test that variability due to factor B was not significant with a $P$-value that was sufficiently large, one would have the option of pooling factor B with the residual (*i.e.*, ignoring factor B and permuting all observations randomly) and doing a one-way test of factor A, as would be the case for the normal-theory test. The decision "to pool or not to pool" in such situations is an important one, and a determination of what constitutes a $P$-value that is "sufficiently large" may not be straightforward. This issue is the same for permutation tests as for the normal-theory test and won't be discussed further here (Winer *et al.*, 1991).

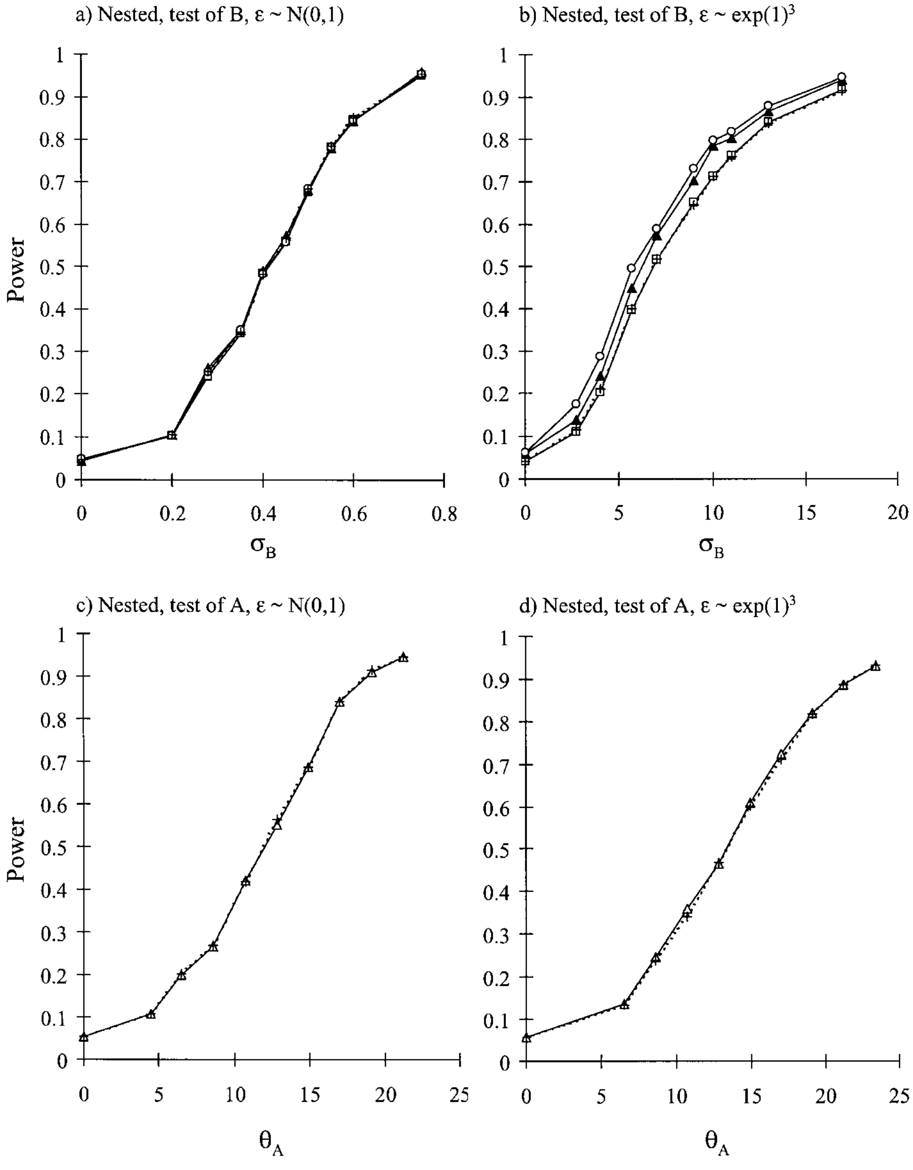


FIGURE 2 Power curves for permutation tests and the normal-theory $F$-test for a two-way nested design:
□ = Y (permutation of raw data),
▲ = R (permutation of residuals),
○ = Y(A) (permutation of raw data within levels of A),
△ = Yab (permutation of raw data as *ab* units),
+ = normal-theory $F$-test.

Permutation of residuals under a reduced model for a test of A might be done using the residuals $r_{ijk} = y_{ijk} - \bar{y}_{ij.} + \bar{y}_{i..} - \bar{y}_{...}$, where $\bar{y}_{ij.}$ is the mean for the *ij*th cell, $\bar{y}_{i..}$ is the mean of the observations in level $i$ of factor A and $\bar{y}_{...}$ is the estimated overall mean. This essentially removes the effect of the nested factor, B, if any, independently of A. Residuals under the full model would be $r_{ijk} = y_{ijk} - \bar{y}_{ij.}$ and one might also imagine that an approximate test could be provided by permutation of raw data.

TABLE I Summary of results regarding power of permutation tests, based on pair-wise Wilcoxon's signed-ranks tests for $n = 10$ sets of 1000 simulations used to produce power curves for each of four different error distributions. The permutation methods are ordered according to the greatest number of rejections (greatest overall power) in simulations: " = " indicates not significantly different ($P > 0.05$), while inequalities indicate significant differences ($P < 0.05$). Symbols in bold indicate the exact permutation test procedure in each case, where possible.

| | $\varepsilon \sim N(0,1)$ | $\varepsilon \sim \mathit{Uniform}(1,10)$ | $\varepsilon \sim \mathit{Lognormal} \sim \exp\{N(0,1)\}$ | $\varepsilon \sim \exp(1)^3$ |
|---|---|---|---|---|
| Nested | | | | |
| Test of B: | R = F = **Y(A)** = Y | F > R = **Y(A)** > Y | **Y(A)** = R > F > Y | **Y(A)** > R > Y = F |
| Test of A: | F = **Yab** = Y | F > Y = **Yab** | F = **Yab** = Y | Y* > **Yab** > F |
| Crossed, Fixed | | | | |
| Test of AB: | F > R = Y | F > R = Y | R > F > Y | R > Y > F |
| Test of A: | F = R = Y > **Y(B)** | F > R = Y > **Y(B)** | R > F > Y > **Y(B)** | R > Y = F = **Y(B)** |
| Crossed, Mixed, No Interaction | | | | |
| Test of A:<br>(fixed factor) | F > R = Rab = Y = **Y(B)ab** > Y(B) | F > R > Y = **Y(B)ab** = Rab > Y(B) | R > **Y(B)ab** > Rab = F = Y > Y(B) | R > **Y(B)ab** = Y > Rab > F > Y(B) |
| Crossed, Mixed, Interaction Present | | | | |
| Test of A:<br>(fixed factor) | F > R = Y = **Y(B)ab** = Rab > Y(B) | F = Y = R = **Y(B)ab** > Rab > Y(B) | Y = F = R > **Y(B)ab** = Rab > Y(B) | Y* > R* > **Y(B)ab** > Rab > F > Y(B) |

*Type I error was inflated for these tests – see Tables II and III.
Y = permutation of raw data; R = permutation of residuals;
Y(A) = permutation of raw data restricted within levels of A;
Y(B) = permutation of raw data restricted within levels of B;
Yab = permutation of raw data, as *ab* units (induced by nested factor B);
Rab = permutation of residuals, as *ab* units;
Y(B)ab = permutation of raw data, as *ab* units, restricted within levels of B;
F = normal-theory *F*-test (tabled values).

Simulations were done to compare the relative type I error and power of the exact test, each of the approximate tests (unrestricted permutation of raw data, permutation of residuals under the reduced or under the full model) and the normal-theory $F$-test. Data were simulated as described in Section 4.1, but this time variability due to B was kept constant (and large) while the fixed effect of factor A was gradually increased. For each level of factor A, separate levels of B were chosen randomly from a normal distribution with mean zero and standard deviation = 20.0.

There are an infinite number of possible values the fixed effects ($A_i$, $i = 1, \dots, a$) may take in an investigation of power. This is the nature of ANOVA: it is overparameterised (*e.g.*, Scheffé, 1959). However, an investigation of power (for a given sample size) can be indexed uniquely by the measure of effect size:

$$f_{\mathrm{A}} = \frac{\sqrt{\sum_{i=1}^{a}(A_i - \bar{A})^2/a}}{\sigma_\varepsilon}$$

with $\bar{A} = \sum_{i=1}^{a} A_i/a$ (*e.g.*, Winer *et al.*, 1991). For simplicity, as $\sigma_\varepsilon$ is a constant for any set of simulations, we shall use $\theta_{\mathrm{A}} = \sigma_\varepsilon \cdot f_{\mathrm{A}}$ as a measure of effect size. Note that $\theta_{\mathrm{A}}$ for a fixed factor A is analogous to $\sigma_{\mathrm{B}}$ for a random factor B. This has the consequence that different ranges of $\theta_{\mathrm{A}}$ (or $\theta_{\mathrm{AB}}$ or $\sigma_{\mathrm{B}}$) are required to obtain power curves for different error distributions used in simulations.

When data were simulated with normal, uniform or lognormal errors, all tests had type I error close to the nominal significance level (Tab. II). However, when data were simulated with the highly skewed exponential cubed errors the normal-theory $F$-test had type I error significantly lower than the significance level of 0.05 (Tab. II). Furthermore, when levels of the nested factor were drawn from a distribution with large variance, all of the approximate methods, including permutation of raw data and permutation of any form of residuals, had significantly inflated type I error (Tab. II). Thus, the use of any of these approximate tests is unwise for the test of a higher ranked factor in a nested design.

In comparisons of power, we only included those methods that did not suffer from inflated type I error: namely, the normal-theory $F$-test and the exact permutation test. There was not a striking visual difference in power between these two methods evidenced in the power plots (Fig. 2c,d). However, Wilcoxon's pair-wise tests indicated that the exact permutation test had significantly greater power than the normal-theory $F$-test for the situation with exponential cubed errors, while there was no significant difference in the power of these two tests for normal or lognormal errors (Tab. I), and for uniform errors, the $F$-test was most powerful.

Clearly, the exact permutation test should be used for any test of a higher ranked factor in a nested hierarchical design where non-normality is an issue. None of the approximate permutation tests provides an alternative that can be trusted to maintain type I error for this situation.

Researchers would be well-advised to increase replication of the levels of nested random factors, as the lack of such replication may severely sacrifice the power of the permutation test, as for the normal-theory test. For example, where A has 2 levels, there must be at least 4 levels of the nested factor B in order to obtain an exact permutation test for the effect of A whose $P$-value can exceed a significance level of 0.05. In other words, 3 levels of the nested factor B in each of 2 levels of A gives only 10 possible permutations yielding a unique value for $F$, while 4 levels of B would give 35 possible permutations. The lack of a sufficient number of possible permutations (due to only a small number of exchangeable units) is potentially a serious drawback and can only be remedied by increasing the number of levels of the nested factor, since none of the approximate methods of permutation will do. This is

TABLE II Proportion of rejections of the true null hypothesis (type I error) out of 1000 simulations with significance level = 0.05 for the test of a higher ranked factor (A) in a two-way nested ANOVA design, with $F_A = MS_A/MS_B$. Factor B is random and nested within levels of A. Levels of B were chosen randomly from a normal distribution with mean zero and standard deviation $\sigma_B$. Values that lie outside the 95% confidence interval for type I error (which has a binomial ($n$, $p$) distribution with $n = 1000$ simulations and $p = 0.05$, *i.e.*, 0.036–0.064) are indicated with an asterisk.

| *a* | *b* | *n* | $\sigma_B$ | *Yab* | *Y* | *R(Reduced)* | *R(Full)* | *F* |
|---|---|---|---|---|---|---|---|---|
| $\varepsilon \sim N(0, 1)$ | | | | | | | | |
| 4 | 5 | 2 | 1.0 | 0.060 | 0.057 | 0.054 | 0.059 | 0.058 |
| 4 | 5 | 5 | 1.0 | 0.045 | 0.045 | 0.045 | 0.047 | 0.045 |
| 4 | 5 | 10 | 1.0 | 0.047 | 0.048 | 0.050 | 0.051 | 0.052 |
| 4 | 5 | 2 | 20.0 | 0.048 | 0.045 | 0.043 | 0.048 | 0.051 |
| 4 | 5 | 5 | 20.0 | 0.042 | 0.040 | 0.043 | 0.038 | 0.042 |
| 4 | 5 | 10 | 20.0 | 0.050 | 0.057 | 0.050 | 0.053 | 0.051 |
| $\varepsilon \sim$ *Uniform*(1, 10) | | | | | | | | |
| 4 | 5 | 2 | 1.0 | 0.050 | 0.056 | 0.055 | 0.052 | 0.055 |
| 4 | 5 | 5 | 1.0 | 0.049 | 0.055 | 0.050 | 0.051 | 0.052 |
| 4 | 5 | 10 | 1.0 | 0.058 | 0.058 | 0.059 | 0.061 | 0.060 |
| 4 | 5 | 2 | 20.0 | 0.048 | 0.042 | 0.040 | 0.045 | 0.045 |
| 4 | 5 | 5 | 20.0 | 0.046 | 0.044 | 0.045 | 0.045 | 0.047 |
| 4 | 5 | 10 | 20.0 | 0.051 | 0.050 | 0.055 | 0.053 | 0.053 |
| $\varepsilon \sim$ *Lognormal* | | | | | | | | |
| 4 | 5 | 2 | 1.0 | 0.036 | 0.032* | 0.034* | 0.034* | 0.031* |
| 4 | 5 | 5 | 1.0 | 0.062 | 0.053 | 0.062 | 0.062 | 0.056 |
| 4 | 5 | 10 | 1.0 | 0.052 | 0.045 | 0.047 | 0.047 | 0.044 |
| 4 | 5 | 2 | 20.0 | 0.054 | 0.052 | 0.052 | 0.060 | 0.053 |
| 4 | 5 | 5 | 20.0 | 0.047 | 0.048 | 0.050 | 0.057 | 0.050 |
| 4 | 5 | 10 | 20.0 | 0.040 | 0.036 | 0.036 | 0.045 | 0.039 |
| $\varepsilon \sim \exp(1)^3$ | | | | | | | | |
| 4 | 5 | 2 | 1.0 | 0.049 | 0.050 | 0.033* | 0.038 | 0.026* |
| 4 | 5 | 5 | 1.0 | 0.043 | 0.045 | 0.032* | 0.034* | 0.024* |
| 4 | 5 | 10 | 1.0 | 0.055 | 0.052 | 0.045 | 0.044 | 0.038 |
| 4 | 5 | 2 | 20.0 | 0.060 | 0.072* | 0.061 | 0.104* | 0.057 |
| 4 | 5 | 5 | 20.0 | 0.059 | 0.068* | 0.070* | 0.085* | 0.057 |
| 4 | 5 | 10 | 20.0 | 0.057 | 0.067* | 0.069* | 0.071* | 0.058 |

Yab = permutation of raw data as *ab* units, which provides an exact test in this case;
Y = unrestricted permutation of raw data;
R(reduced) = permutation of residuals under the reduced model;
R(full) = permutation of residuals under the full model;
F = normal-theory *F*-test (tabled values).

somewhat analogous to the situation encountered with parametric ANOVA: greater power can be achieved by increasing the degrees of freedom associated with the denominator mean square of the *F*-ratio.

## 5 CROSSED DESIGN, FIXED EFFECTS

Consider a fixed effects two-way ANOVA design with the following linear model:

$$y_{ijk} = \mu + A_i + B_j + AB_{ij} + \varepsilon_{ijk}$$

where A and B are fixed factors, $\mu$ is the unknown population mean, $AB_{ij}$ indicates the interaction effect of the $ij$th cell and $\varepsilon_{ijk}$ is the unknown error associated with observation $y_{ijk}$. As before, for the permutation test we assume that the $\varepsilon$'s are i.i.d.

## 5.1 Test of Interaction

There is no exact permutation test for an interaction using an *F*-ratio (but see some alternatives in Pesarin, 2001). This is so because there are no possible permutations left that would give an *F*-ratio different to the observed value when permutations are restricted to occur within levels of each of the main effects. For an approximate test that attempts to control for main effects, one can calculate and permute residuals $r_{ijk} = y_{ijk} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}$, where $\bar{y}_{i..}$ and $\bar{y}_{...}$ are as previously described and $\bar{y}_{.j.}$ is the mean of observations in level *j* of factor B. The method of permutation asymptotically approaches the exact test because, although $\mathrm{SS_A}$ and $\mathrm{SS_B}$ are not kept constant, variability due to A and B are estimated and removed by subtracting means.

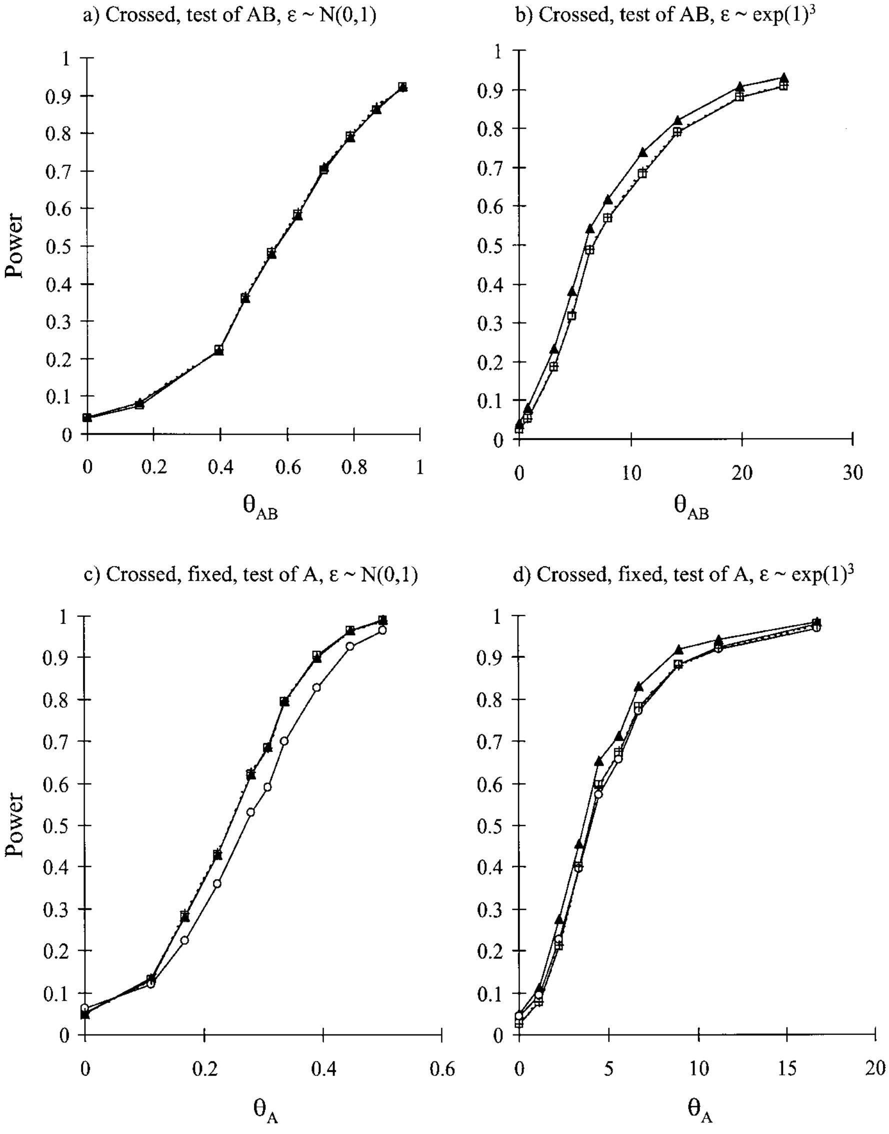


FIGURE 3 Power curves for permutation tests and the normal-theory *F*-test for a two-way crossed fixed effects design. Symbols are as in Figure 2, except O = Y(B) (permutation of raw data within levels of B).

Unrestricted permutation of raw data is another approximate test that can be used. It does not attempt to control for the main effects of A or B, but, as in all cases with complex designs, relies on the pivotal properties of the $F$-ratio. This approach mixes variability under permutation among $\mathrm{SS_A}$, $\mathrm{SS_B}$, $\mathrm{SS_{AB}}$ and $\mathrm{SS_R}$.

Simulations were done to examine type I error and relative power of these two approximate tests and the normal-theory $F$-test, as in previous situations. Here, $a = 2, b = 4, n = 5$, non-zero levels of A were set at $A_i = \{50, -50\}$ and of B were set at $B_j = \{50, -50, 20, -20\}$. Fixed levels of $AB_{ij}$ were chosen to provide a gradual increase in the interaction effect, denoted by $\theta_{\mathrm{AB}}$, where $\theta_{\mathrm{AB}} = \sqrt{\sum_{i=1}^{a}\sum_{j=1}^{b}(AB_{ij} - \overline{AB})^2/ab}$ and $\overline{AB} = \sum_{i=1}^{a}\sum_{j=1}^{b} AB_{ij}/ab$.

Type I error did not differ significantly from 0.05 for any of these methods. All of the tests had comparable power for the situation with normal errors (Fig. 3a), although the Wilcoxon test indicated that the normal-theory $F$-ratio was most powerful in this situation and for uniform errors (Tab. I). Permutation of residuals, however, was significantly more powerful than raw data permutation or the traditional $F$-test in the situation with either lognormal or exponential cubed errors (Fig. 3b, Tab. I).

### 5.2 Test of Main Effect

If the test for interaction is not significant, tests of each of the main effects are reasonable. In that case, $\mathrm{E(MS_{AB}) = E(MS_R)} = \sigma_\varepsilon^2$ and the model is $y_{ijk} = \mu + A_i + B_j + \varepsilon_{ijk}$. An exact test for factor A consists in permuting raw observations but restricting them to occur within levels of B and calculating $F_{\mathrm{A}} = \mathrm{MS_A/MS_R}$. This means $\mathrm{SS_B}$, like $\mathrm{SS_T}$, remains constant across all permutations in the test of the effect of A, so the test is exact. Note that this method does not provide an exact test in the situation where a significant AB interaction is present, as $\mathrm{SS_{AB}}$ would not be held constant in this permutation strategy.

Approximate tests are provided by either permutation of the residuals $r_{ijk} = y_{ijk} - \bar{y}_{.j.}$ or unrestricted permutation of raw data. Simulations were done again to compare relative power of these methods. For these, $a = 4, b = 4, n = 5$, $B_j = \{50, -50, 20, -20\}$ and the effect of factor A was gradually increased. All interaction effects were set at zero.

Once again, type I errors of the methods did not differ significantly from 0.05. For this situation, the exact test had significantly less power to detect the false null (hypothesis than any of the other tests, regardless of the error distribution (Tab. I, Fig. 3c). When errors were either lognormal or exponential cubed, permutation of residuals had the greatest power of any of the tests (Tab. I, Fig. 3d), while the normal-theory $F$-test was most powerful for the normal or uniform error distributions (Tab. I).

## 6 CROSSED DESIGN, MIXED MODEL

The crossed mixed model can be denoted in the same way as for the crossed fixed effects design (Section 5 above), but now consider that while factor A is fixed, factor B is a random factor.

### 6.1 Test of Fixed Factor, No Interaction

For such a mixed model, the test of interaction would be as discussed for the fixed effects model in Section 5.1 above. The test for the random effect would be provided by $F_{\mathrm{B}} = \mathrm{MS_B/MS_R}$, and would proceed as for the test of main effects described in Section 5.2 above (*i.e.*, the test of a main effect in the absence of a significant interaction). The important distinction from the completely fixed effects model comes about in the test of the fixed factor, A, in the mixed model. Here, the $F$-ratio is constructed as $F_{\mathrm{A}} = \mathrm{MS_A/MS_{AB}}$.

This indicates that for the exact test of factor A, the exchangeable units are the *ab* units (*i.e.*, Fig. 1c). Furthermore, the exact test requires that permutation of these units be restricted to occur within levels of factor B, according to the guideline.

This provides an exact test for A because (i) permutation within levels of B means $\text{SS}_\text{B}$ remains constant throughout the permutations, (ii) permutation of the *ab* units means $\text{SS}_\text{R}$ remains constant throughout the permutations (iii) $\text{SS}_\text{T} = \text{SS}_\text{A} + \text{SS}_\text{B} + \text{SS}_\text{AB} + \text{SS}_\text{R}$ and (iv) $\text{SS}_\text{T}$ remains constant throughout the permutations. Thus, the permutation strategy described exchanges variability only across $\text{SS}_\text{A}$ and $\text{SS}_\text{AB}$, which is exact for a test of $F_\text{A} = \text{MS}_\text{A}/\text{MS}_\text{AB}$.

Note that the common wisdom for such situations is that an exact test of factor A would consist of simply permuting observations within levels of B (*e.g.*, Edgington, 1995). This ignores the fact that the expected mean square of A includes a component of variability due to the interaction term, due to the fact that B is a random factor. As a consequence, restricting permutations within levels of B does not provide an exact test in this case. We included this method in our simulations (below) for comparison with the exact test according to the guideline.

Several approximate permutation methods could be used for this test. First, one can simply permute raw data values without restriction. Alternatively, one can permute the residuals $r_{ijk} = y_{ijk} - \bar{y}_{.j.}$, which attempt to remove the effect of factor B in the data through subtraction of means in order to isolate the test of factor A. One further possible approximate test is to permute these residuals, but only as *ab* units. The rationale for this test is to remove the effect of factor B by subtracting means, but to use the exchangeable units identified for the exact test, which keeps $\text{SS}_\text{R}$ constant across permutations.

Altogether, there were thus 5 different methods of permutation compared in these simulations, in addition to the traditional ANOVA $F$-statistic (Tab. I). Data were simulated with $a = 4, b = 4$ and $n = 5$. Levels of B were obtained randomly from a normal distribution with mean zero and standard deviation $\sigma_\text{B} = 20.0$. Note that, unlike the nested case, the 4 levels of B ($B_j, j = 1, \dots, b$), although random, were nevertheless the same four levels in each of the levels of A. Levels of factor A were fixed and gradually changed to increase the effect of A, while the interaction effect was set at zero.

The methods did not differ significantly from 0.05 in their empirical type I errors. When errors were normal or uniform, the normal-theory $F$-test was most powerful, followed closely by all of the permutation tests, except for the permutation of raw data within levels of B. This was significantly less powerful than any other method by quite a wide margin (Fig. 4a, Tab. I). When errors were lognormal or exponential cubed, permutation of residuals was most powerful, followed by the exact test, while permutation of raw data within levels of B was still the least powerful of all the tests (Fig. 4c, Tab. I).

## 6.2 Test of Fixed Factor, Presence of Interaction

For the fixed effects model, it is generally considered illogical to test main effects in the presence of a significant interaction. For the mixed model, however, one may be interested in the test for a significant fixed effect, over and above any variability caused by the interaction, which in this case simply contributes a random error component to the model. That is, the random interaction effect may contribute a significant source of variation (much as a nested random factor does), but this may not deter an experimenter from wishing to know, on average, if a consistent fixed main effect can still be detected across levels of the random factor.

Thus, the five different permutation methods outlined in Section 6.1 above were also compared for this situation. First, to investigate type I error, data were generated with $a = 4, b = 4$ and $n = 10$ and levels of A were set at zero, while levels of B were randomly chosen from a normal distribution with mean zero and standard deviation

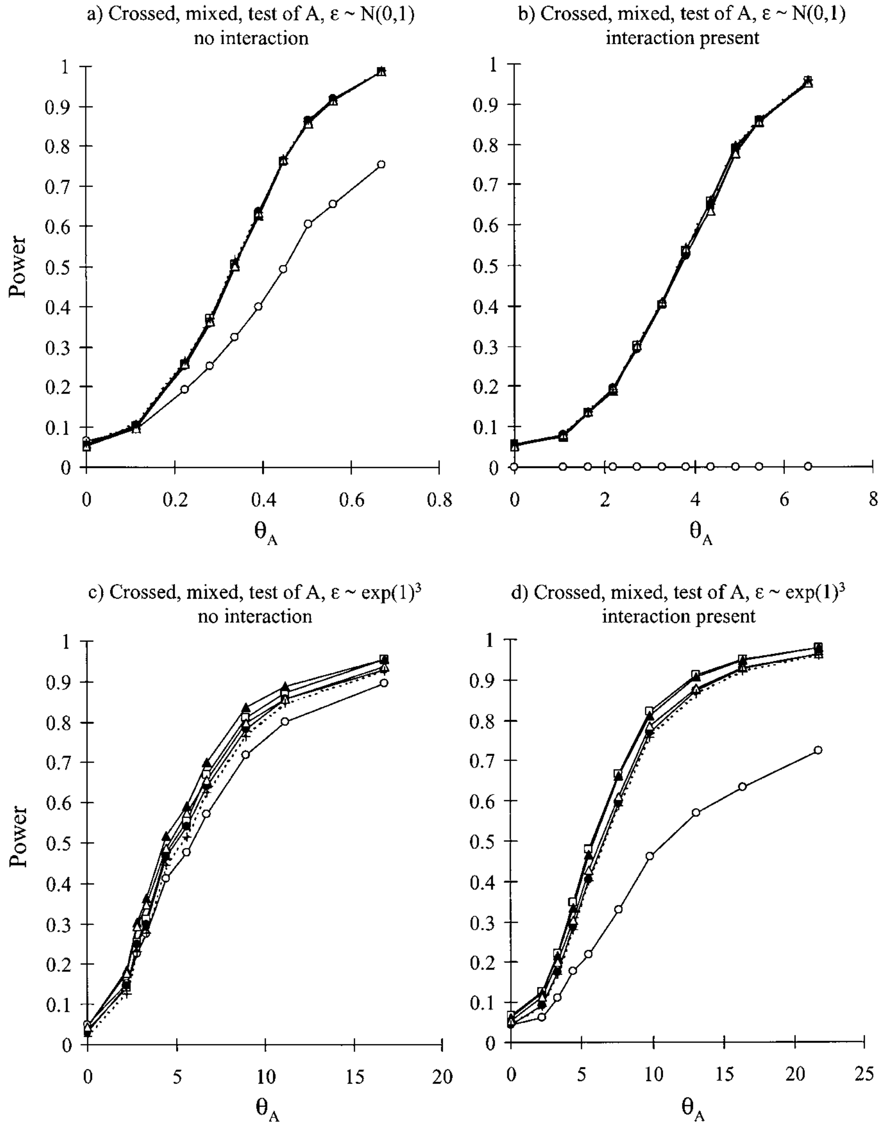


FIGURE 4 Power curves for permutation strategies and the normal-theory $F$-test for the test of a fixed factor (A) in a two-way crossed mixed model (B is random). Symbols are as follows:
□ = Y (permutation of raw data),
▲ = R (permutation of residuals),
○ = Y(B) (permutation of raw data restricted within levels of B),
● = Rab (permutation of residuals as *ab* units),
△ = Y(B)ab (permutation of raw data as *ab* units, restricted within levels of B),
+ = normal-theory $F$-test.

$\sigma_B = \{0, 1, 5, \text{ or } 10\}$. Levels of AB were also chosen from a normal distribution with mean zero and standard deviation $\sigma_{AB} = \{0, 1, 5, \text{ or } 10\}$. To investigate power, parameters were set at $a = 4$, $b = 4$, $n = 5$, $\sigma_B = 1.0$ and $\sigma_{AB} = 5.0$, while the effect of A was gradually increased. Note that if the standard deviation for random factor B in the crossed mixed model (which we may call $\sigma_{B,\text{crossed}}$ for clarity) is equal to zero, then the test of the fixed factor A in the crossed mixed model would be the same as the test of A in the nested design whenever $\sigma_{\mathrm{AB}} = \sigma_{\mathrm{B,nested}}$. However, when $\sigma_{\mathrm{B,crossed}} \neq 0$ and $\sigma_{\mathrm{AB}} \neq 0$, the test of A in the crossed mixed design provides a new situation, which is the case of interest here.

The exact test (*i.e.*, permuting *ab* units within levels of factor B, as specified by the guideline) and permuting residuals of factor B as *ab* units were the only tests that adequately maintained type I error in all situations (Tab. III). The traditional normal-theory $F$-test maintained

TABLE III Proportion of rejections of the true null hypothesis (type I error) out of 1000 simulations with significance level = 0.05 for the test of a fixed factor (A) in a two-way crossed mixed model ANOVA design, with $F_{\mathrm{A}} = \mathrm{MS_A}/\mathrm{MS_{AB}}$. The $b$ levels of the random factor B and the *ab* levels of the interaction AB were each chosen randomly from a normal distribution with mean zero and standard deviations $\sigma_{\mathrm{B}}$ and $\sigma_{\mathrm{AB}}$, respectively. Values that lie outside the 95% confidence interval for type I error (which has a binomial ($n$, $p$) distribution with $n = 1000$ simulations and $p = 0.05$, *i.e.*, 0.036–0.064) are indicated with an asterisk.

| $\sigma_B$ | $\sigma_{AB}$ | *Y(B)ab* | *Y(B)* | *Y* | *R* | *Rab* | *F* |
|---|---|---|---|---|---|---|---|
| | | | | $\varepsilon \sim N(0, 1)$ | | | |
| 0 | 0 | 0.047 | 0.052 | 0.047 | 0.047 | 0.048 | 0.049 |
| 0 | 5 | 0.040 | 0.000* | 0.047 | 0.046 | 0.041 | 0.040 |
| 0 | 10 | 0.049 | 0.000* | 0.045 | 0.047 | 0.047 | 0.047 |
| 5 | 0 | 0.056 | 0.054 | 0.050 | 0.052 | 0.049 | 0.051 |
| 5 | 5 | 0.051 | 0.000* | 0.045 | 0.049 | 0.049 | 0.048 |
| 5 | 10 | 0.042 | 0.000* | 0.039 | 0.040 | 0.041 | 0.039 |
| 10 | 0 | 0.050 | 0.057 | 0.048 | 0.047 | 0.048 | 0.050 |
| 10 | 5 | 0.043 | 0.000* | 0.047 | 0.047 | 0.047 | 0.046 |
| 10 | 10 | 0.057 | 0.000* | 0.058 | 0.061 | 0.057 | 0.059 |
| | | | | $\varepsilon \sim$ *Uniform*(1, 10) | | | |
| 0 | 0 | 0.057 | 0.053 | 0.058 | 0.058 | 0.055 | 0.058 |
| 0 | 5 | 0.062 | 0.000* | 0.057 | 0.061 | 0.055 | 0.060 |
| 0 | 10 | 0.056 | 0.000* | 0.050 | 0.047 | 0.055 | 0.052 |
| 5 | 0 | 0.054 | 0.052 | 0.050 | 0.051 | 0.051 | 0.051 |
| 5 | 5 | 0.044 | 0.000* | 0.045 | 0.044 | 0.040 | 0.047 |
| 5 | 10 | 0.040 | 0.000* | 0.043 | 0.046 | 0.041 | 0.043 |
| 10 | 0 | 0.050 | 0.047 | 0.046 | 0.047 | 0.045 | 0.046 |
| 10 | 5 | 0.051 | 0.001* | 0.046 | 0.049 | 0.050 | 0.046 |
| 10 | 10 | 0.059 | 0.000* | 0.063 | 0.060 | 0.060 | 0.058 |
| | | | | $\varepsilon \sim$ *Lognormal* | | | |
| 0 | 0 | 0.052 | 0.055 | 0.056 | 0.052 | 0.049 | 0.048 |
| 0 | 5 | 0.044 | 0.000* | 0.047 | 0.043 | 0.042 | 0.043 |
| 0 | 10 | 0.052 | 0.000* | 0.054 | 0.053 | 0.050 | 0.049 |
| 5 | 0 | 0.059 | 0.051 | 0.051 | 0.055 | 0.048 | 0.050 |
| 5 | 5 | 0.057 | 0.000* | 0.052 | 0.054 | 0.054 | 0.052 |
| 5 | 10 | 0.048 | 0.000* | 0.046 | 0.048 | 0.041 | 0.045 |
| 10 | 0 | 0.056 | 0.053 | 0.057 | 0.065* | 0.056 | 0.056 |
| 10 | 5 | 0.042 | 0.000* | 0.042 | 0.043 | 0.039 | 0.042 |
| 10 | 10 | 0.045 | 0.000* | 0.043 | 0.042 | 0.040 | 0.042 |
| | | | | $\varepsilon \sim \exp(1)^3$ | | | |
| 0 | 0 | 0.052 | 0.052 | 0.056 | 0.054 | 0.042 | 0.032* |
| 0 | 5 | 0.047 | 0.047 | 0.065* | 0.068* | 0.039 | 0.038 |
| 0 | 10 | 0.058 | 0.042 | 0.076* | 0.074* | 0.055 | 0.055 |
| 5 | 0 | 0.048 | 0.045 | 0.048 | 0.046 | 0.038 | 0.032* |
| 5 | 5 | 0.056 | 0.059 | 0.067* | 0.070* | 0.046 | 0.041 |
| 5 | 10 | 0.056 | 0.039 | 0.067* | 0.073* | 0.051 | 0.050 |
| 10 | 0 | 0.058 | 0.055 | 0.055 | 0.062 | 0.044 | 0.038 |
| 10 | 5 | 0.055 | 0.052 | 0.058 | 0.060 | 0.041 | 0.039 |
| 10 | 10 | 0.042 | 0.033* | 0.059 | 0.062 | 0.042 | 0.041 |

Y = permutation of raw data; R = permutation of residuals;

Y(B) = permutation of raw data within levels of B;

Rab = permutation of residuals, as *ab* units;

Y(B)ab = permutation of raw data, as *ab* units, within levels of B, which gives the exact test in this case;

F = normal-theory $F$-test (tabled values).

type I error for situations with normal errors, but tended to be too conservative when the errors were exponential cubed (Tab. III). In contrast, the method of permutation usually advocated for such tests (*i.e.*, permuting raw data within levels of factor B) gave empirical type I errors that were extremely conservative in the case of normal, uniform or lognormal errors when simulated data included variation due to the interaction term (*i.e.*, for non-zero $\sigma_{\mathrm{AB}}$, Tab. III). In contrast, the approximate methods of permuting residuals of factor B or permuting raw data showed inflated type I error in certain circumstances with exponential cubed errors and non-zero $\sigma_{\mathrm{AB}}$ (Tab. III).

In terms of power, with normal errors, the traditional *F*-test was significantly more powerful than the other tests (Tab. I, Fig. 4b). There were no significant differences among the other tests, except all of them were more powerful than the permutation of observation units within levels of the other factor (*i.e.*, Y(B), Tab. I). In fact, with these parameters (non-zero interaction and normal errors), this test failed to find any significant effects of factor A at all, even when the power of the other tests was near 100% (Tab. III, Fig. 4b).

When errors were radically non-normal (*i.e.*, exponential cubed), then the permutation of raw data and permutation of residuals had the greatest power (Tab. I, Fig. 4d), however, neither of these methods maintained type I error at nominal levels for all situations (Tab. III). The next most powerful test was the exact test (*i.e.*, Y(B)ab), followed by permutation of residuals of B as *ab* units, followed by the *F*-test (Tab. I, Fig. 4d). Although the permutation of raw data within levels of factor B did have some power in this situation, in contrast to the extreme results obtained using normal errors, this method was still severely lacking in power compared to all other approaches (Fig. 4d). This method was also the least powerful when either uniform or lognormal errors were simulated (Tab. I).

## 7 DISCUSSION

We have provided a guideline that indicates, on the basis of expected mean squares, the construction of an exact permutation test for individual terms in ANOVA models. This may include permutation of groups of units as well as restricted permutations within levels of factors. We have provided examples for two-way ANOVA designs that demonstrate how the exact tests and various approximate permutation methods can be done and we have compared their power empirically for these two-way designs using simulations.

Results from simulations have shown that, in general, when errors were either normal or uniformly distributed (both symmetric distributions), the normal-theory *F*-test (using traditional tabled values) provided the most powerful test, in addition to maintaining correct type I error. However, when errors were lognormal or exponential cubed (both right-skewed distributions), permutation of residuals under a reduced model (*e.g.*, Still and White, 1981; Freedman and Lane, 1983) provided the most powerful test, while maintaining type I error, for virtually all situations with crossed designs. The exception to this was the test of a fixed factor in the presence of a non-zero interaction in a mixed model. Here, as for the nested designs with such error distributions, the exact test was best for maintaining type I error and having reasonable power.

Restricted permutation methods virtually always resulted in tests with very low power. Relatively low power for restricted permutations was also found by Gonzalez and Manly (1998). The fact that permutation of residuals provided greater power than restricted permutations for ANOVA designs challenges the current wisdom with regard to appropriate permutation procedures. We hasten to add that this result was not caused by the number of restricted permutations being too few to obtain reasonable power. For example, in our simulations, the number of possible restricted permutations for the test of A in the fixed model was $2.52 \times 10^8$ and for the mixed model was $5.71 \times 10^{34}$.

The above observations on restricted permutations apply to all tests of ANOVA terms where the denominator mean square is the residual mean square. Whenever the denominator mean square for the test of the term of interest is not the residual, then this indicates the exchangeable units for the test. These exchangeable units must be used for such tests in order to avoid bias in type I error. This applies, for example, in the case of the test of a higher ranked factor in a nested hierarchy. The units induced by the nested factor must be used as the exchangeable units for the test. It also applies in the case of a test for a significant fixed main effect over any interaction effects in a mixed model. Here, the *ab* units must be permuted for the test. It is important to note that in these situations, none of the approximate permutation methods should be used which do not permute these exchangeable units, as they do not maintain accurate type I error in all situations.

The bottom line from these simulations is that one must control for other terms in the model that are of higher order (*e.g.* of lower rank in a hierarchy), where this is necessary, by permuting appropriate units, which may not be the individual observation units. To control for terms that are of the same or lower order, in contrast, one generally has the option of permuting residuals or of permuting within levels of those factors. The latter will give an exact test while permuting residuals generally results in greater power and good type I error, provided the correct exchangeable units are used.

The specific situation of a test for a significant main effect (A) in the presence of an interaction in a mixed model deserves some more attention here. It was found that restricting permutations within levels of the random factor (B) resulted in no rejections of the null hypothesis, for normal errors (Tab. III, Fig. 4c). Similarly, virtually no rejections occurred for either lognormal or uniform error distributions in this situation. Why should this method of permutation have such low power? When there is significant variability due to an interaction, restricting permutations within levels of B causes the mean square of A, on average, to be inflated under permutation when the interaction effect is not controlled in any way. Thus, the expected value of the mean square of A under permutation is, on average, too large, causing the distribution of the *F*-statistic under permutation to be shifted to the right (Fig. 5). This means the observed value does not appear extreme relative to this distribution under permutation as frequently as it should, individual *P*-values tend to be over-estimated and the resulting distribution of *P*-values is no longer uniform (Fig. 5). So, type I error does not equal $\alpha$ (it is too small) and power is compromised, sometimes severely. Further work is certainly necessary to clarify the role of restrictions in changing the distribution of relevant test statistics under permutation. The above explanation (along with Fig. 5) appears to provide at least some insight into this problem, although we do not pursue it further here.

Certain trade-offs attend the use of an exact test versus permutation of residuals. The exact test obviously ensures that type I error will be maintained at the chosen significance level ($\alpha$). While the type I error of a permutation test using residuals asymptotically approaches this nominal level, even by using radically non-normal errors we were not able to simulate a situation where permutation of residuals resulted in inflated type I error, provided the appropriate denominator units were permuted for the test. Our results are consistent with empirical results obtained for the method of permutation of residuals described by Freedman and Lane (1983) for multiple regression (Anderson and Legendre, 1999). In addition, there are some terms in ANOVA models for which no exact test could ever be constructed. If a permutational approach is to be used in these situations, an approximate permutation test must be chosen.

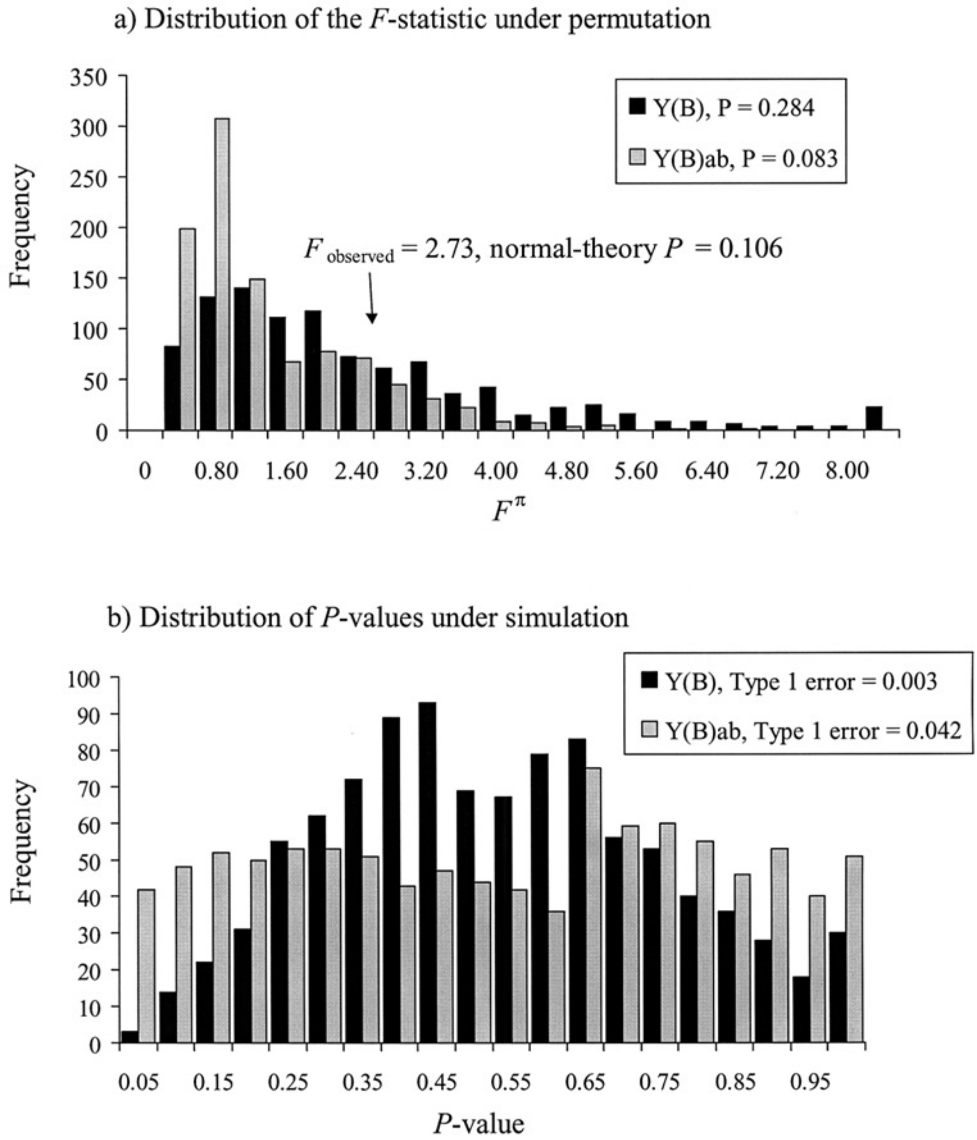


FIGURE 5 Comparison of the distribution of the $F$-statistic under permutation (there were 999 permutations) and the distribution of $P$-values under simulation (1000 simulations) for two permutation methods, denoted as in previous figures. The null hypothesis was true and the data were simulated according to a two-way mixed model with $a = 4, b = 4, n = 10, \varepsilon \sim N(0, 1), \sigma_{\text{B}} = 0$ and $\sigma_{\text{AB}} = 1$.

Unrestricted permutation of raw data had significantly less power than permutation of residuals and also suffered from inflated type I error in the test of a higher ranked factor in a nested hierarchical design (Section 4.2). Although it can provide a reasonable, if somewhat conservative, approximate test (Gonzalez and Manly, 1998), it cannot be used where individual units are not the correct exchangeable units (*i.e.*, when the denominator mean square for the test is not the residual mean square).

For reference, we provide tables in an Appendix (Tabs. AI–AIV) that show how exact tests can be constructed for individual terms in ANOVA for two-way and three-way designs.

The tables also indicate which terms need to be "removed" (through subtraction of cell means) for the approximate test provided by permutation of residuals. They also indicate the situations for which no exact test can be constructed.

We have taken the philosophy here that any ANOVA designed experiment can be analyzed using a permutation test approach, provided due care is given to identifying nested structures and whether individual factors are fixed or random. It is worth noting that the whole notion of treating a factor as "random" might be considered by some as an anathema in the permutation context [*i.e.*, see Manly (1997), Section 7.6, pp. 142–143]. In that framework, the observations as well as any levels of any factors are considered fixed by necessity. This is so because this approach considers randomization tests to obtain their validity by virtue of random assignment of units to combinations of treatments, with the tests being conditional on the factor levels actually included in the experiment, regardless of how levels were chosen.

In contrast, our approach, as stated in the introduction, is that permutation tests gain their validity by virtue of exchangeability of errors under the null hypothesis. This allows for the possibility of there being more than just the errors associated with individual units as possible sources of variability under the null hypothesis, including random interaction terms or nested factors.

Regardless of which philosophical approach an experimenter chooses to take, we have shown here using simulations that taking an approach which ignores expected mean squares in the construction of a test statistic or in the choice of exchangeable units for the test can have undesirable consequences.

Although the experimenter has some choices in the use of permutation tests for complex designs in ANOVA, permutations must not be done indiscriminately or without thought as to the nature of the factors and the structure of the design. The logic attending parametric procedures in ANOVA applies to permutation strategies also, in terms of the expected mean squares, the pivotal $F$-statistic used for the test and the way in which the permutations are to be done. The guideline we provide for the construction of exact tests is consistent with the traditional rules applied in parametric ANOVA and is driven by the logic of the experimental design and the expected mean squares of individual terms. As for parametric tests, the proper logic of experimental design and analysis must be observed with care for valid tests of hypotheses to be obtained using permutations.

The guideline we have provided here has great utility for choosing appropriate permutation strategies for the analysis of multivariate data in complex ANOVA designs (*e.g.*, ter Braak and Šmilauer, 1998; Anderson, 2001). For multivariate data, different variables within the same data set may often have different kinds of distributions, thus making the use of GLMs and other related approaches unfeasible. In contrast, permutation tests based on distance matrices are quite robust in this situation and can be done using the appropriate permutation procedures as described here (*e.g.*, McArdle and Anderson, 2001).

We have not discussed the potential use of permutation methods with more robust statistics, such as for tests of medians or using least absolute deviations (*e.g.*, see Cade and Richards, 1996) rather than least-squares. For univariate data with highly skewed distributions, such approaches may well be more appropriate than using permutation methods with the traditional $F$-ratios. We propose, however, that the same issues we have addressed here (*e.g.*, the choice of exchangeable units and the choice of whether or not to restrict permutations and/or permute residuals) will still be very important to consider in order to develop appropriate permutation distributions for tests in complex ANOVA designs, regardless of the kind of statistic used. It is clear that further work in this area is necessary, including by reference to the utility and relative power of such approaches compared to, for example, GLMs or GLMMs (*e.g.*, Breslow and Clayton, 1993) in different situations.

## 8 CONCLUSION

We summarize and conclude our results by briefly answering questions (a) through (f) posed in the Introduction:

(a) *Should raw data be permuted or some form of residuals?* To obtain an exact test, permutation of raw data can be done using the guideline according to the ANOVA model (*i.e.*, permuting appropriate exchangeable units and restrictions within levels of other factors). Permuting the correct exchangeable units is not optional, but mandatory to maintain level accuracy. Having identified the appropriate exchangeable units, however, one generally then has the choice of whether to further restrict permutations (to accomplish an exact test) or to permute residuals. Our results indicate that permutation of residuals will have power that is greater than (or equal to) the exact test, while maintaining type I error. We therefore recommend permutation of residuals for general applications. Indeed, in situations where an exact test is impossible, permutation of residuals may give the most powerful and reliable approximate test.

(b) *Which units should be permuted?* Exchangeable units for the test are identified by the denominator mean square of the $F$-ratio for the test as determined by expected mean squares.

(c) *When should permutations be restricted to occur within levels of other factors?* When there are factors of similar or lower order, these may be taken into account by restricting permutations within their levels for an exact test. The alternative to such restrictions (yielding an approximate test, generally with greater power) is to permute residuals of those factors.

(d) *How can interaction terms be tested using permutations?* No general exact test exists using the $F$-ratio. Permuting the residuals of main effects provides a powerful and valid approximate permutation test for interaction terms.

(e) *Which tests are exact and which are approximate?* A test is exact when its type I error is exactly equal to the *a priori* significance level chosen for the test. Tests are approximate (asymptotically exact) when their type I error asymptotically approaches the *a priori* significance level chosen for the test. Permutation of any kind of residuals will always give approximate tests. Exact tests are those constructed using the guideline.

(f) *What test statistic should be used?* A pivotal statistic, such as $F$, is necessary for the approximate permutation tests in complex ANOVA designs (*e.g.*, Anderson and Robinson, 2001). For the exact tests, there are simpler statistics, such as the sum of squares or mean square, that are monotonically related to the $F$-ratio, thus yielding equivalent $P$-values under permutation (*e.g.*, Edgington, 1995). As the $F$-ratio has an immediately familiar and interpretable meaning, however, we recommend its use for any test by permutation in ANOVA.

### *Acknowledgements*

This research was supported by an Australian Postgraduate Award and a U2000 Post-doctoral Fellowship to MJA at the University of Sydney. We thank B. H. McArdle and A. J. Underwood for discussions and comments regarding its content.

### *References*

Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral. Ecol.*, **26**, 32–46.

Anderson, M. J. and Legendre, P. (1999). An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. *J. Statist. Comput. Simul.*, **62**, 271–303.

Anderson, M. J. and Robinson, J. (2001). Permutation tests for linear models. *Austral. & New Zealand J. Statist.*, **43**, 75–88.

Boik, R. J. (1987). The Fisher-Pitman permutation test: a non-robust alternative to the normal theory F-test when variances are heterogeneous. *Br. J. Math. Stat. Psychol.*, **40**, 26–42.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.*, **88**, 9–25.

Brown, B. M. and Maritz, J. S. (1982). Distribution-free methods in regression. *Austral. J. Statist.*, **24**, 318–331.

Cade, B. S. and Richards, J. D. (1996). Permutation tests for least absolute deviation regression, *Biometrics*, **52**, 886–902.

Clarke, K. R. (1993). Non-parametric multivariate analysis of changes in community structure. *Austral. J. Ecol.*, **18**, 117–143.

Cochran, W. G. (1974). Some consequences when the assumptions for the analysis of variance are not satisfied. *Biometrics*, **3**, 22–38.

Crowley, P. H. (1992). Resampling methods for computation-intensive data analysis in ecology and evolution. *Annu. Rev. Ecol. Syst.*, **23**, 405–447.

Edgington, E. S. (1995). *Randomization Tests*, 3rd ed. Marcel Dekker, New York.

Excoffier, L., Smouse, P. E. and Quattro, J. M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*, **131**, 479–491.

Fisher, R. A. (1935). *Design of Experiments.* Oliver and Boyd, Edinburgh.

Freedman, D. and Lane, D. (1983). A nonstochastic interpretation of reported significance levels. *J. Bus. Econ. Statist.*, **1**, 292–298.

Gaston, K. J. and McArdle, B. H. (1994). The temporal variability of animal abundances: measures, methods and patterns. *Philos. Trans. R. Soc. London Ser. B*, **345**, 335–358.

Gonzalez, L. and Manly, B. F. J. (1998). Analysis of variance by randomization with small data sets. *Environmetrics*, **9**, 53–65.

Good, P. I. (2000). *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*, 2nd ed. Springer-Verlag, Berlin.

Gower, J. C. and Krzanowski, W. J. (1999). Analysis of distance for structured multivariate data and extensions to multivariate analysis of variance. *J. Roy. Statist. Soc. Ser. C Appl. Statist.*, **48**, 505–519.

Hayes, A. F. (1996). Permutation test is not distribution free. *Psychol. Methods*, **1**, 184–198.

Hope, A. C. (1968). A simplified Monte Carlo significance test procedure. *J. Roy. Statist. Soc. Ser. B*, **30**, 582–598.

Hubert, L. and Schultz, J. (1976). Quadratic assignment as a general data analysis strategy. *Br. J. Math. Stat. Psychol.*, **29**, 190–241.

Johnson, C. R. and Field, C. A. (1993). Using fixed-effects model multivariate analysis of variance in marine biology and ecology. *Oceanogr. Mar. Biol. Ann. Rev.*, **31**, 177–221.

Kempthorne, O. (1952). *The Design and Analysis of Experiments.* John Wiley & Sons, New York.

Kempthorne, O. (1955). The randomization theory of experimental inference. *J. Amer. Statist. Assoc.*, **50**, 946–967.

Kempthorne, O. (1966). Some aspects of experimental inference. *J. Amer. Statist. Assoc.*, **61**, 11–34.

Kennedy, P. E. and Cade, B. S. (1996). Randomization tests for multiple regression. *Commun. Statist. – Simulation Comput.*, **25**, 923–936.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.

Manly, B. F. J. (1997). *Randomization, Bootstrap and Monte Carlo Methods in Biology*, 2nd ed. Chapman and Hall, London.

Mantel, N. and Valand, R. S. (1970). A technique of nonparametric multivariate analysis. *Biometrics*, **26**, 547–558.

Mardia, K. V. (1971). The effect of non-normality on some multivariate tests and robustness to non-normality in the linear model. *Biometrika*, **58**, 105–121.

McArdle, B. H. and Anderson, M. J. (2001). Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology*, **82**, 290–297.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd ed. Chapman and Hall, London.

Mielke, P. W., Berry, K. J. and Johnson, E. S. (1976). Multi-response permutation procedures for a priori classifications. *Commun. Stat. – Theory and Methods A*, **5**(14), 1409–1424.

Nelder, J. A. (1964a). The analysis of randomized experiments with orthogonal block structure. I. Block structure and the null analysis of variance. *Proc. Roy. Statist. Soc. Ser. A*, **283**, 147–162.

Nelder, J. A. (1964b). The analysis of randomized experiments with orthogonal block structure. II. Treatment structure and the general analysis of variance. *Proc. Roy. Statist. Soc. Ser. A*, **283**, 163–178.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *J. Roy. Statist. Soc. Ser. A*, **135**, 370–384.

Olson, C. L. (1974). Comparative robustness of six tests in multivariate analysis of variance. *J. Amer. Statist. Assoc.*, **69**, 894–908.

Pesarin, F. (2001). *Multivariate Permutation Tests: With Applications in Biostatistics.* John Wiley and Sons, Chichester.

Pillar, V. D. P. and Orlóci, L. (1996). On randomization testing in vegetation science: multifactor comparisons of relevé groups. *J. Veg. Sci.*, **7**, 585–592.

Pitman, E. J. G. (1937). Significance tests which may be applied to samples from any populations. III. The analysis of variance test. *Biometrika*, **29**, 322–335.

Scheffé, H. (1959). *The Analysis of Variance.* John Wiley & Sons, New York.

Searle, S. R., Casella, G. and McCulloch, C. E. (1992). *Variance Components.* John Wiley & Sons, New York.

Smith, E. P., Pontasch, K. W. and Cairns, J. (1990). Community Similarity and the analysis of multispecies environmental data: a unified statistical approach. *Water Res.*, **24**, 507–514.

Snedecor, G. W. and Cochran, W. G. (1967). *Statistical Methods*, 6th ed. University of Iowa Press, Ames, Iowa.

Still, A. W. and White, A. P. (1981). The approximate randomization test as an alternative to the *F* test in analysis of variance. *Br. J. Math. Stat. Psychol.*, **34**, 243–252.

ter Braak, C. J. F. (1992). Permutation versus bootstrap significance test in multiple regression and ANOVA. In: Jöckel, K.-H., Rothe, G. and Sendler, W. (Eds.), *Bootstrapping and Related Techniques.* Springer-Verlag, Berlin, pp. 79–86.

ter Braak, C. J. F. and Šmilauer, P. (1998). *CANOCO Reference Manual and User's Guide to Canoco for Windows: Software for Canonical Community Ordination* (version 4). Micro Computer Power, Ithaca, New York.

Taylor, L. R. (1961). Aggregation and the mean. *Nature*, **189**, 732–735.

Welch, W. J. (1990). Construction of permutation tests. *J. Amer. Statist. Assoc.*, **85**, 693–698.

Winer, B. J., Brown, D. R. and Michels, K. M. (1991). *Statistical Principles in Experimental Design*, 3rd ed. McGraw-Hill, New York.

Zeger, S. L. and Liang, K.-Y. (1992). An overview of the methods for the analysis of longitudinal data. *Statistics in Medicine*, **11**, 1825–1839.

## APPENDIX

TABLE AI Permutation strategies for tests of particular terms in two-way analysis of variance. The linear model is: $y_{ijk} = \mu + A_i + B(A)_{j(i)} + \varepsilon_{ijk}$, where B is nested within A (i) and $y_{ijk} = \mu + A_i + B_j + AB_{ij} + \varepsilon_{ijk}$, where factors A and B are crossed. In the latter case, possible designs include: (ii) A is fixed and B is fixed, (iii) A is fixed and B is random, (iv) A is random and B is random. Note that permutation of residuals refers to residuals under a reduced model (*i.e.*, Freedman and Lane, 1983).

| *Term in model being tested* | *Terms contributing components of variation to expected MS of term being tested (R = residual)* | *Term whose MS is used as the denominator for the F-test (R = residual)* | *Exchangeable units for the test* | *Exact test:* *Additional restrictions applied to exchangeable units* | *Permutation of residuals:* *Terms in model to be removed by subtraction of cell means* |
|---|---|---|---|---|---|
| (i) B nested within A | | | | | |
| A | R + B(A) + A | B(A) | *ab*'s | No further restrictions | None |
| B(A) | R + B(A) | R | Replicates | Within levels of A | A |
| (ii) A fixed, B fixed | | | | | |
| A | R + A | R | Replicates | Within levels of B | B |
| B | R + B | R | Replicates | Within levels of A | A |
| A × B | R + A × B | R | Replicates | No test of interaction | A, B |
| (iii) A fixed, B random | | | | | |
| A | R + A × B + A | A × B | *ab*'s | Within levels of B | B |
| B | R + B | R | Replicates | Within levels of A | A |
| A × B | R + A × B | R | Replicates | No test of interaction | A, B |
| (iv) A random, B random | | | | | |
| A | R + A × B + A | A × B | *ab*'s | Within levels of B | B |
| B | R + A × B + B | A × B | *ab*'s | Within levels of A | A |
| A × B | R + A × B | R | Replicates | No test of interaction | A, B |

TABLE AII Permutation strategies for tests of particular terms in three-way analysis of variance where the experimental design involves nesting of one factor in the interaction between the other two factors. The linear model is: $y_{ijkl} = \mu + A_i + B_j + AB_{ij} + C(AB)_{k(ij)} + \varepsilon_{ijkl}$, where factors A and B are crossed and C is nested within A × B. The possible designs include: (i) A is fixed and B is fixed, (ii) A is fixed and B is random, (iii) A is random and B is random. Note that permutation of residuals refers to residuals under a reduced model (*i.e.*, Freedman and Lane, 1983).

| | | | | *Exact test:* | *Permutation of residuals:* |
|---|---|---|---|---|---|
| *Term in model being tested* | *Terms contributing components of variation to expected MS of term being tested (R = residual)* | *Term whose MS is used as the denominator for the F-test (R = residual)* | *Exchangeable units for the test* | *Additional restrictions applied to exchangeable units* | *Terms in model to be removed by subtraction of cell means* |
| (i) A fixed, B fixed | | | | | |
| A | R + C(A × B) + A | C(A × B) | *abc*'s | Within levels of B | B, A × B |
| B | R + C(A × B) + B | C(A × B) | *abc*'s | Within levels of A | A, A × B |
| A × B | R + C(A × B) + A × B | C(A × B) | *abc*'s | No test of interaction | A, B |
| C(A × B) | R + C(A × B) | R | Replicates | Within levels of A and B | A, B, A × B |
| (ii) A fixed, B random | | | | | |
| A | R + C(A × B) + A × B + A | A × B | *ab*'s | No further restrictions | B |
| B | R + C(A × B) + B | C(A × B) | *abc*'s | Within levels of A | A, A × B |
| A × B | R + C(A × B) + A × B | C(A × B) | *abc*'s | No test of interaction | A, B |
| C(A × B) | R + C(A × B) | R | Replicates | Within levels of A and B | A, B, A × B |
| (iii) A random, B random | | | | | |
| A | R + C(A × B) + A × B + A | A × B | *ab*'s | No further restrictions | B |
| B | R + C(A × B) + A × B + B | A × B | *ab*'s | No further restrictions | A |
| A × B | R + C(A × B) + A × B | C(A × B) | *abc*'s | No test of interaction | A, B |
| C(A × B) | R + C(A × B) | R | Replicates | Within levels of A and B | A, B, A × B |

TABLE AIII Permutation strategies for tests of particular terms in three-way ANOVA where the experimental design involves nesting in only one of two crossed factors. The linear model is: $y_{ijkl} = \mu + A_i + B_j + C(B)_{k(j)} + AB_{ij} + AC(B)_{ik(j)} + \varepsilon_{ijkl}$, where factors A and B are crossed and C is nested within B. The possible designs include: (i) A is fixed and B is fixed, (ii) A is fixed and B is random, (iii) A is random and B is fixed, and (iv) A is random and B is random. Note that permutation of residuals refers to residuals under a reduced model (*i.e.*, Freedman and Lane, 1983).

| | | | | *Exact test:* | *Permutation of residuals:* |
|---|---|---|---|---|---|
| *Term in model being tested* | *Terms contributing components of variation to expected MS of term being tested (R = residual)* | *Term whose MS is used as the denominator for the F-test (R = residual)* | *Exchangeable units for the test* | *Additional restrictions applied to exchangeable units* | *Terms in model to be removed by subtraction of cell means* |
| (i) A fixed, B fixed | | | | | |
| A | R + A × C(B) + A | A × C(B) | *abc*'s | Within levels of B | B, A × B |
| B | R + C(B) + B | C(B) | *bc*'s | Within levels of A | A, A × B |
| C(B) | R + C(B) | R | Replicates | Within levels of A and B | A, B, A × B, A × C(B) |
| A × B | R + A × C(B) + A × B | A × C(B) | *abc*'s | No test of interaction | A, B |
| A × C(B) | R + A × C(B) | R | Replicates | No test of interaction | A, B, C(B), A × B |
| (ii) A fixed, B random | | | | | |
| A | R + A × C(B) + A × B + A | A × B | *abc*'s | No further restrictions | B |
| B | R + C(B) + B | C(B) | *bc*'s | Within levels of A | A, A × B |
| C(B) | R + C(B) | R | Replicates | Within levels of A and B | A, B, A × B, A × C(B) |
| A × B | R + A × C(B) + A × B | A × C(B) | *abc*'s | No test of interaction | A, B |
| A × C(B) | R + A × C(B) | R | Replicates | No test of interaction | A, B, C(B), A × B |
| (iii) A random, B fixed | | | | | |
| A | R + A × C(B) + A | A × C(B) | *abc*'s | Within levels of B | B, A × B |
| B | R + A × C(B) + A × B + C(B) + B | None – pooling required | – | – | – |
| C(B) | R + A × C(B) + C(B) | A × C(B) | *abc*'s | Within levels of B | A, B, A × B |
| A × B | R + A × C(B) + A × B | A × C(B) | *abc*'s | No test of interaction | A, B |
| A × C(B) | R + A × C(B) | R | Replicates | No test of interaction | A, B, C(B), A × B |
| (iv) A random, B random | | | | | |
| A | R + A × C(B) + A × B + A | A × B | *ab*'s | No further restrictions | B |
| B | R + A × C(B) + A × B + C(B) + B | None – pooling required | – | – | – |
| C(B) | R + A × C(B) + C(B) | A × C(B) | *abc*'s | Within levels of B | A, B, A × B |
| A × B | R + A × C(B) + A × B | A × C(B) | *abc*'s | No test of interaction | A, B |
| A × C(B) | R + A × C(B) | R | Replicates | No test of interaction | A, B, C(B), A × B |

TABLE AIV Permutation strategies for tests of particular terms in three-way analysis of variance where the experimental design involves three crossed factors. The linear model is: $y_{ijkl} = \mu + A_i + B_j + C_k + AB_{ij} + AC_{ik} + BC_{jk} + ABC_{ijk} + \varepsilon_{ijkl}$. The possible designs include: (i) A, B and C are fixed, (ii) A and B are fixed and C is random, (iii) A is fixed and B and C are random, and (iv) A, B and C are random. Note that permutation of residuals refers to residuals under a reduced model (*i.e.*, Freedman and Lane, 1983).

| *Term in model being tested* | *Terms contributing components of variation to expected MS of term being tested (R = residual)* | *Term whose MS is used as the denominator for the F-test (R = residual)* | *Exchangeable units for the test* | *Exact test: Additional restrictions applied to exchangeable units* | *Permutation of residuals: Terms in model to be removed by subtraction of cell means* |
|---|---|---|---|---|---|
| (i) A fixed, B fixed, C fixed | | | | | |
| A | R + A | R | Replicates | Within levels of B and C | B, C, A × B, A × C, B × C, A × B × C |
| B | R + B | R | Replicates | Within levels of A and C | A, C, A × B, A × C, B × C, A × B × C |
| C | R + C | R | Replicates | Within levels of A and B | A, B, A × B, A × C, B × C, A × B × C |
| A × B | R + A × B | R | Replicates | No test of interaction | A, B, C, A × C, B × C, A × B × C |
| A × C | R + A × C | R | Replicates | No test of interaction | A, B, C, A × B, B × C, A × B × C |
| B × C | R + B × C | R | Replicates | No test of interaction | A, B, C, A × B, A × C, A × B × C |
| A × B × C | R + A × B × C | R | Replicates | No test of interaction | A, B, C, A × B, A × C, B × C |
| (ii) A fixed, B fixed, C random | | | | | |
| A | R + A × C + A | A × C | *ac*'s | Within levels of B | B, C, A × B, B × C |
| B | R + B × C + B | B × C | *bc*'s | Within levels of A | A, C, A × B, A × C |
| C | R + C | R | Replicates | Within levels of A and B | A, B, A × B, A × C, B × C, A × B × C |
| A × B | R + A × B × C + A × B | A × B × C | *abc*'s | No test of interaction | A, B, C, A × C, B × C |
| A × C | R + A × C | R | Replicates | No test of interaction | A, B, C, A × B, B × C, A × B × C |
| B × C | R + B × C | R | Replicates | No test of interaction | A, B, C, A × B, A × C, A × B × C |
| A × B × C | R + A × B × C | R | Replicates | No test of interaction | A, B, C, A × B, A × C, B × C |

| | | | | | |
|---|---|---|---|---|---|
| (iii) A fixed, B random, C random | | | | | |
| A | R + A × B × C + A × C + A × B + A | None – pooling required | – | – | – |
| B | R + B × C + B | B × C | *bc*'s | Within levels of A and C | A, C, A × B, A × C |
| C | R + B × C + C | B × C | *bc*'s | Within levels of A and B | A, B, A × B, A × C |
| A × B | R + A × B × C + A × B | A × B × C | *abc*'s | No test of interaction | A, B, C, A × C, B × C |
| A × C | R + A × B × C + A × C | A × B × C | *abc*'s | No test of interaction | A, B, C, A × B, B × C |
| B × C | R + B × C | R | Replicates | No test of interaction | A, B, C, A × B, A × C, A × B × C |
| A × B × C | R + A × B × C | R | Replicates | No test of interaction | A, B, C, A × B, A × C, B × C |
| (iv) A random, B random, C random | | | | | |
| A | R + A × B × C + A × C + A × B + A | None – pooling required | – | – | – |
| B | R + A × B × C + B × C + A × B + B | None – pooling required | – | – | – |
| C | R + A × B × C + B × C + A × C + C | None – pooling required | – | – | – |
| A × B | R + A × B × C + A × B | A × B × C | *abc*'s | No test of interaction | A, B, C, A × C, B × C |
| A × C | R + A × B × C + A × C | A × B × C | *abc*'s | No test of interaction | A, B, C, A × B, B × C |
| B × C | R + A × B × C + B × C | A × B × C | *abc*'s | No test of interaction | A, B, C, A × B, A × C |
| A × B × C | R + A × B × C | R | Replicates | No test of interaction | A, B, C, A × B, A × C, B × C |