Statistica Sinica **17**(2007), 1441-1455

EXACT TESTS FOR NEGLIGIBLE INTERACTION IN TWO-WAY ANALYSIS OF VARIANCE/COVARIANCE

Bin Cheng[1] and Jun Shao[2]

[1] *Columbia University and* [2] *University of Wisconsin*

Abstract: Statistical analysis of the interaction effect in a two-way analysis of a variance/covariance model (typically unbalanced) is often performed to provide some additional support for the assessment of the main effect of interest. For example, an analysis of treatment-by-center interaction, in addition to the assessment of treatment effect, is required by the International Conference on Harmonization Guidance in multicenter clinical studies. For this purpose, the usual test for interaction with zero interaction as the null hypothesis is not useful: rejecting such a null hypothesis does not tell us whether the interaction is large enough to affect the assessment of the main effect of interest; not rejecting such a null hypothesis does not provide any statistical assurance for ignoring interaction. We define a measure of interaction relative to the error variance, and derive some exact tests for testing negligible interaction (i.e., the relative interaction measure is smaller than a given margin) as the alternative hypothesis. If we conclude that interaction is negligible at a given significance level, we can then go on to assess the main effect. An example is presented for illustration.

Key words and phrases: Analysis of variance/covariance, mixed effects, size and power, test for negligible interaction, treatment-by-center interaction, unbalanced models.

1. Introduction

In two-way analysis of variance (ANOVA), an assessment of interaction between the two factors is required before the analysis of the main effect of interest. The traditional size α test for interaction is constructed with the null hypothesis of zero interaction, i.e., H_0: there is no interaction. This test can be used to detect interaction, but it is not useful if the intention is to show whether we can ignore the interaction effect, since we do not have enough statistical evidence (the power of the size α test is unknown) to support any conclusion when the null hypothesis of zero interaction is not rejected. In many applications, a conclusion of truly zero interaction may be unrealistic as well as unnecessary. Therefore, there is a need to advance the statistical theory for two-way linear models to include methodology for testing interaction with the alternative hypothesis being

the hypothesis that there is a negligible (not exactly zero) interaction. We call these tests for negligible interaction.

Our study is initially motivated by testing treatment-by-center interaction in muticenter clinical trials. Multicenter trials are commonly employed in clinical research. See the discussion in the International Conference on Harmonization (ICH) Guidance (*Statistical Principles for Clinical Trials*, known as E9) issued in 1998. A two-way ANOVA with treatment as one factor and center as the other is often adopted for multicenter trials. The following statements can be found in the ICH Guidance (E9):

> If positive treatment effects are found in a trial with appreciable numbers of subjects per center, there should generally be an exploration of the heterogeneity of treatment effects across centers, as this may affect the generalizability of the conclusions. Marked heterogeneity may be identified by graphical display of the results of individual centers or by analytical methods, such as a significance test of the treatment-by-center interaction.

Here, heterogeneity of treatment effects across centers can be referred to as the treatment-by-center interaction. If the treatment-by-center interaction is negligible, then we may assess the general treatment effect by averaging treatment effects from all centers. If the treatment-by-center interaction is large, however, we cannot obtain any general conclusion about the treatment effect. Clearly, negligible interaction is a desirable property, but cannot be established by a classical interaction test with the null hypothesis of zero treatment-by-center interaction.

The assessment of interaction in the context of multicenter clinical trials has been considered by several authors. Boos and Brownie (1992) constructed a rank-based test under a mixed effects model. Assessment of treatment-by-center interaction for censored data was considered by Peterson and George (1993) and Potthoff, Peterson and George (2001). Snapinn (1998) provided some insight in the interpretation of interaction effect. However, these authors all considered the null hypothesis of zero interaction and, thus, their tests are tests for detecting interaction, not for detecting negligible interaction. In the special case of two treatments, Gail and Simon (1985) introduced the concept of quantitative treatment-by-center interaction (one treatment is uniformly better than the other among all centers) and qualitative treatment-by-center interaction (one treatment is better than the other treatment in some centers but worse in the others). Ciminera, Heyse, Nguyen and Tukey (1993a,b) developed a push-back procedure for testing qualitative treatment-by-center interaction. However, their

alternative hypothesis is qualitative interaction and, hence, their proposed tests are for detecting qualitative interaction rather than the more desired quantitative interaction. To establish quantitative treatment-by-center interaction, we could switch the null and alternative hypotheses in Gail and Simon (1985) but a new test statistic needs to be derived.

The purpose of this article is to derive exact size α tests for negligible interaction in two-way fixed effects or mixed effects models. Details about models and the forms of hypotheses are given in Section 2. Exact tests for interaction are given in Section 3 for fixed effects and mixed effects ANOVA and analysis of covariance (ANCOVA) models. An example is given in Section 4 for illustration.

In most applications, tests for interaction are used together with tests for main effects. If interaction and some main effects are assessed simultaneously, then a union-intersection type test can be constructed using the results in this article for interaction, and existing results for testing main effects, e.g., Searle (1971) and Speed and Hocking (1976) for fixed effects models, and Khuri and Littell (1987), Gallo and Khuri (1990), Öfversten (1993), Christensen (1996) and Cheng and Shao (2006) for mixed effects models.

2. Hypotheses for Interaction

Let y_{ijk} denote the kth observation of the (i,j)th treatment in the two-way ANOVA model

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \quad i = 1, \ldots, I,\ j = 1, \ldots, J,\ k = 1, \ldots, n_{ij}, \quad (1)$$

where the ε_{ijk} are independent and identically distributed (i.i.d.) $N(0, \sigma^2)$, and μ and σ^2 are unknown parameters. In traditional two-way ANOVA models, the main effects α_i and β_j and the interaction effects γ_{ij} are treated as fixed unknown parameters. The use of mixed effects linear models has received a great deal of attention in recent years. In multicenter clinical trials, the ICH Guidance (E9) states that "mixed models may also be used to explore the heterogeneity of the treatment effects", and "these models consider center and treatment-by-center effects to be random, and are especially relevant when the number of sites is large". Therefore, we also consider mixed effects models in which the α_i's are fixed unknown parameters but the β_j's are i.i.d. random effects distributed as $N(0, \sigma_\beta^2)$, the γ_{ij}'s are i.i.d. random effects distributed as $N(0, \sigma_\gamma^2)$, and the β_j's, γ_{ij}'s and ε_{ijk}'s are mutually independent. The following typical constraints are imposed: $\bar{\alpha}_{\cdot} = 0$, $\bar{\beta}_{\cdot} = 0$ (when the β_j are fixed effects), $\bar{\gamma}_{i\cdot} = 0$, and $\bar{\gamma}_{\cdot j} = 0$ (when the γ_{ij} are fixed effects). Throughout the paper, for any given variable x,

$\bar{x}$ denotes an average and a dot is used in the subscript to denote averaging over the indicated subscript, e.g., $\bar{x}_{\cdot} = I^{-1}\sum_{i=1}^{I} x_i$ and $\bar{x}_{i\cdot} = J^{-1}\sum_{j=1}^{J} x_{ij}$.

We now define a quantitative measure of interaction, δ. Because of the existence of the random error ε_{ijk}, it is reasonable to consider a measure relative to the error variance σ^2. Under the mixed effects ANOVA model, it is natural to use $\delta = \sigma_\gamma^2/\sigma^2$.

Let δ_0 be a tolerance margin for interaction, i.e., the interaction effect is practically negligible if and only if $\delta < \delta_0$. The use of a tolerance margin in assessing treatment effects is not uncommon in clinical studies; for example, noninferiority and equivalence margins are used in noninferiority cancer trials (e.g., Laster and Johnson (2003)) and bioequivalence studies (e.g., FDA (2001)). In multicenter clinical trials, δ_0 is determined by a regulatory agency or by the experimenters based on historical information and/or their understanding of the nature of the study. In any case, δ_0 has to be chosen prior to the study to ensure a fair evaluation. Since δ measures interaction and the main purpose of many studies is to assess main effects, $\sqrt{\delta_0}$ may be chosen to be a fraction (say 20% or 30%) of a meaningful margin for main effects relative to the error standard deviation σ. In the stage of developing a study protocol for a clinical research, a sample size analysis is typically performed to ensure that the power of the test for treatment effects is approximately equal to a desired level for detecting treatment effects when a measure of treatment effects is greater than or equal to a clinically meaningful margin. Then, $\sqrt{\delta_0}$ can be chosen to be 20% or 30% of the clinical meaningful margin for treatment effects.

For a chosen δ_0, we consider the following hypotheses for testing negligible interaction:

$$H_0 : \delta \geq \delta_0 \quad \text{versus} \quad H_1 : \delta < \delta_0. \tag{2}$$

In a multicenter clinical trial, if H_0 is rejected at a given significance level α, then we have statistical evidence that the heterogeneity of treatment effects among all centers is negligible, and consequently the treatment effect, if significant, can be interpreted without being misleading.

Finding a suitable measure for interaction is more difficult under fixed effects ANOVA models. When model (1) is balanced in the sense that $n_{ij} = n$ for all i and j, we can define δ as

$$\delta = \frac{1}{\sigma^2 IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\gamma_{ij}^2. \tag{3}$$

Note that, under mixed effects models, the expectation of the right hand side of (3) is exactly σ_γ^2/σ^2. The usual sum of squares for interaction in textbooks is

$$\text{SSAB} = n\sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdots})^2, \tag{4}$$

which is distributed as σ^2 times a noncentral chi-square random variable with degree of freedom $(I-1)(J-1)$ and the noncentrality parameter $N\delta$, N being the total number of observations ($N = nIJ$ in the balanced case). Thus, the use of δ in (4) as a measure for interaction is consistent with the use of SSAB as a statistic for assessing the interaction effect.

Under unbalanced fixed effects models, however, an exact test for (2) with δ defined by (3) is not available if we consider test statistics that are quadratic forms of the cell mean vector

$$\bar{\mathbf{y}} = (\bar{y}_{11\cdot}, \ldots, \bar{y}_{I1\cdot}, \ldots, \bar{y}_{1J\cdot}, \ldots, \bar{y}_{IJ\cdot})'$$

(such as the SSAB defined in (4)). This is because, under an unbalanced model, a necessary condition for a nonzero quadratic form $\bar{\mathbf{y}}'\mathbf{D}\bar{\mathbf{y}}/\sigma^2$ to have a noncentral chi-square distribution with noncentrality parameter $\boldsymbol{\mu}'\mathbf{D}\boldsymbol{\mu}/\sigma^2$ is that $\mathbf{D}$ depends on the n_{ij}'s, where $\boldsymbol{\mu} = E(\bar{\mathbf{y}})$. Note that $\bar{\mathbf{y}}'\mathbf{D}\bar{\mathbf{y}}/\sigma^2$ has a (noncentral) chi-square distribution if and only if $\mathbf{D}\boldsymbol{\Lambda}\mathbf{D} = \mathbf{D}$, where

$$\boldsymbol{\Lambda} = \text{diag}(n_{11}^{-1}, \ldots, n_{I1}^{-1}, \ldots, n_{1J}^{-1}, \ldots, n_{IJ}^{-1}) = \sigma^{-2}\text{Var}\,(\bar{\mathbf{y}}). \tag{5}$$

Write $\mathbf{D} = (\mathbf{D}_{11}, \ldots, \mathbf{D}_{IJ})$, where $\mathbf{D}_{ij}$'s are the columns of $\mathbf{D}$. If $\mathbf{D}$ does not depend on n_{ij}, then by taking derivatives of both sides of $\mathbf{D}\boldsymbol{\Lambda}\mathbf{D} = \mathbf{D}$ with respect to n_{ij}^{-1}, we conclude that $\mathbf{D}_{ij} = \mathbf{0}$. Therefore, if $\mathbf{D}$ does not depend on n_{ij}'s, we must have $\mathbf{D} = \mathbf{0}$.

Because constructing exact tests not using quadratic forms of the cell mean vector $\bar{\mathbf{y}}$ is difficult, we consider a different interaction measure for unbalanced fixed effects models.

For assessing interaction under unbalanced models, the following statistic is commonly used to replace SSAB:

$$R(\gamma|\mu, \alpha, \beta) = R(\mu, \alpha, \beta, \gamma) - R(\mu, \alpha, \beta),$$

where $R(\mu, \alpha, \beta, \gamma)$ is the reduction in the total sum of squares due to fitting model (1), and $R(\mu, \alpha, \beta)$ is the reduction in the total sum of squares due to model (1) without γ-terms (or $\gamma_{ij} = 0$ for all i and j). When the model is balanced, $R(\gamma|\mu, \alpha, \beta)$ is the same as SSAB in (4). Detailed discussions of the use of $R(\ )$-notation can be found in Searle (1971) and Speed and Hocking (1976).

After some algebra, it turns out that $R(\gamma|\mu,\alpha,\beta) = \bar{\mathbf{y}}'\mathbf{L}(\mathbf{L}'\mathbf{\Lambda}\mathbf{L})^{-1}\mathbf{L}'\bar{\mathbf{y}}$, where $\mathbf{\Lambda}$ is given by (5),

$$\mathbf{L} = \begin{pmatrix} \mathbf{J}'_{J-1} \\ -\mathbf{I}_{J-1} \end{pmatrix} \otimes \begin{pmatrix} \mathbf{J}'_{I-1} \\ -\mathbf{I}_{I-1} \end{pmatrix},$$

$\mathbf{I}_a$ denotes the identity matrix of order a, $\mathbf{J}_b$ denotes the vector of ones of order b, and $\otimes$ denotes the Kronecker product for matrices. Since $\mathbf{L}(\mathbf{L}'\mathbf{\Lambda}\mathbf{L})^{-1}\mathbf{L}'\mathbf{\Lambda}\mathbf{L}(\mathbf{L}'\mathbf{\Lambda}\mathbf{L})^{-1}\mathbf{L}' = \mathbf{L}(\mathbf{L}'\mathbf{\Lambda}\mathbf{L})^{-1}\mathbf{L}'$, $R(\gamma|\mu,\alpha,\beta)/\sigma^2$ has the noncentral chi-square distribution with degree of freedom $(I-1)(J-1)$ (the rank of $\mathbf{L}(\mathbf{L}'\mathbf{\Lambda}\mathbf{L})^{-1}\mathbf{L}'$) and noncentrality parameter $\lambda = \sigma^{-2}\boldsymbol{\mu}'\mathbf{L}(\mathbf{L}'\mathbf{\Lambda}\mathbf{L})^{-1}\mathbf{L}'\boldsymbol{\mu} = \sigma^{-2}\boldsymbol{\gamma}'\mathbf{L}(\mathbf{L}'\mathbf{\Lambda}\mathbf{L})^{-1}\mathbf{L}'\boldsymbol{\gamma}$ (since $\mathbf{L}'\boldsymbol{\mu} = \mathbf{L}'\boldsymbol{\gamma}$), where $\boldsymbol{\gamma} = (\gamma_{11},\ldots,\gamma_{I1},\ldots,\gamma_{1J},\ldots,\gamma_{IJ})'$. This suggests the following measure of interaction:

$$\delta = \frac{\lambda}{N} = \frac{\boldsymbol{\gamma}'\mathbf{L}(\mathbf{L}'\bar{n}_{\cdot\cdot}\mathbf{\Lambda}\mathbf{L})^{-1}\mathbf{L}'\boldsymbol{\gamma}}{\sigma^2 IJ}, \tag{6}$$

which reduces to the δ in (3) in the balanced case where $\bar{n}_{\cdot\cdot}\mathbf{\Lambda} = \mathbf{I}_{IJ}$.

Unlike the interaction measure δ in (3), δ in (6) depends on the sample sizes n_{ij}, although both of them are quadratic forms of $\boldsymbol{\gamma}$ and are identical in the balanced case. As the previous discussion indicated, however, it is impossible in an unbalanced model to derive a reasonable interaction measure not depending on n_{ij}'s and an associated exact test based on a quadratic form of $\bar{\mathbf{y}}$. Furthermore, it is not uncommon that hypotheses depending on n_{ij}'s are tested under unbalanced models; for example, the well-known type II analysis for treatment effects in an unbalanced ANOVA model considers a null hypothesis depending on n_{ij}'s (Speed and Hocking (1976)).

It should be noted that there may be other reasonable measures of interaction. The δ we choose allows us to derive the exact statistical tests, given in the next section. Other choices of δ typically yield conservative tests. Also, we consider a single aggregated measure of interaction. Under mixed effects models, the γ_{ij} are assumed to be i.i.d. so that $\delta = \sigma^2_\gamma/\sigma^2$ is reasonable. Under fixed effects models, the use of δ defined by (3) or (6) regards the γ_{ij}'s as exchangeable, as in the random γ_{ij} case. It is tempting to consider multiple non-aggregated measures of interaction. However, statistical analysis based on such measures are usually difficult and exact tests with reasonable power are typically not available.

3. Exact Tests

Some exact tests for (2) are given in this section for several different situations. If the null hypothesis in (2) is rejected, i.e., interaction is negligible, we can then test main effects using available results in the literature (see the discussion

in Section 1). If the null hypothesis in (2) is not rejected, applying tests for main effects may result in misleading conclusions.

3.1. Fixed effects ANOVA models

Consider model (1) with fixed effects. Define

$$\text{SSE} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n_{ij}}(y_{ijk} - \bar{y}_{ij\cdot})^2, \tag{7}$$

$$F_{\text{AB}} = \frac{R(\gamma|\mu,\alpha,\beta)[(I-1)(J-1)]^{-1}}{\text{SSE}(N-IJ)^{-1}}. \tag{8}$$

Note that SSE and $R(\gamma|\mu,\alpha,\beta)$ are independent. SSE/σ^2 is distributed as chi-square with $N-IJ$ degrees of freedom. It is shown in the previous section that $R(\gamma|\mu,\alpha,\beta)/\sigma^2$ is distributed as the noncentral chi-square distribution with degrees of freedom $(I-1)(J-1)$ and the noncentrality parameter $N\delta$, where δ is given by (6). Thus, F_{AB} is noncentral F with degrees of freedom $(I-1)(J-1)$ and $N-IJ$ and noncentrality parameter $N\delta$, and $P\{F_{\text{AB}} < t\}$ is a decreasing function of δ for any fixed t and N (a property of noncentral F-distributions). Hence, if $F_{(I-1)(J-1),IJ(n-1),\alpha}(N\delta_0)$ is the αth quantile of the noncentral F-distribution with degrees of freedom $(I-1)(J-1)$ and $N-IJ$ and the noncentrality parameter $N\delta_0$, then

$$\begin{aligned}&\sup_{\delta\geq\delta_0} P_\delta\{F_{\text{AB}} < F_{(I-1)(J-1),IJ(n-1),\alpha}(N\delta_0)\}\\ &\quad = P_{\delta_0}\{F_{\text{AB}} < F_{(I-1)(J-1),IJ(n-1),\alpha}(N\delta_0)\} = \alpha.\end{aligned}$$

Consequently, for testing hypotheses (2), a size α test rejects H_0 if and only if

$$F_{\text{AB}} < F_{(I-1)(J-1),N-IJ,\alpha}(N\delta_0). \tag{9}$$

The power of this test is $P_\delta\{F_{\text{AB}} < F_{(I-1)(J-1),IJ(n-1),\alpha}(N\delta_0)\}$ with $\delta < \delta_0$, which is larger than α by the monotone property of the noncentral F-distribution. The cumulative distribution function and quantiles of the noncentral F-distribution can be determined using statistical software. In SAS, for example, CDF("F", x, a, b, c) is used to compute the cumulative distribution evaluated at x and FINV(p, a, b, c) is used to compute the pth quantile, where (a, b) are the degrees of freedom and c is the noncentrality parameter of the noncentral F-distribution.

For some balanced models, where $R(\gamma|\mu,\alpha,\beta)$ reduces to SSAB in (4) and δ reduces to the one defined by (3), the power of test rule (9) is computed and plotted in Figure 1. The test rule (9) is quite powerful for small sample sizes. For example, the first panel of Figure 1 indicates that with 100 subjects equally

randomized to 2 treatments and 5 centers (i.e., $I = 2$, $J = 5$, $n = 10$) and $\sqrt{\delta_0}$ chosen as 0.5, we have an 80% power at the negligible interaction $\sqrt{\delta} = 0.25$.

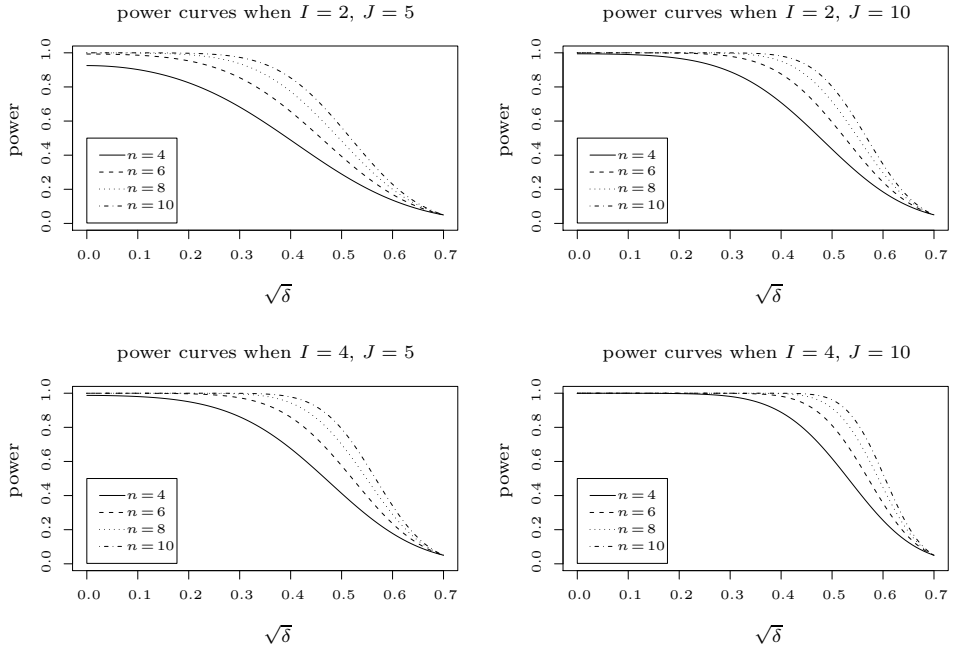

Figure 1. Power of Test Rule (9). Balanced Fixed Effects Models With $\sqrt{\delta_0} = 0.5$. Rejection Region: $F_{\mathrm{AB}} < F_{(I-1)(J-1),IJ(n-1),0.05}(0.25nIJ)$.

3.2. Mixed effects ANOVA models

Under mixed effects ANOVA models, $\delta = \sigma_\gamma^2/\sigma^2$ can be used as the measure of interaction in (2) for both balanced and unbalanced models. However, the derivation of an exact test for (2) under mixed effects models is not so simple. Thomsen (1975) derived an exact test for testing the hypotheses $\delta \le \delta_0$ versus $\delta > \delta_0$ under (1), with α_i's assumed to be normally distributed random effects. His test statistic does not have an explicit form since it involves determination of an orthogonal transformation which simultaneously diagonalizes two semi-positive definite matrices. Although Thomsen's method may be modified to test (2), some additional restrictions on the choice of the orthogonal matrix are required when the α_i's are fixed effects, which leads to an even more computationally complicated procedure. We derive two exact tests that are explicit and simple.

The first test statistic is the F_{AB} defined in (8). Under mixed effects models, $R(\gamma|\mu,\alpha,\beta)/\sigma^2$ is no longer chi-square distributed, although it is independent of SSE defined by (7). Hence test rule (9) cannot be used because the F-percentile is not the right percentile to use. However, F_{AB} can still be used if its correct

percentile can be found when $\delta = \delta_0$. Since $\mathbf{L}'\bar{\mathbf{y}}$ is normally distributed with mean $\mathbf{0}$ and

$$\text{Var}\,(\mathbf{L}'\bar{\mathbf{y}}) = \mathbf{L}'(\sigma^2\mathbf{\Lambda} + \sigma_\gamma^2\mathbf{I}_{IJ} + \sigma_\beta^2\mathbf{I}_J \otimes \mathbf{J}_I\mathbf{J}_I')\mathbf{L} = \sigma^2\mathbf{L}'(\mathbf{\Lambda} + \delta\mathbf{I}_{IJ})\mathbf{L},$$

$\mathbf{L}'\bar{\mathbf{y}}/\sigma$ has the same distribution as $[\mathbf{L}'(\mathbf{\Lambda}+\delta\mathbf{I}_{IJ})\mathbf{L}]^{1/2}\mathbf{z}$, where $\mathbf{z}$ is a multivariate normal random vector with mean $\mathbf{0}$ and covariance matrix $\mathbf{I}_{(I-1)(J-1)}$. Note that $R(\gamma|\mu,\alpha,\beta)/\sigma^2 = \bar{\mathbf{y}}'\mathbf{L}(\mathbf{L}'\mathbf{\Lambda}\mathbf{L})^{-1}\mathbf{L}'\bar{\mathbf{y}}$. Then, F_{AB} has the same distribution as

$$\begin{aligned} G(\delta) &= \frac{\mathbf{z}'[\mathbf{L}'(\mathbf{\Lambda}+\delta\mathbf{I}_{IJ})\mathbf{L}]^{\frac{1}{2}}(\mathbf{L}'\mathbf{\Lambda}\mathbf{L})^{-1}[\mathbf{L}'(\mathbf{\Lambda}+\delta\mathbf{I}_{IJ})\mathbf{L}]^{\frac{1}{2}}\mathbf{z}[(I-1)(J-1)]^{-1}}{\chi^2_{N-IJ}(N-IJ)^{-1}} \\ &= \frac{\sum_{k=1}^{(I-1)(J-1)}\lambda_k(\delta)\chi^2_{(k)}[(I-1)(J-1)]^{-1}}{\chi^2_{N-IJ}(N-IJ)^{-1}}, \end{aligned} \tag{10}$$

where χ^2_{N-IJ} is a central chi-square random variable with degree of freedom $N-IJ$ independent of $\mathbf{z}$, $\chi^2_{(k)}$, $k = 1,\ldots,(I-1)(J-1)$, are independent central chi-square random variables with degree of freedom 1 independent of χ^2_{N-IJ}, and $\lambda_k(\delta)$, $k = 1,\ldots,(I-1)(J-1)$, are the eigenvalues of the matrix $[\mathbf{L}'(\mathbf{\Lambda}+\delta\mathbf{I}_{IJ})\mathbf{L}]^{1/2}(\mathbf{L}'\mathbf{\Lambda}\mathbf{L})^{-1}[\mathbf{L}'(\mathbf{\Lambda}+\delta\mathbf{I}_{IJ})\mathbf{L}]^{1/2}$. Although the distribution of $G(\delta)$ is of an unfamiliar form, $P\{G(\delta) < t\}$ is decreasing in δ for any fixed t because $G(\delta)$ in (10) is increasing in δ. Also, the distribution of $G(\delta)$ is known when δ is known. Consequently, a test of size α for (2) rejects H_0 if and only if

$$F_{\text{AB}} < G_\alpha(\delta_0), \tag{11}$$

where $G_\alpha(\delta)$ is the αth quantile of the distribution of $G(\delta)$ in (10). Note that test rule (11) is the same as test rule (9) except that the F-percentile is replaced by $G_\alpha(\delta_0)$. The percentile $G_\alpha(\delta_0)$ has to be numerically computed for each set of sample sizes. For example, we can apply the Monte Carlo simulation method using (10).

Our second exact test is motivated by the search for a simple F-test. It suffices to find a quadratic form that is independent of SSE and has a chi-square distribution when $\delta = \delta_0$. Consider the quadratic form

$$R_{\delta_0}(\gamma|\mu,\alpha,\beta) = \bar{\mathbf{y}}'\mathbf{L}[\mathbf{L}'(\mathbf{\Lambda}+\delta_0\mathbf{I}_{IJ})\mathbf{L}]^{-\mathbf{1}}\mathbf{L}'\bar{\mathbf{y}},$$

which reduces to $R(\gamma|\mu,\alpha,\beta)$ when $\delta_0 = 0$. Since $\bar{y}_{ij\cdot}$ and $y_{ijk} - \bar{y}_{ij\cdot}$ are independent, $R_{\delta_0}(\gamma|\mu,\alpha,\beta)$ and SSE defined in (7) are independent. Define

$$F_{\text{AB}}(\delta_0) = \frac{R_{\delta_0}(\gamma|\mu,\alpha,\beta)[(I-1)(J-1)]^{-1}}{\text{SSE}(N-IJ)^{-1}}.$$

Then, it can be shown that (i) the distribution of $F_{\mathrm{AB}}(\delta_0)$ is the same as that of

$$H(\delta) = \frac{\mathbf{z}'[\mathbf{L}'(\mathbf{\Lambda} + \delta \mathbf{I}_{IJ})\mathbf{L}]^{\frac{1}{2}}[\mathbf{L}'(\mathbf{\Lambda} + \delta_0 \mathbf{I}_{IJ})\mathbf{L}]^{-1}[\mathbf{L}'(\mathbf{\Lambda} + \delta \mathbf{I}_{IJ})\mathbf{L}]^{\frac{1}{2}}\mathbf{z}}{(I-1)(J-1)\chi^2_{N-IJ}(N-IJ)^{-1}};$$

(ii) $P\{H(\delta) < t\}$ is decreasing in δ for any fixed t; (iii) when $\delta = \delta_0$ (the common boundary of the hypotheses in (2)), $H(\delta_0)$ has the central F-distribution with degrees of freedom $(I-1)(J-1)$ and $N-IJ$. Consequently, a test of size α for (2) rejects H_0 if and only if

$$F_{\mathrm{AB}}(\delta_0) < F_{(I-1)(J-1),N-IJ,\alpha}. \tag{12}$$

Under a balanced mixed effects model, however, the two tests (11) and (12) are equivalent. In fact, when $n_{ij} = n$ for all i and j, $R_{\delta_0}(\gamma|\mu,\alpha,\beta) = \mathrm{SSAB}/(1+\delta_0 n)$, where SSAB is given in (4), and the test rule in (12) becomes $F_{\mathrm{AB}}/(1+\delta_0 n) < F_{(I-1)(J-1),N-IJ,\alpha}$, where F_{AB} is defined by (8). On the other hand, $R(\gamma|\mu,\alpha,\beta) = \mathrm{SSAB}$ and the test rule (11) becomes $F_{\mathrm{AB}} < G_\alpha(\delta_0) = [(n^{-1}+\delta_0)/(n^{-1})]F_{(I-1)(J-1),N-IJ,\alpha}$. Therefore, the tests are the same. For some balanced models, the power of test rule (11) is computed and plotted in Figure 2. It can be seen that, compared to its fixed effect counterpart, the power under a mixed model is typically lower, which indicates that it is generally harder to test for negligible interaction under a mixed model.

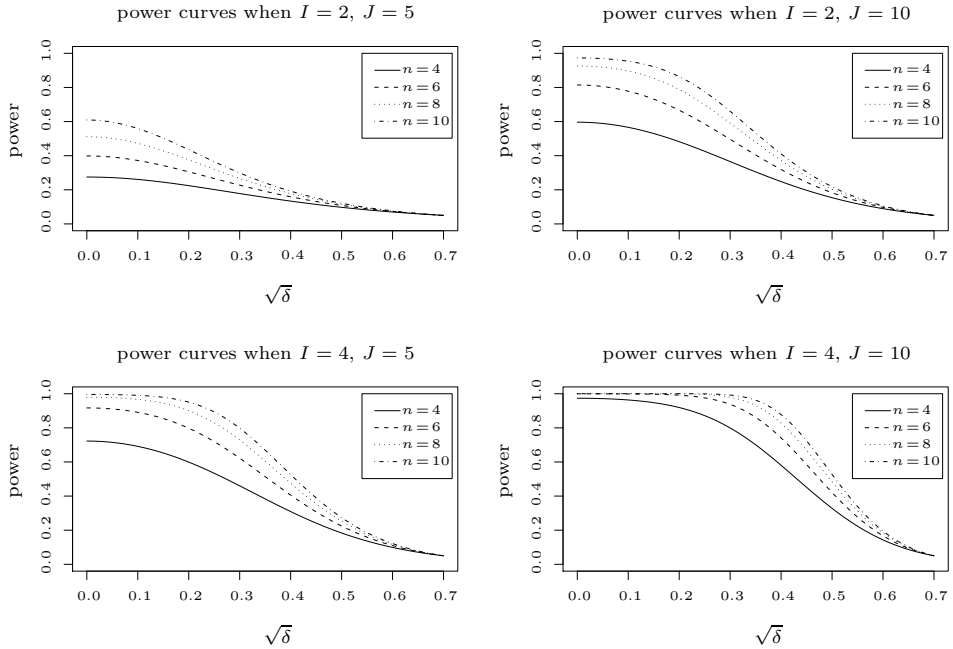

Figure 2. Power of Test Rule (11) or (12). Balanced Mixed Effects Models With $\sqrt{\delta_0} = 0.5$. Rejection Region: $F_{\mathrm{AB}} < \frac{n^{-1}+0.25}{n^{-1}} F_{(I-1)(J-1),IJ(n-1),0.05}$.

Under an unbalanced model, the tests defined by (11) and (12) are generally different. Test rule (11) uses the well-known statistic F_{AB} that is also used in fixed effects ANOVA models (Section 3.1), and its value is directly available from any major statistical software. The critical value $G_\alpha(\delta_0)$ in (11) is not a percentile of a familiar distribution and a numerical computation of $G_\alpha(\delta_0)$ is required for every given set of n_{ij}'s. On the other hand test rule (12) is an F-test, but it uses a statistic that is not familiar to most statisticians, and is not directly available from any major software.

For an unbalanced design with 4 treatments and 5 centers and n_{ij}'s given in Table 2, we compared by simulation the test rules (11) and (12) in terms of their power. In the simulation, $\sigma^2 = 1$ and, hence, $\sqrt{\delta} = \sigma_\gamma$. For $\sqrt{\delta_0} = 0.5$ and 0.7, the critical values $G_{0.05}(\delta_0)$ in (11) computed by Monte Carlo are 0.972 and 1.486, respectively. The estimated powers based on 5,000 simulations are shown in Table 1, indicating that the two tests have very similar power (test rule (11) is slightly better), the choice between them is perhaps best decided by computational convenience.

Table 1. Power Comparison of Tests (11) and (12) with $\sigma = 1.0$.

	$\sqrt{\delta_0} = 0.5$		$\sqrt{\delta_0} = 0.7$	
$\sqrt{\delta} = \sigma_\gamma$	test (11)	test (12)	test (11)	test (12)
0.05	0.4998	0.4864	0.8494	0.8478
0.10	0.4732	0.4632	0.8224	0.8134
0.25	0.2900	0.2824	0.6452	0.6436

Table 2. Summary Statistics.

	Treatment (i)											
	1			2			3			4		
Center(j)	n_{ij}	$\bar{y}_{ij\cdot}$	$\bar{w}_{ij\cdot}$	n_{ij}	$\bar{y}_{ij\cdot}$	$\bar{w}_{ij\cdot}$	n_{ij}	$\bar{y}_{ij\cdot}$	$\bar{w}_{ij\cdot}$	n_{ij}	$\bar{y}_{ij\cdot}$	$\bar{w}_{ij\cdot}$
1	7	29.02	29.09	6	32.37	33.51	4	33.77	32.51	4	36.14	31.88
2	6	34.47	32.18	5	34.26	33.85	6	34.42	30.81	6	40.10	35.02
3	5	35.10	34.87	5	33.28	32.41	5	32.02	29.68	5	37.97	32.53
4	7	36.04	34.14	4	31.62	29.83	4	36.50	33.95	5	38.92	33.62
5	7	31.49	32.25	6	31.29	32.56	4	32.42	32.06	4	34.88	32.45
	Under ANOVA Model (1)						Under ANCOVA Model (13)					
SSE	187.38						71.24					

3.3. ANCOVA models

In some multicenter clinical trials there are covariates, such as patients' demographic variables, medical history, and other baseline characteristics, and some

center characteristics. Including covariates that are related to the response variable reduces error variability and, hence, increases the power of various tests. In some cases non-negligible interaction is caused by the difference in patients' demographics, medical conditions or departures from the protocol (Snapinn (1998)). Incoporation of these inhomogeneity variables in the model may help to reduce the interaction.

Including covariates in the analysis leads to the following popular two-way ANCOVA model:

$$y_{ijk} = \mu+\alpha_i+\beta_j+\gamma_{ij}+\boldsymbol{\eta}'\mathbf{w}_{ijk}+\varepsilon_{ijk},\ i = 1,\ldots,I,\ j = 1,\ldots,J,\ k = 1,\ldots,n_{ij}, \quad (13)$$

where $\boldsymbol{\eta}$ is a q-dimensional unknown parameter vector, the $\mathbf{w}_{ijk}$'s are q-dimensional covariate vectors, and the ε_{ijk}'s are $N(0,\sigma^2)$ random errors. The assumptions on the α_i's β_j's and γ_{ij}'s in model (13) are the same as those under model (1). Let $\mathbf{y}$ be the vector formed by listing y_{ijk} in the order j, i and k. Then (13) can be written in the matrix form as $\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \mathbf{W}\boldsymbol{\eta} + \boldsymbol{\varepsilon}$, where $\mathbf{X}$ is the usual design matrix for the two-way ANOVA model, $\mathbf{W}$ is the design matrix containing $\mathbf{w}_{ijk}$'s, $\boldsymbol{\theta} = (\mu, \alpha_1, \ldots, \alpha_I, \beta_1, \ldots, \beta_J, \gamma_{11}, \ldots, \gamma_{IJ})'$, and $\boldsymbol{\varepsilon}$ is the error vector. The least squares estimator of $\boldsymbol{\eta}$ is $\hat{\boldsymbol{\eta}} = [\mathbf{W}'(\mathbf{I}-\mathbf{P}_{\mathbf{X}})\mathbf{W}]^{-1}\mathbf{W}'(\mathbf{I}-\mathbf{P}_{\mathbf{X}})\mathbf{y}$, where $\mathbf{P}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'$. Define $u_{ijk} = y_{ijk} - \hat{\boldsymbol{\eta}}'\mathbf{w}_{ijk}$ and $\bar{\mathbf{u}} = (\bar{u}_{11\cdot}, \ldots, \bar{u}_{I1\cdot}, \ldots, \bar{u}_{1J\cdot}, \ldots, \bar{u}_{IJ\cdot})'$, which can be called the adjusted cell mean vector. Let

$$\mathrm{SSE} = \mathbf{y}'(\mathbf{I}-\mathbf{P}_{(\mathbf{X},\mathbf{W})})\mathbf{y} = \boldsymbol{\varepsilon}'(\mathbf{I}-\mathbf{P}_{(\mathbf{X},\mathbf{W})})\boldsymbol{\varepsilon}, \quad (14)$$

where $\mathbf{P}_{(\mathbf{X},\mathbf{W})}$ is the same as $\mathbf{P}_{\mathbf{X}}$ with $\mathbf{X}$ replaced by the matrix $(\mathbf{X},\mathbf{W})$. Note that $\hat{\boldsymbol{\eta}} = \boldsymbol{\eta}+[\mathbf{W}'(\mathbf{I}-\mathbf{P}_{\mathbf{X}})\mathbf{W}]^{-1}\mathbf{W}'(\mathbf{I}-\mathbf{P}_{\mathbf{X}})\boldsymbol{\varepsilon}$. Since $\mathbf{W}'(\mathbf{I}-\mathbf{P}_{\mathbf{X}})(\mathbf{I}-\mathbf{P}_{(\mathbf{X},\mathbf{W})}) = \mathbf{W}'(\mathbf{I}-\mathbf{P}_{(\mathbf{X},\mathbf{W})}) = \mathbf{0}$, $\hat{\boldsymbol{\eta}}$ and $(\mathbf{I}-\mathbf{P}_{(\mathbf{X},\mathbf{W})})\boldsymbol{\varepsilon}$ are independent. Furthermore, $\bar{\varepsilon}_{ij\cdot}$ is independent of $(\mathbf{I}-\mathbf{P}_{(\mathbf{X},\mathbf{W})})\boldsymbol{\varepsilon}$ since $(\mathbf{I}-\mathbf{P}_{(\mathbf{X},\mathbf{W})})\mathbf{P}_{\mathbf{X}} = \mathbf{P}_{\mathbf{X}} - \mathbf{P}_{(\mathbf{X},\mathbf{W})}\mathbf{P}_{\mathbf{X}} = \mathbf{0}$. From $\bar{u}_{ij\cdot} = \mu + \alpha_i + \beta_j + \gamma_{ij} + (\boldsymbol{\eta}-\hat{\boldsymbol{\eta}})'\bar{\mathbf{w}}_{ij\cdot} + \bar{\varepsilon}_{ij\cdot}$, we conclude that SSE and $\bar{\mathbf{u}}$ are independent.

Note that $\bar{\mathbf{u}}$ is normally distributed with covariance matrix $\mathrm{Var}(\bar{\mathbf{u}}) = (\mathbf{V}+\boldsymbol{\Lambda})\sigma^2$ for fixed effects models and $\mathrm{Var}(\bar{\mathbf{u}}) = (\mathbf{I}_J \otimes (\mathbf{J}_I\mathbf{J}_I'))\sigma_\beta^2 + \mathbf{I}_{IJ}\sigma_\gamma^2 + (\mathbf{V}+\boldsymbol{\Lambda})\sigma^2$ for mixed effects models, where $\mathbf{V} = \bar{\mathbf{w}}'[\mathbf{W}'(\mathbf{I}-\mathbf{P}_{\mathbf{X}})\mathbf{W}]^{-1}\bar{\mathbf{w}}$ and $\bar{\mathbf{w}} = (\bar{\mathbf{w}}_{11\cdot}, \ldots, \bar{\mathbf{w}}_{I1\cdot}, \ldots, \bar{\mathbf{w}}_{1J\cdot}, \ldots, \bar{\mathbf{w}}_{IJ\cdot})'$. It can be seen that results in the previous sections are derived based on the key fact that $R(\gamma|\mu,\alpha,\beta)$ and $R_{\delta_0}(\gamma|\mu,\alpha,\beta)$ are quadratic functions of $\bar{\mathbf{y}}$ independent of SSE. Under the ANCOVA model, the adjusted cell mean vector $\bar{\mathbf{u}}$ plays the role of $\bar{\mathbf{y}}$. As an illustration, we consider testing (2) under (13). We modify the test statistic in (12) as

$$F_{\mathrm{AB}}(\delta_0) = \frac{\bar{\mathbf{u}}'\mathbf{L}[\mathbf{L}'(\boldsymbol{\Lambda}+\mathbf{V}+\delta_0\mathbf{I}_{IJ})\mathbf{L}]^{-1}\mathbf{L}'\bar{\mathbf{u}}[(I-1)(J-1)]^{-1}}{\mathrm{SSE}(N-IJ-q)^{-1}}, \quad (15)$$

where SSE is defined in (14). Since under $\delta = \delta_0$, $\mathrm{Var}\,(\mathbf{L}'\bar{\mathbf{u}}) = \sigma^2\mathbf{L}'(\sigma^2(\mathbf{\Lambda} + \mathbf{V}) + \delta_0\mathbf{I}_{IJ})\mathbf{L}$, we conclude that $\bar{\mathbf{u}}'\mathbf{L}[\mathbf{L}'(\mathbf{\Lambda} + \mathbf{V} + \delta_0\mathbf{I}_{IJ})\mathbf{L}]^{-1}\mathbf{L}'\bar{\mathbf{u}}/\sigma^2 \sim \chi^2_{(I-1)(J-1)}$ independent of $\mathrm{SSE}/\sigma^2 \sim \chi^2_{N-IJ-q}$ since we have shown that $\bar{\mathbf{u}}$ and SSE are independent. Therefore, $F_{\mathrm{AB}}(\delta_0)$ defined in (15) is distributed as $F_{(I-1)(J-1),N-IJ-q}$, and the test rule is modified accordingly as $F_{\mathrm{AB}} < F_{(I-1)(J-1),N-IJ-q,\alpha}$. Similarly, the results in Sections 3.1 and 3.2 still hold under two-way ANCOVA models with the following modifications: (i) $\bar{\mathbf{y}}$ is replaced by $\bar{\mathbf{u}}$; (ii) SSE is defined by (14); (iii) the degree of freedom for SSE, which appears as denominator degree of freedom in some tests, is changed from $N - IJ$ to $N - IJ - q$ due to estimation of $\boldsymbol{\eta}$; (iv) δ in (3) or (6) is replaced by $\gamma'\mathbf{L}[\mathbf{L}'\bar{n}_{\cdot\cdot}(\mathbf{\Lambda} + \mathbf{V})\mathbf{L}]^{-1}\mathbf{L}'\boldsymbol{\gamma}/\sigma^2 IJ$; and (v) the matrix $\mathbf{\Lambda}$ in any statistic is replaced by $\mathbf{\Lambda} + \mathbf{V}$. For example, the statistic F_{AB} in (8) is defined as

$$\frac{\bar{\mathbf{u}}'\mathbf{L}[\mathbf{L}'(\mathbf{\Lambda} + \mathbf{V})\mathbf{L}]^{-1}\mathbf{L}'\bar{\mathbf{u}}[(I-1)(J-1)]^{-1}}{\mathbf{y}'(\mathbf{I} - \mathbf{P}_{(\mathbf{X},\mathbf{W})})\mathbf{y}(N - IJ - q)^{-1}}.$$

4. An Example

An example from a Phase II clinical trial is presented in this section. Since the compound is currently only at Phase II development stage, all background information is concealed. The trial was conducted in five centers with a total of 105 patients, randomized to one of the four treatment groups. The statistical model specified in the study protocol for a continuous primary response variable is a two-way ANCOVA model (i.e., model (13)) with patients' baseline responses as a univariate covariate. Sample sizes n_{ij}'s and some summary statistics for the data are given in Table 2. The model is unbalanced with the n_{ij} ranging from 4 to 7. Note that the SSE under the ANCOVA model is 71.24, while the SSE under the ANOVA model ignoring the covariate is 187.38.

The textbook test for interaction with zero interaction ($\gamma_{ij} = 0$ for all i and j) as the null hypothesis uses the statistic F_{AB} given in (8). Under both fixed effects and mixed effects models (with or without the covariate), F_{AB} has a central F-distribution when the null hypothesis of zero interaction is true. Based on the data, the p-value under the ANOVA model (ignoring the covariate) is less than 0.0001, whereas the p-value under the ANCOVA model is 0.1850. As discussed in Section 1, these results are useless for assessing treatment effects in the presence of possible treatment-by-center interaction. When the covariate is ignored, there is strong evidence that the treatment-by-center interaction is not zero, but we do not know whether the treatment-by-center interaction is large enough that assessing treatment main effects is inappropriate. When the covariate is included in the analysis, a p-value of 0.1850 indicates that the null hypothesis of zero interaction

cannot be rejected at the typical 5% significance level, but it does not provide any statistical assurance in concluding zero treatment-by-center interaction.

We now consider the tests for negligible interaction proposed in the previous section. First, we need to determine δ_0, a margin for treatment-by-center interaction. For illustrative purpose (not data analysis), we choose $\sqrt{\delta_0}$ to be 0.5 and 0.7, which are respectively about 20% and 30% of the clinical meaningful margin for treatment effects relative to the standard error σ. With these choices of δ_0, results for testing treatment-by-center interaction with hypotheses given by (2) are listed in Table 3 for ANOVA and ANCOVA models with fixed and mixed effects. For mixed effects models, both test rules (11) and (12) are used. Consider first the fixed effects models. From Table 3, we cannot reject the null hypothesis in (2) if the covariate is ignored. However, under the ANCOVA model, for both δ_0 values we can reject the null hypothesis in (2) at a significance level of 5%. Ignoring the covariate may substantially increase the error variability and, hence, there is not enough power in the interaction test under the ANOVA model. This is also true for the analysis under the mixed effects models. Under the mixed effects ANCOVA model, we cannot reject the null hypothesis in (2) when $\sqrt{\delta_0} = 0.5$, although we can reject the null hypothesis in (2) at 5% level of significance when $\sqrt{\delta_0} = 0.7$. It is reasonable to believe that the interaction test under mixed effects models is not as powerful as that under fixed effects models (see Figures 1-2).

Table 3. Results for Testing Negligible Interaction.

	$\sqrt{\delta_0} = 0.5$		$\sqrt{\delta_0} = 0.7$	
Model	Test Statistic	p-value	Test Statistic	p-value
Fixed Effect, ANOVA	6.0500	0.9823	6.0500	0.6943
Fixed Effect, ANCOVA	1.3900	0.0166	1.3900	0.0001
Mixed Effect, ANOVA, Test (11)	6.0500	0.9958	6.0500	0.9176
Mixed Effect, ANCOVA, Test (11)	1.3900	0.1754	1.3900	0.0368
Mixed Effect, ANOVA, Test (12)	2.6026	0.9947	1.6864	0.9159
Mixed Effect, ANCOVA, Test (12)	0.5531	0.1270	0.3590	0.0260

References

Boos, D. D. and Brownie, C. (1992). A rank-based mixed model approach to multisite clinical trials. *Biometrics* **48**, 61-72.

Ciminera, J. L., Heyse, J., Nguyen, H. and Tukey, J. W. (1993a). Tests for qualitative treatment-by-centre interaction using a "Pushback" procedure. *Statist. Medicine* **12**, 1033-1045.

Ciminera, J. L., Heyse, J., Nguyen, H. and Tukey, J. W. (1993b). Evaluation of multicentre clinical trial data using adaptations of the Mosteller-Tukey Procedure. *Statist. Medicine* **12**, 1047-1061.

Cheng, B. and Shao, J. (2006). Testing treatment effects in two-way linear models: Additive or full model? *Sankhyā* **68**, 392-408.

Christensen, R. (1996). Exact tests for variance components. *Biometrics* **52**, 309-314.

FDA (2001). *Guidance for Industry on Statistical Approaches to Establishing Bioequivalence*. Center for Drug Evaluation and Research, Food and Drug Administration, Rockville, Maryland.

Gail, M. and Simon, R. (1985). Testing for Qualitative Interactions Between Treatment Effects and Patient Subsets. *Biometrics* **41**, 361-372.

Gallo, J. and Khuri, A. I. (1990). Exact Tests for the Random and Fixed Effects in an Unbalanced Mixed Two-Way Cross-Classification Model. *Biometrics* **46**, 1087-1095.

ICH (1998). Statistical Principles for Clinical Trials. Tripartite International Conference on Harmonization Guideline, E9.

Khuri, A. I. and Littell, R. C. (1987). Exact Tests for the Main Effects Variance Components in an Unbalanced Random Two-Way Model. *Biometrics* **43**, 545-560.

Laster, L. L. and Johnson, M. F. (2003). Non-inferiority trials: the 'at least as good as' criterion. *Statist. Medicine* **22**, 187-200.

Öfversten, J. (1993). Exact Tests for Variance Components in Unbalanced Mixed Linear Models. *Biometrics* **49**, 45-57.

Peterson, B. and George, S. L. (1993). Sample size Requirements and Length of Study for Testing Interaction in a $1 \times k$ Factorial Design When Time-to-Failure is the Outcome. *Controlled Clinical Trials* **14**, 511-522.

Potthoff, R., Peterson, B. L. and George, S. L. (2001). Detecting Treatment-by-Centre Interaction in Multi-centre Clinical Trials. *Statist. Medicine* **20**, 193-213.

Searle, S. R. (1971). *Linear Models*. Wiley, New York.

Snapinn, S. M. (1998). Interpreting Interaction: the Classical Approach. *Drug Infor. J.* **32**, 433-438.

Speed, F. M. and Hocking, R. R. (1976). The Use of the $R(\ )$-Notation with Unbalanced Data. *Amer. Statist.* **30**, 30-33.

Thomsen, I. (1975). Testing Hypotheses in Unbalanced Variance Components Models for Two-Way Layouts. *Ann. Statist.* **3**, 257-265.

Department of Biostatistics, Columbia University, New York, NY 10032, U.S.A.

E-mail: bc2159@columbia.edu

Department of Statistics, University of Wisconsin, Madison, WI 53706, U.S.A.

E-mail: shao@stat.wisc.edu

(Received November 2004; accepted February 2006)