Journal of Modern Applied Statistical Methods

Volume 6 | Issue 1 Article 7

5-1-2007

Analyses of Unbalanced Groups-Versus-Individual Research Designs Using Three Alternative Approximate Degrees of Freedom Tests: Test Development and Type I Error Rates

Stephanie Wehry
University of North Florida, swehry@unf.edu

James Algina
University of Florida, algina@ufl.edu

Recommended Citation

Wehry, Stephanie and Algina, James (2007) "Analyses of Unbalanced Groups-Versus-Individual Research Designs Using Three Alternative Approximate Degrees of Freedom Tests: Test Development and Type I Error Rates," *Journal of Modern Applied Statistical Methods*: Vol. 6 : Iss. 1 , Article 7.
DOI: 10.22237/jmasm/1177992360

Journal of Modern Applied Statistical Methods
May, 2007, Vol. 6, No. 1, 53-65

1538 – 9472/07/$95.00

Analyses of Unbalanced Groups-Versus-Individual Research Designs Using Three Alternative Approximate Degrees of Freedom Tests: Test Development and Type I Error Rates

Stephanie Wehry
University of North Florida

James Algina
University of Florida

Three approximate degrees of freedom quasi-F tests of treatment effectiveness were developed for use in research designs when one treatment is individually delivered and the other is delivered to individuals nested in groups of unequal size. Imbalance in the data was studied from the prospective of subject attrition. The results indicated the test that best controls the Type I error rate depends on the number of groups in the group-administered treatment but does not depend on the subject attrition rates included in the study.

Key words: Groups versus-individuals, approximate degrees of freedom, unbalanced designs, Type I error rate

Introduction

In the simplest groups-versus-individuals research design, two treatments are compared, one of which is administered to J groups. The jth group $\left(j=1,\ldots,J\right)$ has n_j participants, for a total of $N_G=\sum_{j=1}^{J}n_j$ such participants. The other treatment is administered individually to N_I participants. For example, psychotherapy researchers investigating the efficacy of group therapy often use a wait-list control group (Burlingame, Kircher, & Taylor, 1994). The therapy is provided to participants in groups because the researcher believes group processes will enhance the effectiveness of the therapy. Group processes do not affect the participants in the wait-list control group because they do not receive a treatment. In comparative studies, the effectiveness of an active treatment delivered to groups is compared to the effectiveness of an active treatment delivered individually. For example, Bates, Thompson, and Flanagan (1999) compared the effectiveness of a mood induction procedure administered to groups to the effectiveness of the same procedure administered to individuals. Using a more complex groups-versus-individuals research design, Boling and Robinson (1999) investigated the effects of three types of study environment on a measure of knowledge following a distance-learning lecture. The three types of study environment included a printed study guide accessed by individuals, an interactive multi-media study guide accessed by individuals, and a printed study guide accessed by cooperative study groups.

Burlingame, Kircher, and Taylor (1994) reported that independent samples t tests, ANOVAs, and ANCOVAs were the most commonly used methods for analyzing data in group psychotherapy research. It is well known that the independent samples t test requires scores be independently distributed both between and within treatments—an assumption that is most likely violated in the groups-versus-individual research design. This lack of

Stephanie Wehry is Assistant Director for Research and Evaluation at the Florida Institute for Education. Her research interests are evaluating early childhood education programs, applied statistics, and psychometrics. Email her at swehry@unf.edu. James Algina is Professor of Educational Psychology. His interests are in psychometric theory and applied statistics. Email him at algina@ufl.edu.

independence is indicated by a non-zero intraclass correlation coefficient for participants who receive the group-administered treatment. Myers, Dicecco, and Lorch (1981), using simulated data, showed that the Type I error rates for independent samples *t* test is greater than nominal alpha when the intraclass correlation is positive. Burlingame, Kircher, and Honts (1994) reported similar results.

The Myers, Dicecco, & Lorch (1981) Quasi-F Test Statistic

Myers et al. (1981) developed a quasi-F statistic that takes into account the lack of independence of data collected from the participants in the same group in a groups-versus-individuals research design. The Myers et al. test statistic is based on the two models for the data. The model for the *i*th $\left(i=1,\ldots,N_I\right)$ participant within the individually administered treatment $\left(T_I\right)$ is

$$Y_{i/T_I}=\mu_I+\varepsilon_{i/T_I} \qquad (1)$$

and the model for the *i*th participant $\left(i=1,\ldots,n_j\right)$ within the *j*th group $\left(j=1,\ldots,J\right)$ within the group-administered treatment $\left(T_G\right)$ is

$$Y_{i/j/T_G}=\mu_G+\alpha_{j/T_G}+\varepsilon_{i/j/T_G}. \qquad (2)$$

Myers et al. assumed that $\varepsilon_{I/T_I}\sim N\left(0,\sigma^2_{S/T_I}\right)$, $\alpha_{j/T_G}\sim N\left(0,\tau^2\right)$, and $\varepsilon_{i/j/T_G}\sim N\left(0,\sigma^2_{S/G/T_G}\right)$. The assumption about the α_{j/T_G} implies that the groups in the group-administered treatment are considered to be representative of an infinitely large number of groups. Therefore, the Myers et al. method permits generalization of the result to this larger number of groups. In addition, Myers et al. assumed that the groups within the group-administered treatments were balanced $\left(n_1=,\cdots,=n_J\right)$.

Formulated as an approximate degrees of freedom (APDF) *t* statistic, the Myers et al. test statistic is

$$t_{APDF}=\frac{\bar{Y}_I-\bar{Y}_G}{\sqrt{a_1 MS_{S/T_I}+a_2 MS_{G/T_G}}}$$

where a_1 is $\left(1/N_I\right)$ and a_2 is $\left(1/N_G\right)$. The mean

$$\bar{Y}_I=\frac{1}{N_I}\sum_{i=1}^{N_I}Y_{i/T_I}$$

is the mean of the criterion scores for the participants in the individually administered treatment $\left(T_I\right)$,

$$MS_{S/T_I}=\frac{\sum_{i=1}^{N_I}\left(Y_{i/T_I}-\bar{Y}_I\right)^2}{N_I-1}$$

is the variance for participants who received the individually-administered treatment,

$$\bar{Y}_G=\frac{1}{N_G}\sum_{j=1}^{J}\sum_{i=1}^{n_j}Y_{i/j/T_G}$$

is the mean of the criterion scores of participants who received the group-administered treatment, and

$$MS_{G/T_G}=\frac{\sum_{j=1}^{J}n_j\left(\bar{Y}_{j/T_G}-\bar{Y}_G\right)^2}{J-1}$$

is the between-group mean square for these participants. It can be shown that the squared denominator of the *t* statistic estimates the sampling variance of the numerator given the assumptions made by Myers et al. about the random effect and residuals. The estimated Satterthwaite (1941) approximate degrees of freedom are

$$\hat{f}_2 = \frac{\left(a_1 MS_{S/T_I} + a_2 MS_{G/T_G}\right)^2}{\frac{\left(a_1 MS_{S/T_I}\right)^2}{N_I - 1} + \frac{\left(a_2 MS_{G/T_G}\right)^2}{J - 1}}.$$

An Alternative Approximation for the Degrees of Freedom

The Satterthwaite (1941) approximation of the distribution of the linear combination of mean squares in the denominator of the *t* statistic is based on the assumptions that MS_{S/T_I} and MS_{G/T_G} are independent random variables that are distributed as multiples of chi-square distributions. The distribution of the sum is approximated as chi-square with degrees of freedom estimated by equating the first two moments of the sample and the approximating chi-square distribution.

The discussion in Satterthwaite (1941) implied that this approximation of the distribution of the denominator improves as $J-1$ or N_I-1 increases and as

$$\frac{\left(N_I - 1\right)\left(n\tau^2 + \sigma^2_{S/G/T_G}\right)}{\left(J-1\right)\sigma^2_{S/T_I}} \tag{3}$$

becomes closer to 1.0. When there are two groups in the group-administered treatment level, *J* is as small as possible and the ratio of equation (3) is typically larger than 1.0 and increases as the number of participants in the two groups increases and as the intraclass correlation increases. Scarino and Davenport (1986) studied the Type I error rate of the Welch APDF *t* test and found it could be seriously inflated when (a) there is a negative relationship between the sampling variances of the means and the degrees of freedom for the estimated sampling variances and (b) the smaller of the two degrees of freedom is small. Wehry and Algina (2003) applied the work of Scarino and Davenport to the Myers et al. (1981) quasi-F test and showed that when *J* equals two or three and $\tau > 0$, the Satterthwaite approximation of the denominator degrees of freedom also resulted in a quasi-F test that does not control the Type I error rate at nominal alpha.

Scarino and Davenport (1986) developed a four-moment approximation of the degrees of freedom for use with the Welch *t* when the ratio of the sampling variances is large and the corresponding ratio of degrees of freedom is small. Wehry and Algina (2003) adapted the four-moment approximation for use with the groups-versus-individual research design. The four-moment approximation to the degrees of freedom is

$$\hat{f}_4 = \frac{\left\{\frac{u^2}{m_1} + \frac{1}{m_2}\right\}^3}{\left(\frac{u^3}{m_1^2} + \frac{1}{m_2^2}\right)^2} \tag{4}$$

where $u = a_2 MS_{G/T_G} \big/ a_1 MS_{S/T_I}$, $m_1 = J-1$, and $m_2 = N_I - 1$. Like the Satterthwaite approximation employed by Myers et al. (1981), the four-moment degrees of freedom is based on the assumption of a balanced design.

Scarino and Davenport (1986) reported that the four-moment APDF test is conservative under some conditions and suggested using an average of the two-moment and four-moment approximations of the degrees of freedom. Wehry and Algina (2003) conducted a study of the APDF quasi-F test with the two-moment, four-moment, and an arithmetic average of the two- and four-moment approximations of the degrees of freedom using both analytical results and simulated data. They concluded that when the group-administered treatment is delivered to two groups, the four-moment APDF quasi-F test should be used and when the group-administered treatment is delivered to three or more groups, the average-moment APDF quasi-F test should be used. However, the two-moment APDF quasi-F test is only slightly liberal in conditions involving more than three groups.

Quasi-F Statistics For Use When Data Are Not Balanced Across Groups In The Group-Administered Treatment Level

The purpose of the present study is to extend the work of Myers et al. (1981) and Wehry and Algina (2003) to include groups-versus-individuals research designs that are not balanced across either treatment levels (i.e., $N_I \neq N_G$) or the groups in the group-administered treatment level (i.e., $n_j \neq n_{j'}$ for at least one pair of j and j'). Usually in experimental research an equal number of participants are randomly assigned to each treatment level; however, N_I and N_G, as well as the n_j can be affected by attrition of participants. Burlingame, Kircher, and Taylor (1994) found 18% subject attrition was the median reported attrition rate of subjects in a survey of psychotherapy literature. Clarke (1998) suggested that the attrition rate in wait-list control groups could even be higher than that of the active treatment level.

Imbalance can also result from studying naturally occurring groups such as family units and classrooms. Methods that accommodate imbalance across groups in the group-administered treatment level have not been developed. A possible solution to the imbalance across groups in the group-administered treatment level is to randomly eliminate participants until balance is achieved. However, eliminating data results in a loss of statistical power.

APDF Quasi-F Test for Unbalanced Data

As is well known, if the variances of $\overline{Y}_I$ and $\overline{Y}_G$ were known, the hypothesis $H_O\!:\mu_I - \mu_G = 0$ could be tested by

$$\chi^2 = \frac{\left(\overline{Y}_I - \overline{Y}_G\right)^2}{Var\left(\overline{Y}_I - \overline{Y}_G\right)}. \qquad (5)$$

Because observations are independent across treatment levels, substituting the variances of $\overline{Y}_I$ and $\overline{Y}_G$ into equation (5) results in

$$\chi^2 = \frac{\left(\overline{Y}_I - \overline{Y}_G\right)^2}{\dfrac{\sigma^2_{S:T_I}}{N_I} + \dfrac{\tau^2 \sum_{j=1}^{J} n_j^2}{N_G^2} + \dfrac{\sigma^2_{S/G/T_G}}{N_G}}. \qquad (6)$$

However, the variances are not known, and, in order to develop a test statistic that can be used in practice, two steps must be completed: Develop estimators of the variance components in equation (6) and approximate the distribution of the resulting test statistic. Approximating the distribution of the denominator by a chi-square distribution and the distribution of the test statistic by an F distribution is a common practice in statistics.

Variance Component Estimates

There are numerous methods for estimating the variance components. Perhaps the most commonly used method is the method of moments, also called the ANOVA estimation of variance components (Milliken & Johnson, 1992). Meyers et al. (1981) used the method of moments variance component estimators in formulating the quasi-F test statistic. The method of moments procedure is based on equating the expected values of the sums of squares to their respective observed values.

Other estimation methods include maximum likelihood (ML), restricted maximum likelihood (REML), minimum norm quadratic unbiased (MINQUE), and minimum variance quadratic unbiased (MIVQUE) estimators. ML estimators are values of the parameter space that maximize the likelihood function. In REML, the likelihood equations are partitioned into two parts, one part that is free of fixed effects. REML maximizes the part that has no fixed effects. MINQUE and MIVQUE are iterative and the researcher must provide initial values of the components. All methods produce the same results when the design is balanced (Milliken & Johnson, 1992; Swallow & Monahan, 1984).

Swallow and Monahan (1984) conducted a Monte Carlo study of ANOVA, ML, REML, MIVQUE and MINQUE methods of estimating the variance components of a one-way unbalanced, random effects design. All simulated data were normal, and the variables manipulated were the degree of imbalance, the number of groups, and the ratio of $\tau^2 / \sigma^2_{S/G/T_G}$. In terms of bias of the estimates, the results indicated, except in cases of extreme patterns of imbalance, $n_j = (1,1,1,1,13,\text{and } 13)$ and $n_j = (1,1,1,1,1,1,1,19,\text{and } 19)$, ANOVA, REML, and MINQUE estimators showed little difference. However, the results indicated that

ML methods were the best estimators of τ^2 when $\tau^2 \big/ \sigma^2_{S/G/T_G} \leq .5$ because of the small bias and the low mean square error of the estimate. When $\tau^2 \big/ \sigma^2_{S/G/T_G}$ is large, Swallow and Monahan indicated there may be a substantial downward bias and that ML methods have no superiority over the other methods. There was little difference among the methods studied when estimating $\sigma^2_{S/G/T_G}$. Milliken and Johnson (1992) suggested that ANOVA estimates should have good properties for nearly balanced data, and Swallow and Monahan concluded that unless the data are severely unbalanced and $\tau \big/ \sigma^2_G > 1$, ANOVA estimates are adequate.

The results of the Swallow and Monahan (1984) study and the recommendations of Milliken and Johnson (1992) suggested that ANOVA estimates of the variance components are likely to be adequate for the groups-versus-individuals research design. Data as extreme as that simulated in the Swallow and Monahan study seems likely to be rare in group research; therefore, method of moments estimators of the variance components are used for the quasi-F test for comparing the effectiveness of two treatment levels when data are unbalanced.

The expected values for the mean squares for groups (henceforth when the term groups is used, it will refer to the groups within the group-administered treatments) are

$$EMS_{G/T_G} = \sigma^2_{S/G/T_G} + n_o \tau^2, \qquad (7)$$

where

$$n_o = \frac{1}{J-1}\left(N_G - \frac{\sum_{j=1}^{J} n_j^2}{N_G} \right)$$

(Snedecor & Cochran, 1956). The other two expected values are

$$EMS_{S/G/T_G} = \sigma^2_{S/G/T_G} \qquad (8)$$

and

$$EMS_{S/T_I} = \sigma^2_{S/T_I}.$$

The mean squares are equated with their respective expected values of equations (7), (8), and (9) are the resulting equations are solved for the ANOVA variance component estimates. The variance component estimates are then substituted into equation (6) to obtain the quasi-F test statistic for comparing weighted treatment level means.

The Quasi-F Test Statistic

Using the estimated variance components the quasi-F test statistic is

$$\hat{F}_{quasi} = \frac{\left(\bar{Y}_I - \bar{Y}_G\right)^2}{\left\{ \begin{array}{l} \dfrac{MS_{S/T_I}}{N_I} \\ + \dfrac{\left(MS_{G/T_G} - MS_{S/G/T_G}\right)\sum_{j=1}^{J} n_j^2 / n_o}{N_G^2} \\ + \dfrac{MS_{S/G/T_G}}{N_G} \end{array} \right\}}, \qquad (9)$$

which simplifies to

$$\hat{F}_{quasi} = \frac{\left(\bar{Y}_I - \bar{Y}_G\right)^2}{\left\{ \begin{array}{l} \dfrac{MS_{S/T_I}}{N_I} \\ + \dfrac{MS_{G/T_G} \sum_{j=1}^{J} n_j^2}{n_o N_G^2} \\ + \dfrac{MS_{S/G/T_G}\left(n_o N_G - \sum_{j=1}^{J} n_j^2\right)}{n_o N_G^2} \end{array} \right\}}.$$

The denominator of the quasi-F statistic is a synthetic mean square in the form of

$$MS = a_1 MS_{S/T_I} + a_2 MS_{G/T_G} + a_3 MS_{S/G/T_G}, \qquad (10)$$

where

$$a_1 = \frac{1}{N_I},$$

$$a_2 = \frac{\sum_{j=1}^{J} n_j^2}{n_o N_G^2},$$

and

$$a_3 = \left(\frac{n_o N_G - \sum_{j=1}^{J} n_j^2}{n_o N_G^2} \right).$$

Approximating Chi-Square Distribution

The model for the group-administered treatment is a random effects ANOVA model [see equation (2)]. For a design that is balanced across classes, Searle (1992) showed the mean squares between and within classes are independent and are distributed as multiples of chi-square distributions. When the data are not balanced across classes, the mean squares within and between are still independent; however, the mean square between classes is not distributed as a multiple of a chi-square distribution. Nevertheless, Burdick, and Graybill (1988) indicated as long as τ is not too large, approximating the mean square between as a multiple of a chi-square distribution does not result in a large error.

Two-Moment Approximation of the Degrees of Freedom

The Satterthwaite (1941) approximation for the degrees of freedom for the linear combination in equation (10) is

$$\hat{f}_2 = \frac{\left(a_1 MS_{S/T_I} + a_2 MS_{G/T_G} + a_3 MS_{S/G/T_G} \right)^2}{\frac{\left(a_1 MS_{S/T_I} \right)^2}{N_I - 1} + \frac{\left(a_2 MS_{G/T_G} \right)^2}{J - 1} + \frac{\left(a_3 MS_{S/G/T_G} \right)^2}{N_G - J}}$$

.

It should be noted that $a_3 \le 0$, with equality holding only when $n_o = n$. Therefore, when data are not balanced across groups in the group-administered treatment level, it is possible for the denominator of the quasi-F statistic to be less than or equal to zero when the estimate of τ^2 is substantially smaller than zero. In these cases, as suggested by Searle (1992), it is reasonable to assume τ^2 is zero and replace the quasi-F statistic by the Welch *t*-test where

$$t_{W_{APDF}} = \frac{\left(\bar{Y}_I - \bar{Y}_G \right)}{\sqrt{\frac{MS_{S/T_I}}{N_I} + \frac{MS_{S/T_G}}{N_G}}}$$

and

$$MS_{S/T_G} = \frac{\sum_{j=1}^{J} \sum_{i=1}^{n_j} \left(Y_{i/j/T_G} - \bar{Y}_G \right)^2}{\left(N_G - 1 \right)}$$

with two-moment degrees of freedom

$$\hat{df} = \frac{\left(\frac{MS_{S/T_I}}{N_I} + \frac{MS_{S/T_G}}{N_G} \right)^2}{\frac{\left(MS_{S/T_I} \right)^2}{N_I^{\,2} \left(N_I - 1 \right)} + \frac{\left(MS_{S/T_G} \right)^2}{N_G^{\,2} \left(N_G - 1 \right)}}$$

(Welch, 1938).

Modified Four-Moment Approximation of the Degrees of Freedom

Because the coefficients of the variance component terms in the synthetic error term for unbalanced data are not all positive and because of the occurrence of conditions in which the ratio of the degrees of freedom is less than one when the ratio of the corresponding sampling variances is greater than one, the two-moment quasi-F test may not control the Type I error rate at the nominal level. The four–moment approximation was developed by Scariano and Davenport (1986) for a synthetic mean square that is the sum of two positive terms. Rather than expanding the four-moment approach to three terms including one that is negative, a simpler approach that combines the two-moment and four-moment approximations was used in this study.

In order to compute the modified four-moment approximation, the degrees of freedom for $a_2MS_{G/T_G} + a_3MS_{S/G/T_G}$ are first approximated using the two-moment approach. As noted previously, Searle (1992) showed MS_{G/T_G} and $MS_{S/G/T_G}$ are independent when data are unbalanced, Burdick and Graybill (1988) indicated as long as τ^2 is not too large MS_{G/T_G} can be approximated as a multiple of chi-square distribution, and Swallow and Monahan (1984) showed that method of moments estimation works well in one-way, random effects, unbalanced ANOVA designs as long as $\tau^2 / \sigma^2_{S/G/T_G} \le 1$. The two-moment degrees of freedom for

$$MS_{error_{T_G}} = a_2MS_{G/T_G} + a_3MS_{S/G/T_G}$$

are

$$\hat{f}_{2_G} = \frac{\left(MS_{error_{T_G}}\right)^2}{\left[\frac{(a_2MS_{G/T_G})^2}{(J-1)} + \frac{\left(a_3MS_{S/G/T_G}\right)^2}{\left(N_G - J\right)}\right]}.$$

This value of $\hat{f}_{2_G}$ along with $MS_{error_{T_G}}$ and the estimate of the individual treatment level variance, MS_{S/T_I}, are used in the four-moment approximation of equation (4). In the modified four-moment approximation, $u = MS_{error_{T_G}} / a_1MS_{S/T_I}$, $m_1 = \hat{f}_{2_G}$, and $m_2 = \left(N_I - 1\right)$. When $MS_{error_{T_G}} \le 0$, the quasi-F statistic is replaced by the Welch *t*-test.

Modified Averaged Degrees of Freedom Approximation of the Degrees of Freedom

Scariano and Davenport (1986) reported that, with completely balanced data, the four-moment quasi-F test is conservative under some conditions. Therefore, an arithmetic average of the two-moment and the modified four-moment approximations was also included in the present study. When $MS_{error_G} \le 0$, data were analyzed using the Welch *t* test; otherwise, the two-moment approximation and the modified four-moment approximation to the degrees of freedom were arithmetically averaged resulting in an averaged degrees of freedom quasi-F test.

Example 1

Participants were randomly assigned to two conditions and completed three trials of the prisoner's dilemma. The data are the number of competitive choices across the three trials. In one condition, participants completed the three trials independently. In the second condition, participants worked in teams and discussed how to respond to each trial. However, participants within a team responded individually. For participants in the individual treatment the relevant results are $N_I = 32, \bar{Y}_I = .469$, $MS_{S/T_I} = .773$. For participants in the group-administered treatment, the results are $N_G = 48$, $J = 15$, $\sum_{j=1}^{J} n_j^2 = 141$, $\bar{Y}_G = .905$, $MS_{G/T_G} = 1.896$, and $MS_{S/G/T_G} = .833$. The calculated *t* statistic is -1.86. The degrees of freedom are $\hat{f}_2 = 56.37$, $\hat{f}_4 = 56.04$, and $\hat{f}_a = 54.70$. For all three degrees of freedom, $\left(Prob > |t|\right) = .068$. Because the theory predicts more competitive response following group discussion, the results are in support of the theory.

Example 2

In an evaluation of a pre-school literacy program, the evaluators were interested in whether reading achievement was different in single-classroom sites and multiple-classroom sites. The available data are mean end-of year reading achievement for each of the classrooms. For single-classroom sites the relevant results are $N_I = 38, \bar{Y}_I = 88.85$, $MS_{S/T_I} = 57.84$. For participants in the multiple-classroom sites, the results are $N_G = 63$, $J = 29$, $\sum_{j=1}^{J} n_j^2 = 216$, $\bar{Y}_G = 87.52$, $MS_{G/T_G} = 69.09$, and $MS_{S/G/T_G} = 22.22$. The calculated *t* statistic is

0.71. The degrees of freedom are $\hat{f}_2 = 30.87$, $\hat{f}_4 = 17.76$, and $\hat{f}_a = 24.31$. For all three degrees of freedom, $\left(Prob > |t|\right) = .76$. The results do not support the belief that mean reading achievement is different in single-classroom and multiple classroom sites.

Methodology

Variables Manipulated in the Monte Carlo Study

The design of the Monte Carlo study had five between-subjects factors and one within-subjects factor. There were a total of 2700 conditions. The design included the three approaches to the approximation of the error term degrees of freedom as levels of the within-subjects factor. The number of groups, planned size of the groups, level of the intraclass correlation, ratio of the group to individual treatment level variances, and the rate of subject attrition were the five between-subjects factors. There were five levels of the number of groups, $J =$ 2, 3, 4, 5, and 6; five levels of planned group size, $n =$ 4, 8, 12, 16, and 20 subjects nested in the groups; three levels of intraclass correlation, $\tau^2 \big/ \left(\tau^2 + \sigma^2_{S/G/T_G}\right) =$ .0, .2, and .4; three levels of the ratio of group to individual treatment level variances, $\left(\tau^2 + \sigma^2_{S/G/T_G}\right) / \sigma^2_{S/I} =$ 0.75, 1.00, and 1.25; and four combinations of individual and group treatment level attrition rates, .15 and .15, .15 and .25, .25 and .15, and .25 and .25.

Data Generation

The simulation in the study was carried out using the random number generation functions of SAS, Release 6.12. Scores for simulated participants in the individually administered treatment level were generated using the equation

$$Y_{i/I} = \mu_I + \varepsilon_{i/T_I}$$

where μ_I was arbitrarily set at 100 and the $\varepsilon_{i:T_I}$s were pseudorandom standard normal deviates generated using RANNOR. The variable $Y_{i:T_I}$ was set to the missing data indicator if $U_{i/T_I} < p_{T_I}$ where p_{T_I} is the individually administered treatment level attrition rate and U_{i/T_I} was a pseudorandom uniform deviate generated using RANUNI. However, N_I was not permitted to be smaller than two.

Scores for simulated participants in the group-administered treatment level were generated using the equation

$$Y_{i/j/T_G} = \mu_G + \alpha_{j/T_G} + \varepsilon_{i/j/T_G}$$

where μ_G was arbitrarily set at 100, α_{j/T_G} was a pseudorandom normal deviate with mean zero and variance τ^2, and $\varepsilon_{i/j/T_G}$ was a pseudorandom normal deviate with mean zero and variance $\sigma^2_{S/G/T_G}$. The variable $Y_{i/j/T_G}$ was set to a missing value indicator if $U_{i/j/T_G} < p_{T_G}$, where p_{T_G} is the group-administered treatment level attrition rate and $U_{i/j/T_G}$ was a pseudorandom uniform deviate generated using RANUNI. However, in all cases n_j was not permitted to be smaller than two.

Each of the conditions was replicated 10,000 times, and the Type I errors of the three tests were counted over the replications of each condition. All tests were conduted at $\alpha = .05$.

Results

A Number of Groups (5) × Planned Group Size (5) × Intraclass Correlation (3) × Ratio of Variance (3) × Attrition Rate (4) × Degrees of Freedom Approximation (3), with repeated measures on the last factor, ANOVA was used to analyze the Type I error rate data. Because there was only one data point for each combination of the six factors, the five-way interaction of the first five factors was used as the error term for between-replications effects and the six-way interaction was used as the error term for all within-replications effects. For each effect omega squared was used to express the size of the effect as a proportion of the total variance. An effect was considered important if it was significant at $\alpha = .05$ and accounted for more than 1% of the total variance in the Type I error rate. Table 1 presents the omega squares for all significant effects. The sum of the omega squares for all of the important effects was 0.929. All factors except subject attrition rate were involved in an effect that met our criterion for an important influence on the Type I error rate.

Averaged over all factors, other than number of groups in the group-administered treatment, the average Type I error rate of the two-moment test was greater than that for the averaged degrees of freedom test. Also the average Type I error rate of the averaged degrees of freedom test was greater than that for the modified four-moment test. When there were two groups only the modified four-moment test controlled the Type I error rate near nominal alpha; however, the modified four-moment test resulted in a conservative quasi-F test with three or more groups. In all conditions involving two groups, increasing the planned size of the groups, the ratio of treatment level variances, or the intraclass correlation increased the Type I error rate. Under conditions involving three or more groups, increasing the intraclass correlation increased the Type I error rate of all three tests and increasing the ratio of treatment level variances and the planned size of the groups increased the Type I error rate of the two-moment and averaged degrees of freedom tests. As the number of groups increased the effect of increasing the ICC or the planned size of the groups declined. However, under conditions of three groups or more groups, increasing the ratio of the treatment level variances and the planned size of the groups decreased the Type I error rate of the modified four-moment quasi-F test.

Table 2 contains the minimum and maximum Type I error rate averaged over subject attrition by number of groups, approximate degrees of freedom approach, and intraclass correlation. Minima and maxima were computed over planned size of groups and ratio of treatment level variances. In Table 2 bold and italicized figures indicate the degrees of freedom approach that resulted in better control of Type I error rate for a particular number of groups and ICC. When both bold figures and italicized figures are presented, the italicized figures indicate the degrees of freedom approximation that tended to result in a higher Type I error rate. Tests are considered unacceptable if the maximum Type I error rate is above .075, the upper limit of Bradley's (1978) liberal criterion for a robust test or if the minimum Type I error rate is below .025 the lower limit of Bradley's (1978) liberal criterion.

Inspection of Table 2 indicates that when there are two groups, the modified four-moment test should be used at the risk of a conservative test when the ICC is near zero. The averaged degrees of freedom test may be more attractive with a low ICC, but the fact that it has a strong liberal tendency when the ICC is 0.20 raises the question of how the two tests function for ICCs between 0.00 and 0.20. Supplementary results are shown in Table 3 for ICCs of 0.05, 0.10, and 0.15. In the simulations conducted to obtain these results, all other conditions were the same as in the original study. The findings that the averaged degree of freedom test has a liberal tendency for an ICC of 0.10 and that the conservative tendency of the modified four-moment test is less marked with an ICC of 0.05 than with an ICC of 0.00 suggest the modified four-moment test should be used when there are two groups in the group-administered treatment.

When there are three groups, the results in Tables 2 and 3 suggest the averaged degree of freedom test should be used at the risk of a slightly conservative test when the ICC is near zero. Then the two-moment test may be more attractive. However, it is not clear how valid an estimated ICC will be in selecting between the two tests. Given the very mild conservative tendency for the averaged degrees of freedom test, it is recommended when there are three groups.

When there are four or more groups either the two-moment test or the averaged degrees of freedom test might be used. The former can be somewhat liberal, with the tendency increasing as the ICC increased, but decreasing as the number of groups increased. The averaged degrees of freedom test can be somewhat conservative, with the tendency decreasing as the ICC increased and as the number of groups decreased.

Table 1. Mean Square Components and $\hat{\omega}_j^2$ for the Important Effects $\left(\hat{\omega}_j^2 > .01\right)$

Source of MS	$\hat{\omega}_j^2$
Between Replications Effects	
Number of Groups – *g*	0.239
Planned Size of Groups - *n*	0.024
Intraclass Correlation - *icc*	0.073
$g \times n$	0.056
$g \times icc$	0.079
Within-Replication Effects	
Approximation – *t*	0.390
$t \times g$	0.020
$t \times n$	0.031
t×*ratio of treatment level variance*	0.017

Table 2. Minimum and Maximum Average Type I Error Rate by Number of Groups, Test, and ICC

Number of Groups	Test	ICC 0.00	0.20	0.40
2	$\hat{f}_2$	.0470, .0759	.0532, .1118	.0571, .1204
	$\hat{f}_{ave}$	**.0390, .0572**	.0459, .0907	.0496, .1005
	$\hat{f}_4$	.0338, .0401	**.0412, .0580**	**.0437, .0663**
3	$\hat{f}_2$	*.0471, .0589*	.0514, .0770	.0537, .0776
	$\hat{f}_{ave}$	**.0411, .0488**	**.0450, .0634**	**.0476, .0637**
	$\hat{f}_4$	.0299, .0362	.0322, .0404	.0319, .0403
4	$\hat{f}_2$	*.0488, .0560*	*.0506, .0631*	*.0520, .0637*
	$\hat{f}_{ave}$	**.0422, .0481**	**.0459, .0528**	**.0458, .0539**
	$\hat{f}_4$	.0281, .0400	.0283, .0393	.0298, .0390
5	$\hat{f}_2$	*.0473, .0533*	*.0513, .0585*	*.0491, .0603*
	$\hat{f}_{ave}$	**.0436, .0467**	**.0469, .0499**	**.0451, .0509**
	$\hat{f}_4$	.0282, .0417	.0303, .0411	.0326, .0410
6	$\hat{f}_2$	*.0480, .0557*	*.0488, .0568*	*.0507, .0557*
	$\hat{f}_{ave}$	**.0442, .0491**	**.0451, .0500**	**.0464, .0505**
	$\hat{f}_4$	.0299, .0436	.0317, .0423	.0326, .0405

Table 3. Minimum and Maximum Average Type I Error Rate by Number of Groups, Test, and ICC: Supplemental Conditions

Number of Groups	Test	ICC 0.05	0.10	0.15
2	$\hat{f}_2$	.0482, .0908	.0489, .0990	.0508, .1066
	$\hat{f}_{ave}$	*.0396, .0705*	.0404, .0784	.0430, .0870
	$\hat{f}_4$	**.0352, .0481**	**.0360, .0513**	**.0383, .0562**
3	$\hat{f}_2$	*.0492, .0660*	*.0472, .0711*	*.0495, .0733*
	$\hat{f}_{ave}$	**.0418, .0538**	**.0416, .0585**	**.0436, .0604**
	$\hat{f}_4$	.0296, .0377	.0310, .0373	.0307, .0389

Conclusion

Myers et al. (1981) presented a two-moment, quasi-F test for use when one treatment is delivered to individuals and one is delivered to groups of participants and the data are balanced for the groups in the group-administered treatment. Wehry and Algina (2003) extended that quasi-F test to include a four-moment and an averaged degrees of freedom quasi-F test for use when data are balanced across the group-administered treatment level.

In this study, the two-moment approach developed by Myers et al. (1981) and the four-moment and averaged degrees of freedom approaches developed by Wehry and Algina (2003) were extended to include groups versus individual research designs in which data are not necessarily balanced across treatment levels or across groups in the group-administered treatment level. In addition, Type I error rates of the resulting tests were estimated. The results indicated the modified four-moment test should be used when the group-administered treatment is delivered to two groups and the averaged degrees of freedom approach should be used when the group-administered treatment is delivered to three groups. When there are four or more groups, either test could be used—the averaged degrees of freedom test is has a slightly conservative tendency and the two-moment test has a slightly liberal tendency. When there are four or five groups the Type I error rate for the averaged degrees of freedom test is between .040 and .055. The Type I error for two-moment test can be larger than .06. When there are six groups, the averaged degrees of freedom test controls the Type I error rate between .044 and .051; the two-moment test controls it between .048 and .057.

Although, it is recommended to use the four-moment test when there are two groups, researchers should be very cautious about using a group-versus-individuals design with only a few groups. For a balanced design, Wehry and Algina (2003) showed that power is likely to be very low when there are just two groups and there is no reason for the design to be more powerful when the design is unbalanced. More generally, Myers et al. (1981) have shown that the number of groups can have a larger effect on power than the number of participants per groups and therefore recommended designs with as large a number of groups as possible.

At least four lines of additional research are attractive. Comparison of the three approximate degrees of freedom tests to mixed model tests using Satterthwaite or Kenward-Rogers degrees of freedom might be investigated. One difference between the current approaches and the mixed-model approach is the estimate of the mean for the group-administered treatment. In the present approach the estimated mean is computed by weighting the group means by the group sample sizes. In the mixed model approach, the mean for the group-administered treatment would be estimated by generalized least squares and would have a sampling variance that is not larger than the sampling variance of the mean used in the present approach. This may make the mixed model approach more powerful. However, Wehry and Algina (2003) found that with balanced designs, the mixed model approach had poor control of the Type I error rate in some situations and this problem may generalize to unbalanced designs.

The performance of the three tests when data are not normal is important. Micceri (1987) reported that a wide variety of psychometric distributions may not be normal and that random-effects ANOVA tests may not be robust to departures from normality, especially when conditions involve unbalanced designs or small sample sizes. Developing robust versions of the tests is important. Finally extension of the tests to more than two groups and to multivariate designs would be useful.

References

Bates, G. W., Thompson, J. C., & Flanagan, C. (1999). The effectiveness of individual versus group induction of depressed mood. *The Journal of Psychology*, *33*, 245-252.

Boling, N. C., & Robinson, D. H. (1999). Individual study, interactive multimedia, or cooperative learning: Which activity best supplements lecture-based distance education? *Journal of Educational Psychology*, *91*, 169-174.

Bradley, J. V, (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*, 144-152.

Burdick, R. K., & Graybill, F. A. (1988). The present status of confidence interval estimations on variance components in balanced and unbalanced random models. *Communications in statistics: theory and methods*, *17*, 1165-1195.

Burlingame, G. M., Kircher, J. C., & Honts, C. R. (1994). Analysis of variance versus bootstrap procedures for analyzing dependent observations in small group research. *Small Group Research*, *25*, 486-501.

Burlingame, G. M., Kircher, J. C., & Taylor, S. (1994). Methodological considerations in group psychotherapy research: Past, present, and future practices. In A. Fuhriman & G. Burlingame (Eds.), *Handbook of group psychotherapy and counseling: An empirical and clinical synthesis*. (pp. 41-80). New York: Wiley.

Clarke, G. N. (1998). Improving the transition from basic efficacy research to effectiveness studies: Methodological issues and procedures. In A. E. Kazdin (Ed.), *Methodological issues and strategies in clinical research*, (2nd ed.) (pp. 541-559). New York: Wiley.

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

Myers, J., Dicecco, J., & Lorch, Jr., J. (1981). Group dynamics and individual performances: Pseudogroup and Quasi-F analyses. *Journal of Personality and Social Psychology*, *40*, 86-98.

Milliken, G. A., & Johnson, D. E. (1992). *Analysis of messy data volume 1; Designed experiments*. Boca Raton, FL: Chapman & Hall.

Satterthwaite, F. W. (1941). Synthesis of variance. *Psychometrika*, *6*, 309-316.

Scariano, S. M., & Davenport, J. M. (1986). A four-moment approach and other practical solutions to the Behrens-Fisher problem. *Communications in statistics: theory and methods*, *15*, 1467-1504.

Searle, S. R. (1992). *Variance components*. New York: Wiley.

Snedecor, G. W., & Cochran, W. G. (1956). *Statistical methods applied to experiments in agriculture and biology* (5th Ed.). Ames, IA: Iowa State Coll. Press.

Swallow, W. H., & Monahan, J. F. (1984). Monte Carlo comparisons of ANOVA, MIVQUE, REML, and ML estimators of variance components. Technometrics, 26, 47-57.

Wehry, S. & Algina, J. (2003). Type I Error Rates of Four Methods for Analyzing Data Collected in a Groups-Versus-Individuals Design. *Journal of Modern Applied Statistical Methods*, *2*, 400-413

Welch, B. L. (1938). On the comparison of several mean values: An alternative approach. *Biometrika*, *38*, 330-336.