

Type I Error Rates of Four Methods for Analyzing Data Collected in a Groups-Versus-  
Individuals Design

by

Stephanie Wehry

University of North Florida

and

James Algina

University of Florida

## ABSTRACT

Using previous work on the Behrens-Fisher problem, two approximate degrees of freedom tests, that can be used when one treatment is individually administered and one is administered to groups, were developed. Type I error rates are presented for these tests, an additional approximate degrees of freedom test developed by Myers, Dicecco, and Lorch (1981), and a mixed model test. The results indicate that the test that best controls the Type I error rate depends on the number of groups in the group-administered treatment. The mixed model test should be avoided.

Keywords: groups-versus-individuals design, approximate degrees of freedom tests, mixed models

## Type I Error Rates of Four Methods for Analyzing Data Collected in a Groups-Versus-Individuals Design

### Introduction

When a groups-versus-individuals design is used to compare two treatments, one treatment is administered to  $J$  groups of  $n$  participants (for a total of  $N_G$  such participants) and one treatment is individually administered to  $N_I$  participants or the individual participants may be in a no-treatment control group. For example, psychotherapy researchers investigating the efficacy of group therapy often use a wait-list control group (Burlingame, Kircher, and Taylor, 1994). The therapy is provided to participants in groups because the researchers believe 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, much less meet in groups. According to Clarke (1998) the most common design in psychotherapy research involves the use of a randomly assigned control condition, which can feature a variety of no-treatment control schemes.

The groups-versus-individuals design is also used when the purpose is to compare the effectiveness of an active treatment delivered to groups to 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. Boling and Robinson (1999) investigated the effects of study environment on a measure of knowledge following a distance-learning lecture. The three levels of study environment included a printed study guide accessed by individuals, an interactive multimedia study guide accessed by individuals, and a printed study guide accessed by cooperative study groups.

A possible model for the data collected in a groups-versus-individuals design consists of two submodels. For participants in the individually administered treatment the submodel is

$$Y_{i:T_I} = \mu_I + \varepsilon_{i:T_I} \quad (1)$$

where  $i : T_I$  ( $i = 1, \dots, N_I$ ) denotes the  $i$ th participant within the individually-administered treatment. For participants in the group-administered treatment

$$Y_{i:j:T_G} = \mu_G + \alpha_{j:T_G} + \varepsilon_{i:j:T_G} \quad (2)$$

where  $i : j : T_G$  ( $i = 1, \dots, n$ ) denotes the  $i$ th participant within the  $j$ th group ( $j = 1, \dots, J$ ) in the group-administered treatment. An important question is whether to treat the  $\alpha_{j:T_G}$  as fixed or random. When the researcher views the groups in the group-administered treatment as representative of a larger number of groups,  $\alpha_{j:T_G}$  should be treated as random. In the remainder of the paper we assume that the groups in the group-administered treatment comprise a random factor with the groups in the study representing an infinitely large number of groups.

Burlingame, Kircher, and Taylor (1994) reported that the independent samples  $t$  test, ANOVA, and ANCOVA were the most commonly used methods for analyzing data in group psychotherapy research. It is well known that these procedures require the scores for individual to be independently distributed both between and within treatments, an assumption that is likely to be violated for the participants in the group-administered treatment when  $\alpha_{j:T_G}$  is random. It is also well known that these procedures are not robust to violations of the independence assumption (see, for example, Scheffe, 1958). When the groups-versus-individuals design is used, lack of independence is indicated by a non-zero intraclass correlation coefficient for the participants who receive the group-administered treatments. Myers, Dicecco, and Lorch (1981), using simulated data, showed that the Type I error rates for the independent samples  $t$  test is above the nominal alpha level when the intraclass correlation is positive. Burlingame, Kircher, and Honts (1994) reported similar results. In passing we note that if the researcher believes it is appropriate to treat the  $\alpha_{j:T_G}$  as fixed, if both error terms are normally distributed, and if the error

terms have equal variances, the treatments can be compared by using an independent samples ANOVA and testing the hypothesis

$$H_0 : \mu_I = \mu_G \quad (3)$$

but generalization of the results to additional groups is not warranted.

Myers et al. (1981) developed two statistical tests of the hypothesis given in equation (3). These tests take the lack of independence into account and allow generalization of the results to the population of groups represented by the groups in the group-administered treatment. (In the following, groups will always refer to the groups in the group-administered treatment.) One of these procedures used a quasi-F statistic and degrees of freedom approximated by the Satterthwaite (1941) method. Formulated as an approximate degrees of freedom (APDF)  $t$  statistic, the Myers et al. test statistic is

$$t_{APDF} = \frac{\bar{Y}_I - \bar{Y}_G}{\sqrt{\frac{MS_{S/T_I}}{N_I} + \frac{MS_{G/T_G}}{N_G}}} \quad (4)$$

where  $\bar{Y}_I = \sum_{i=1}^{N_I} Y_{i:T_I} / N_I$  is the mean of the criterion scores and

$$MS_{S/T_I} = \frac{\sum_{i=1}^{N_I} (Y_{i:T_I} - \bar{Y}_I)^2}{N_I - 1} \quad (5)$$

is the variance for participants who received the individually administered treatment;

$\bar{Y}_G = \sum_{j=1}^J \sum_{i=1}^n Y_{i:j:T_G} / N_G$  is the mean of the criterion scores for participants who received the group-

administered treatment ( $i : j : T_G$ ) and

$$MS_{G/T_G} = \frac{\sum_{j=1}^J n (\bar{Y}_{j:T_G} - \bar{Y}_G)^2}{J - 1} \quad (6)$$

is the between-group mean square for these participants. It can be shown that the squared denominator of  $t_{APDF}$  estimates the sampling variance of the numerator assuming a correct model for the data is given by equations (1) and (2) and  $\alpha_{j:T_G}$  is random. Assuming that  $\varepsilon_{i:T_I} \sim N(0, \sigma_I^2)$ ,  $\alpha_{j:T_G} \sim N(0, \tau^2)$ , and  $\varepsilon_{i,j:T_G} \sim N(0, \sigma_G^2)$ , the estimated approximate degrees of freedom are

$$\hat{f}_2 = \frac{\left( \frac{MS_{S/T_I}}{N_I} + \frac{MS_{G/T_G}}{N_G} \right)^2}{\frac{\left( \frac{MS_{S/T_I}}{N_I} \right)^2}{N_I - 1} + \frac{\left( \frac{MS_{G/T_G}}{N_G} \right)^2}{J - 1}} \quad (7)$$

It should be noted that in using the Satterthwaite method, the distribution of the square of the denominator of  $t_{APDF}$  is approximated as a multiple of a chi-square distribution with degrees of freedom estimated by  $\hat{f}_2$ .

Based on simulated data, Myers et al. (1981) reported estimated Type I error rates for their APDF test, including results for  $J = 4$  and  $J = 8$  groups in the group-administered treatment. For both numbers of groups, estimated Type I error rates were very similar to the nominal level. While these results indicate that the APDF has adequate control of the Type I error rate when  $J \geq 4$ , it leaves open the question of how well the test works with a smaller number of groups and the discussion in Satterthwaite (1941) and results in Scariano and Davenport (1986) suggest the test may not control the Type I error rate for  $J \leq 3$ .

The discussion in Satterthwaite (1941) implies that the approximation of the square of the denominator of  $t_{APDF}$  by a multiple of a chi-square distribution improves as  $J - 1$  or  $N_I - 1$  increases and as

$$\frac{(N_I - 1)(n\tau^2 + \sigma_G^2)}{(J - 1)\sigma_I^2} \quad (8)$$

becomes closer 1.0. When there are two groups in the group-administered treatment,  $J - 1$  is as small as it possibly can be. In addition, calculating the ratio in equation (4) for conditions in which  $\sigma_I^2 = \tau^2 + \sigma_G^2$  shows that the ratio can be much larger than 1. Therefore, the discussion in Satterthwaite would lead one to expect that the APDF  $t$  test in Myers et al. (1981) would not work well when there are just two groups.

Scariano and Davenport (1986) studied Type I error rates for the APDF  $t$  test that Welch (1938) proposed as a solution to the Behrens-Fisher problem:

$$t = \frac{\bar{Y}_a - \bar{Y}_b}{\sqrt{\frac{S_a^2}{N_a} + \frac{S_b^2}{N_b}}}. \quad (9)$$

In  $t$ ,  $\bar{Y}_a$  and  $\bar{Y}_b$  are means for two individually administered treatments,  $S_a^2$  and  $S_b^2$  are the sample variances, and the square of the denominator estimates the sampling variance of the numerator. The distribution of the Welch  $t$  can be approximated by a  $t$  distribution with degrees of freedom approximated the by the Satterthwaite (1941) method. Thus, the Myers et al. (1981) APDF test and the Welch APDF solution to the Behrens-Fisher problem are both based on the same theoretical approach to approximating the sampling distribution of the test statistic.

Scariano and Davenport (1986) developed an analytic procedure for calculating the Type I error rate of the Welch APDF test and showed its Type I error rate can 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. In the Myers et al. (1981) APDF test, the sampling variances of the means are  $(n\tau^2 + \sigma_G^2)/N_G$  and  $\sigma_I^2/N_I$  and the degrees of freedom for estimates of these variance are  $J - 1$  and  $N_I - 1$ . When  $N_I = N_G$  and  $\sigma_I^2 = \tau^2 + \sigma_G^2$ , for example, the relationship will be negative and, when  $J \leq 3$ , the degrees of freedom will be small. Consequently, the APDF test may not

work well in these conditions. One purpose of the study is to study Type I error rates when  $J$  is small.

Satterthwaite (1941) showed how to approximate the distribution of a sum of two chi-square distributed random variables by another chi-square distribution. He determined the degrees of freedom for the approximating distribution by equating the mean and variance of the sum with the mean and variance of the approximating chi-square distribution. Thus, the Satterthwaite approach is a two-moment approach to determining the degrees of freedom. Scariano and Davenport (1986) developed a four-moment approach and showed analytically that it provides a more conservative test than does the two-moment approach. In the four-moment approach the estimated approximate degrees of freedom are

$$\hat{f}_4 = \frac{\left\{ \frac{u^2}{J-1} + \frac{1}{N_I-1} \right\}^3}{\left( \frac{u^3}{(J-1)^2} + \frac{1}{(N_I-1)^3} \right)^2} \quad (10)$$

where, in the groups-versus individuals design,

$$u = \frac{MS_{G/T_g} / N_G}{MS_{S/T_I} / N_I}. \quad (11)$$

A second purpose of the present study was to calculate the actual Type I error rate for the four-moment approach.

In Scariano and Davenport (1986), the two-moment approach was sometimes liberal when the four-moment approach was conservative. As a result, they suggested using an average of the estimated degrees of freedom produced by the two approaches. Thus, a third purpose was to analytically evaluate the actual Type I error rate for this averaged degrees of freedom approach.

An alternative to the preceding approaches is based on a mixed model with a proper inference space (McLean, Sanders, & Stroup, 1991) and Satterthwaite degrees of freedom.

When the restricted maximum likelihood estimate (RMLE) of  $\tau^2$  is larger than zero and there are an equal number of participants in the groups, the mixed model test is equivalent to the Myers et al. (1981) two-moment test. However, if the RMLE is zero,  $MS_{G/T_G}$  and  $MS_{S/G/T_G}$  are pooled and replace  $MS_{G/T_G}$  in equation (1). This statistic, which is equivalent to the Welch  $t$  test, is smaller than  $t_{APDF}$  and may be more conservative than the two-moment test. However, it tends to have larger degrees of freedom, which may make it more liberal than the two-moment test.

When there are an equal number of participants in the groups, the RMLE of  $\tau^2$  is zero when the method of moments estimate of  $\tau^2$  is  $\leq 0$  (McCulloch & Searle, 2001). The probability that the method of moments estimate of  $\tau^2$  is  $\leq 0$  is

$$prob\{MS_{G/T_G} \leq MS_{S/G/T_G}\} = prob\left\{F[J-1, J(n-1)] \leq \frac{1 - \rho_{ICC}}{\rho_{ICC}(n-1) + 1}\right\}, \quad (12)$$

where  $\rho_{ICC} = \tau^2 / (\tau^2 + \sigma_G^2)$ . Figure 1 displays the probability as a function of  $J$ ,  $\rho_{ICC}$ , and  $n$ . The probability can be quite substantial and, in some conditions, we would expect the mixed model test to perform differently than the two-moment, four-moment, and averaged degrees of freedom tests. Thus, a fourth purpose of the study is to compare these tests to the mixed model test.

The research was carried out in two studies. In the first study, actual Type I error rates were calculated for the two-moment approach, the four-moment approach, and the averaged degrees of freedom approach. In the second study, simulated data were used to estimate the actual Type I error rate for the mixed model approach as well as for the two-moment approach, the four-moment approach, and the averaged degrees of freedom approach. Taken together, the purposes of the studies were to compare Type I error rates for the two-moment, four-moment, averaged degrees of freedom, and mixed model approaches when the number of groups in the group administered treatment is small and

to study the influence of the number of groups, number of participants in a group, and intraclass correlation on the Type I error rates for these methods.

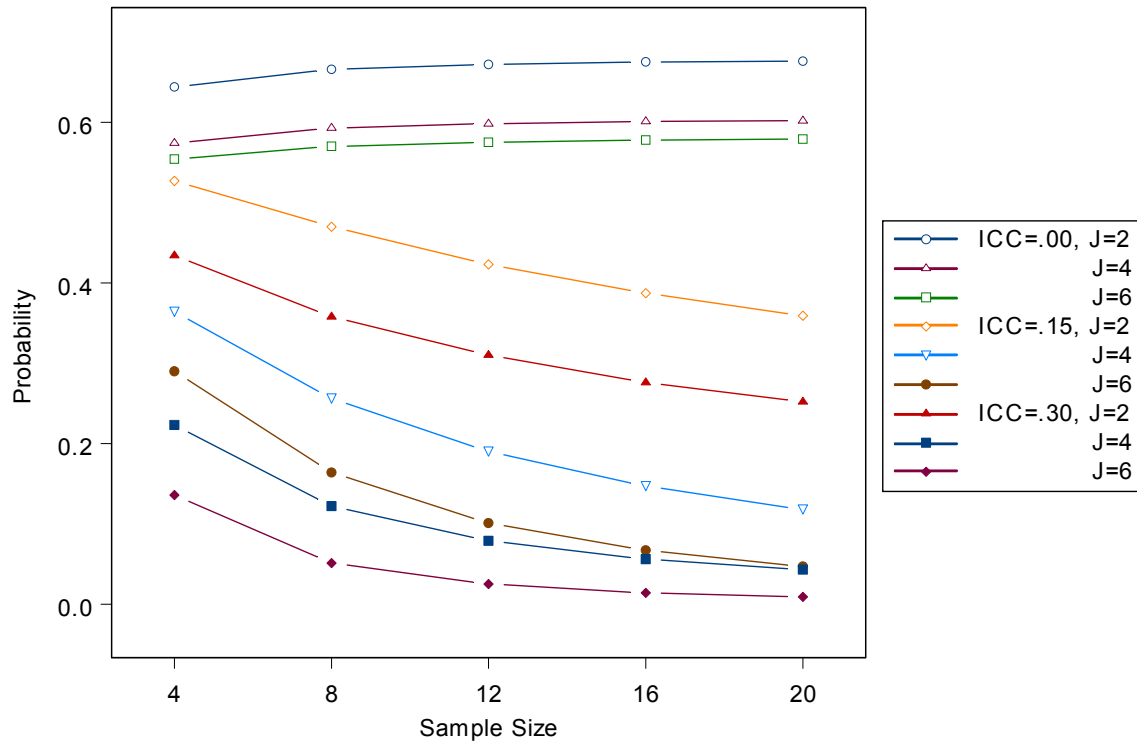


Figure 1. Probability of a Negative Estimate for  $\tau^2$

Methodology: Study 1

Actual Type I error rates were calculated for each condition in a 5 (Number of Groups)  $\times$  4 (Intraclass Correlation)  $\times$  15 (Number of Participants in a Group) completely crossed factorial design. The levels of the factors were  $J = 2$  to 6 for the number of groups;  $n = 3$  and 4, and 6 to 30 in steps of 2 for the number of participants in a group; and  $\rho_{ICC} = .00, .20, .40,$  and  $.80$  for the intraclass correlation. In all conditions,  $(\tau^2 + \sigma_G^2) / \sigma_I^2 = 1$  and, because the design was balanced across treatments,  $N_I = J(n)$ . For all calculations the nominal alpha level was .05. In the following, when we use the term Type I error rate without the actual or nominal modifier, we refer to the actual Type I error rate.

Calculating Type I Error Rates

Scariano and Davenport (1976) developed a method to calculate Type I error rates for the Welch  $t$  test. We applied their method, which we describe below, to the three APDF tests considered in this paper. It should be noted that although the method we applied was developed in the context of the Behrens-Fisher problem, that is, comparing means of independently distributed scores for two groups when the variance are not equal for the groups, we did not apply the method to the Behrens-Fisher problem. Rather we applied the method to comparison of means for two groups, when scores are not independently distributed within the sub-groups in the group-administered treatment. Thus, our work is not subject to Sawilowsky's (2002) criticisms of research on the Behrens-Fisher problem.

The Type I error rate for the APDF  $t$  test is

$$\Pr\left[t_{quasi}^2 > F_{\alpha,1,\hat{f}}\right] = \int_0^\infty \Pr\left[t_{quasi}^2 > F_{\alpha,1,\hat{f}} \mid u\right] g(u) du \tag{13}$$

where  $\hat{f}$  is the two-moment, four-moment, or averaged degrees of freedom and  $\alpha$  is the nominal Type I error rate. Cochran (1951) has shown that  $t_{quasi}^2$  is the ratio of  $Q$  to  $C$  where

$$Q \sim F_{1,m_1+m_2}, \quad m_1 = J - 1, \quad m_2 = N_I - 1,$$

$$C = \frac{(1+u)(m_1+m_2)}{(1+U)\left(\frac{m_1u}{U} + m_2\right)}, \tag{14}$$

and

$$U = \frac{(n\tau^2 + \sigma_G^2/N_G)}{\sigma_I^2/N_I}. \tag{15}$$

To facilitate numerical integration the variable  $u$  can be transformed to

$$s = \frac{\left(\frac{m_1u}{m_2}\right)}{\left(1 + \frac{m_1u}{m_2}\right)} \tag{16}$$

and the Type I error rate is found by numerically integrating

$$\int_0^1 \Pr[Q > C \times F_{\alpha, 1, \hat{\rho}} | s] f(s) ds \quad (17)$$

where

$$f(s) = \frac{\Gamma\left(\frac{m_1 + m_2}{2}\right) U^{m_2/2} s^{(m_1-2)/2} (1-s)^{(m_2-2)/2}}{\Gamma\left(\frac{m_1}{2}\right) \Gamma\left(\frac{m_2}{2}\right) [U(1-s) + s]^{(m_1+m_2/2)}}. \quad (18)$$

Numerical integration was performed using the trapezoid rule. For  $J = 2$  a singularity occurs at  $s = 0$ . Therefore, the limits of integration were .0001 and 1. The interval was divided into 1000 segments of equal width. For  $J = 3$  a removable singularity occurs at  $s = 0$ . For  $J \geq 3$  the limits of integration were 0 and 1 and this interval was also divided into 1000 segments. As a check on the calculations, Type I error rates were estimated by using simulated data with 100,000 replications. The results from the simulation were consistent with the results determined by numerical integration.

#### Results: Study 1

Figures 2 to 6 contain plots of the Type I error rates against size of groups. The five plots are for two, three, four, five, and six groups, respectively. Plots within a figure are organized by the intraclass correlation coefficient. Inspection of Figure 2 indicates that when there are two groups, the four-moment degrees of freedom should be used, except perhaps when  $\rho_{ICC} = 0$ . Then the averaged degrees of freedom might be used. When there are three groups (see Figure 3), the averaged degrees of freedom might be used at the risk of a slightly liberal test when  $\rho_{ICC}$  is at .20 or greater. The two-moment degrees of freedom results in a test that is too liberal and the four-moment degrees of freedom results in a test that is too conservative. When there are four groups (see Figure 4), the two-moment degrees of freedom provides a test that has a slight liberal tendency that increases as  $\rho_{ICC}$  get larger and as the size of the groups get larger. Use of the

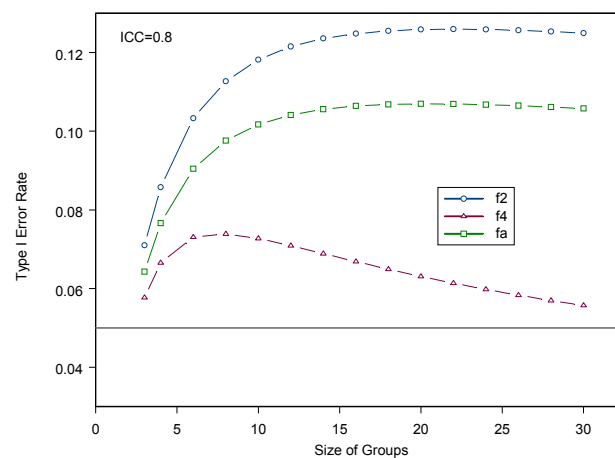
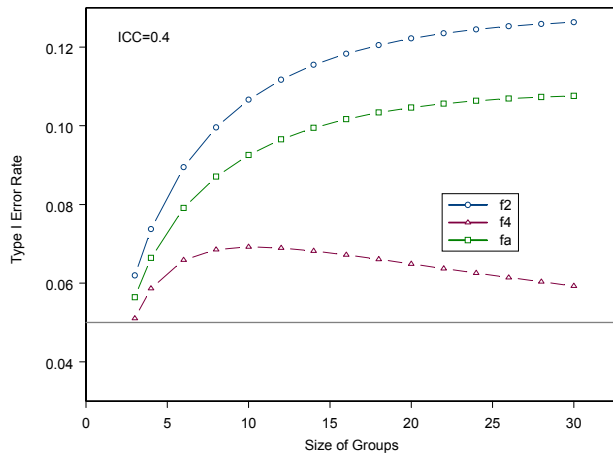
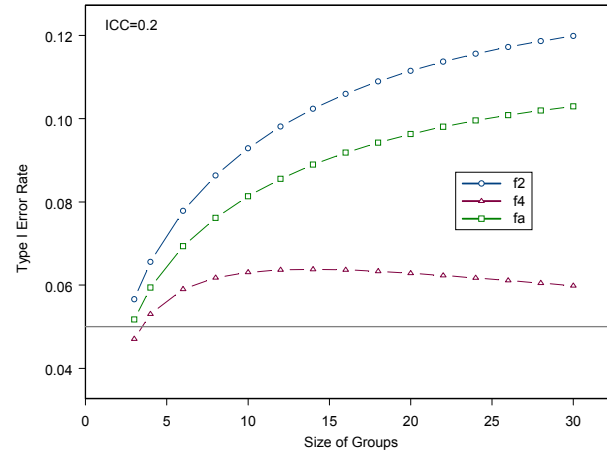
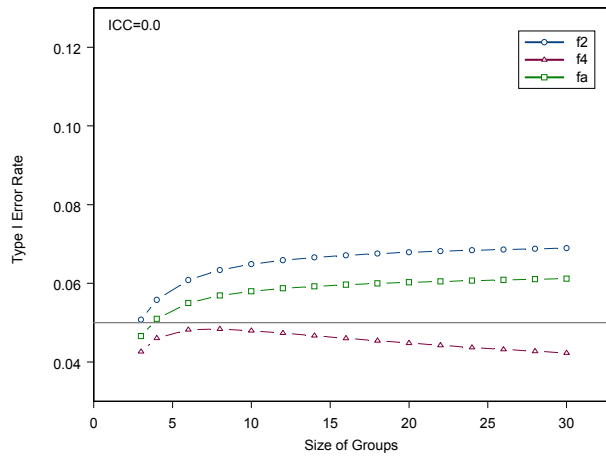


Figure 2. Plots of Type I Error Rates by Size of Group for Two Groups

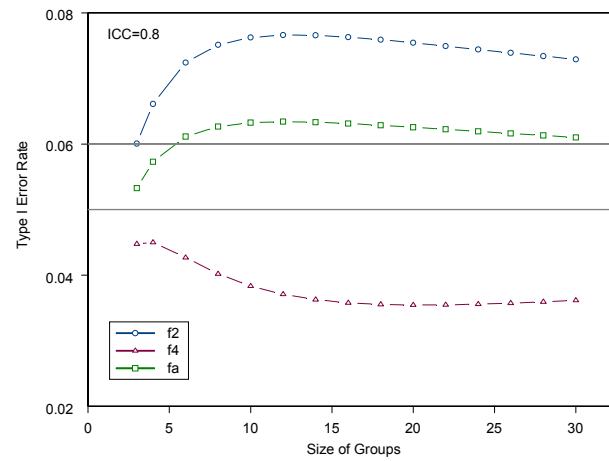
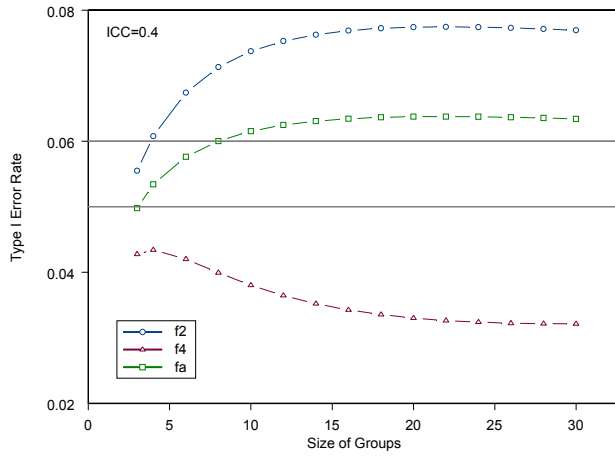
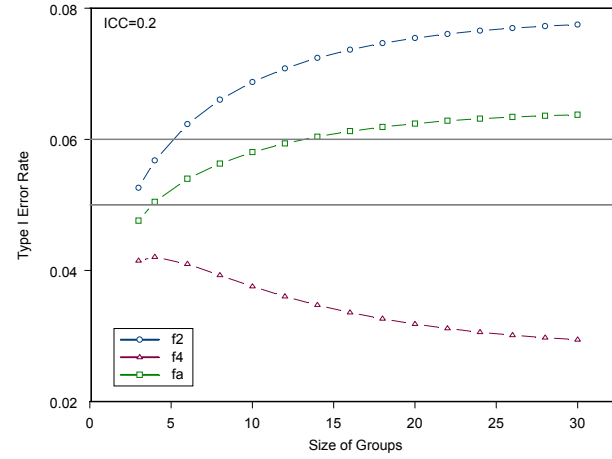
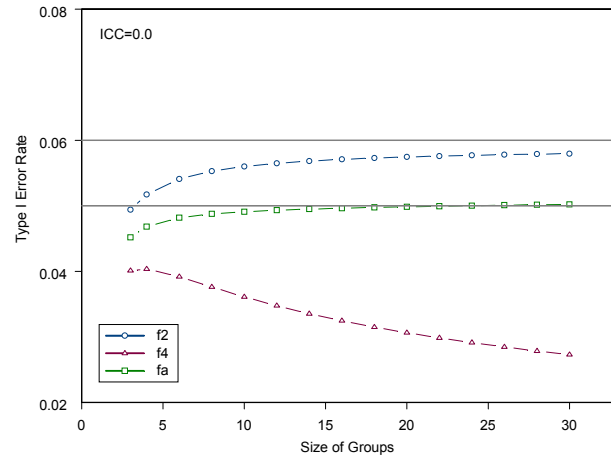


Figure 3. Plots of Type I Error Rates by Size of Group for Three Groups

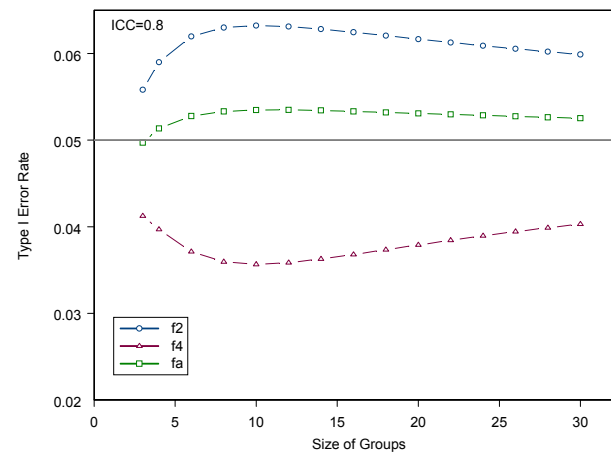
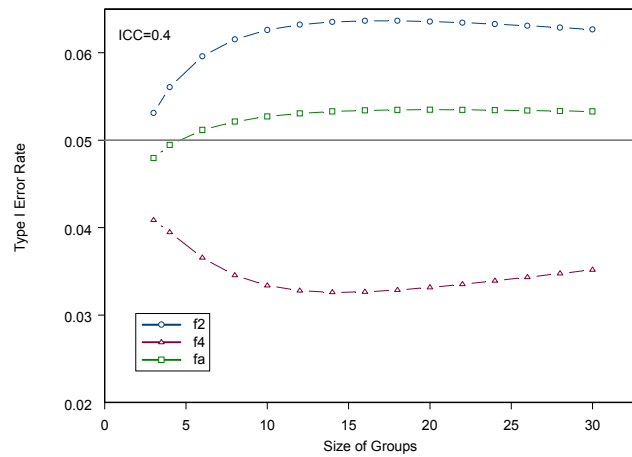
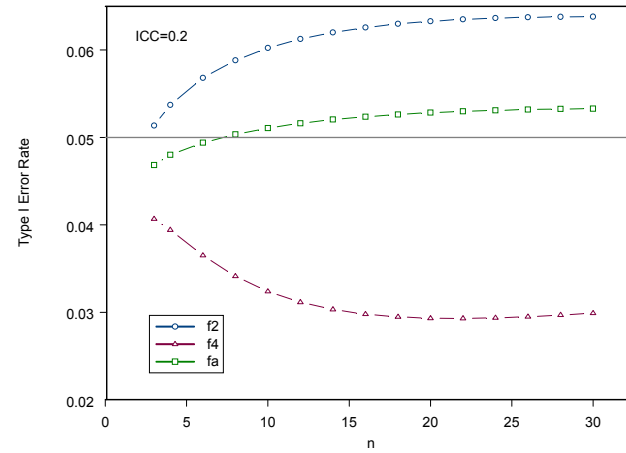
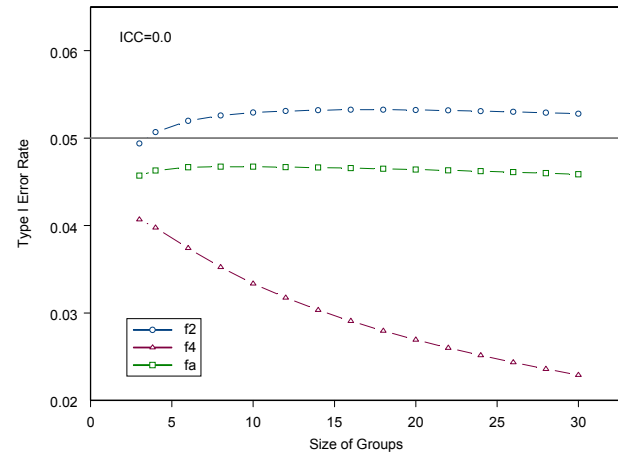


Figure 4. Plots of Type I Error Rates by Size of Group for Four Groups

averaged degrees of freedom provides a test that is slightly conservative when  $\rho_{ICC}$  is small, but controls the Type I error rate well as  $\rho_{ICC}$  increases. Plots for five or more groups (see Figures 5 and 6) are similar to those for four groups. However, the use of either the two-moment degrees of freedom or and average degrees of freedom provide reasonable control of the Type I error rate. Use of the former can result in a slightly liberal test, whereas use of the latter can result in a slightly conservative test.

#### Methodology: Study 2

As noted in the introduction, simulated data were used to compare the three APDF tests and the mixed model test. The design had four factors: the four tests, the number of groups, size of the groups, and level of the intraclass correlation. There were five levels of the number of groups,  $J = 2, 3, 4, 5,$  and  $6$ ; five levels of group size,  $n = 4, 8, 12, 16,$  and  $20$  subjects nested in the groups; and seven levels of intraclass correlation,  $\rho_{ICC} = .00$  to  $.30$  in steps of  $.05$ .

The simulation was carried out using the random number generation functions of SAS, Release 8.2. Scores for simulated participants in the individually administered treatment level were generated using the equation (1), where  $\mu_I$  was arbitrarily set at 100 and the  $\varepsilon_{i;T_I}$ s were pseudorandom standard normal deviates generated using RANNOR. Scores for simulated participants in the group-administered treatment level were generated using equation (2), where  $\mu_G$  was arbitrarily set at 100,  $\alpha_{j;T_G}$  was a pseudorandom normal deviate with mean zero and variance  $\tau^2$  and  $\varepsilon_{i;j;T_G}$  was a pseudorandom normal deviate with mean zero and variance  $\sigma_G^2$ . Each of the conditions was replicated 5,000 times and the Type I errors of the four tests were counted over the replications of each condition. The nominal type I error rate was  $.05$  in all conditions.

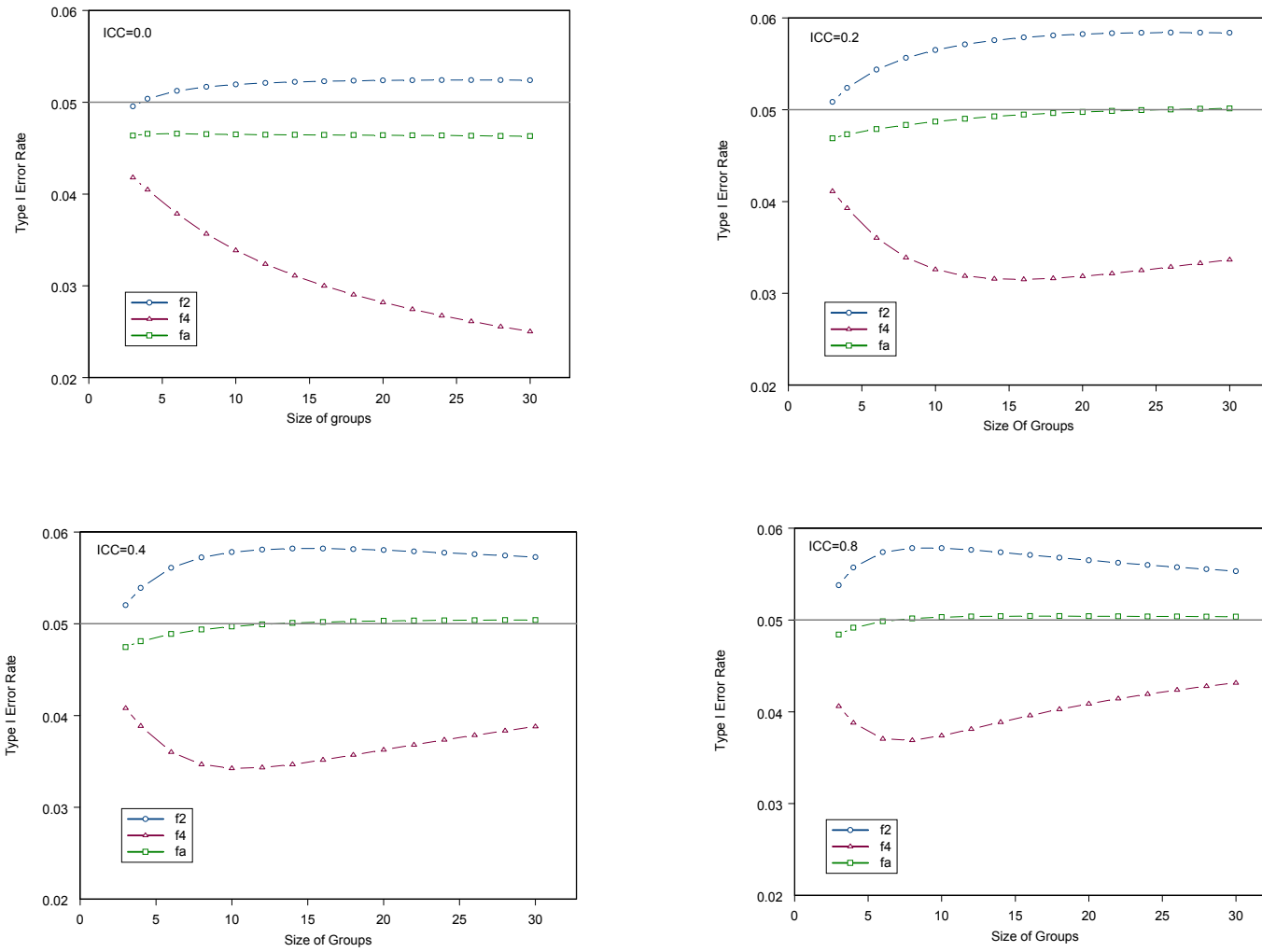


Figure 5. Plots of Type I Error Rates by Size of Group for Five Groups

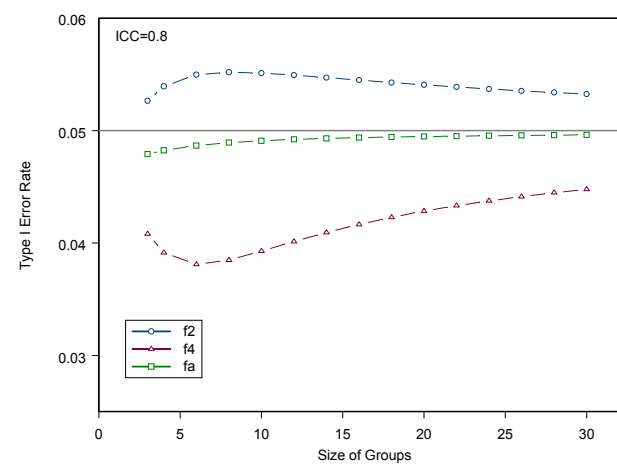
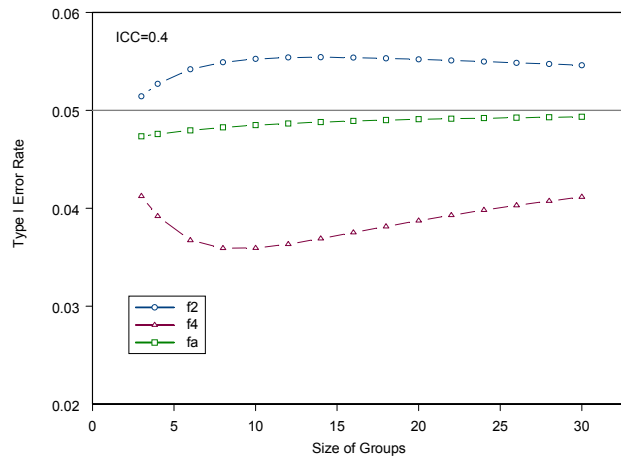
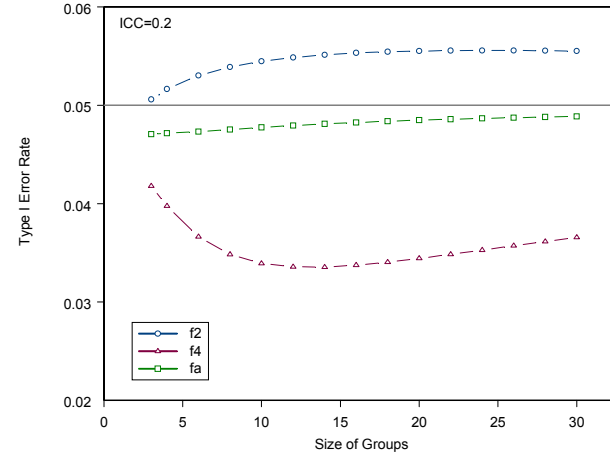
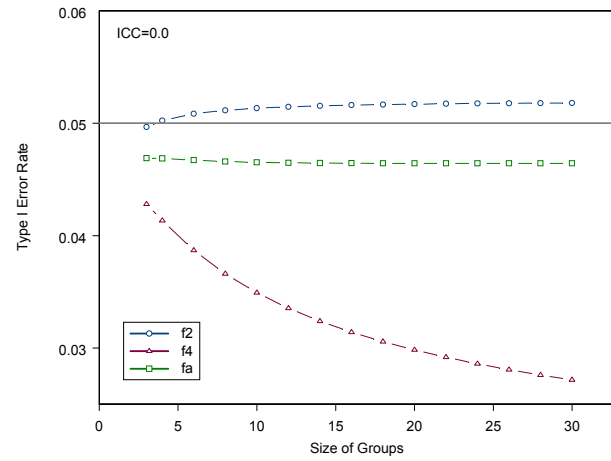


Figure 6. Plots of Type I Error Rates by Size of Group for Six Groups

The mixed model specified in equations (1) and (2) was implemented by using the following is a SAS program. The individually administered treatment is coded 1 on the TRT code.

```
PROC MIXED;
CLASS TRT GROUP;
MODEL SCORE=TRT/SOLUTION DDFM=SATTERTHWAITE;
RANDOM GROUP/GROUP=TRT;
REPEATED/GROUP=TRT;
PARMS (0) (1) (1) (1)/EQCONS=1
ESTIMATE 'COMP' TRT 1 -1;
```

The APDF tests are easily carried out in proc iml as the only required statistics are the means for the two groups, the variance for the treatment administered to individuals, and the mean squares within and between subgroups for the group-administered treatments.

Results: Study 2

The analytic results showed that, when there were two groups, the APDF test statistic with the four-moment degrees of freedom provided the best control of the Type I error rate. Figure 7 compares Type I error rate for the four-moment test and the mixed model test for  $\rho_{ICC} = 0.00$  and 0.30. Results for the APDF test statistic and the two-moment degrees of freedom are also included because the mixed model test is equivalent to the two-moment test when the estimate of  $\tau^2$  is non-zero. The four-moment degree of freedom test still provides the best control of the Type I error rate. The mixed model test is more conservative than the two-moment test and is substantially more conservative in conditions in which the probability of a zero estimate for  $\tau^2$  is large.

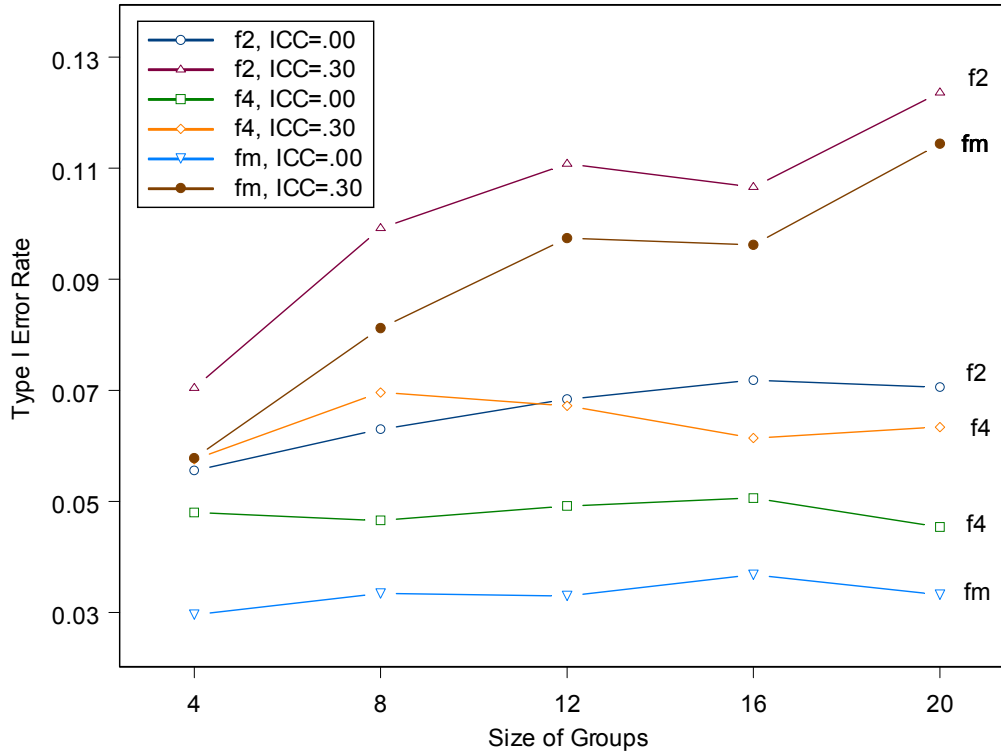


Figure 7. Type I Error Rates for Two Groups

When there were three groups, the analytic results showed that the APDF test statistic with the averaged degrees of freedom provided the best control of the Type I error rate. Type I error rates for the two-moment test tended to be too large. Figure 8 compares Type I error rates for the mixed model test and the APDF tests with two-moment and averaged degrees of freedom when  $\rho_{ICC} = .00$  and  $.30$ . The results indicate that the averaged degrees of freedom test still provides the best control of the Type I error rate.

According to the analytic results, there were four or more groups, both the two-moment and averaged degrees of freedom tests provided good control over the Type I error rate, with the former test being slightly more liberal. Type I error rates are depicted in Figure 9 for the two-moment, four-moment test and the mixed model tests for  $\rho_{ICC} = .00$  and  $.30$ . The results indicate that the mixed-model test is conservative and less adequate than the other tests when

$\rho_{ICC}$  is zero. Inspection of the results for other values of  $\rho_{ICC}$  indicate that when  $\rho_{ICC} = .10$  the performance of the averaged degrees of freedom and the mixed model tests is very similar and as  $\rho_{ICC}$  increases the Type I error rates for the mixed model test become slightly larger than those for the averaged degrees of freedom test. A similar pattern of results emerged for five or six groups. In particular, when  $\rho_{ICC}$  was near zero the mixed model test was too conservative.

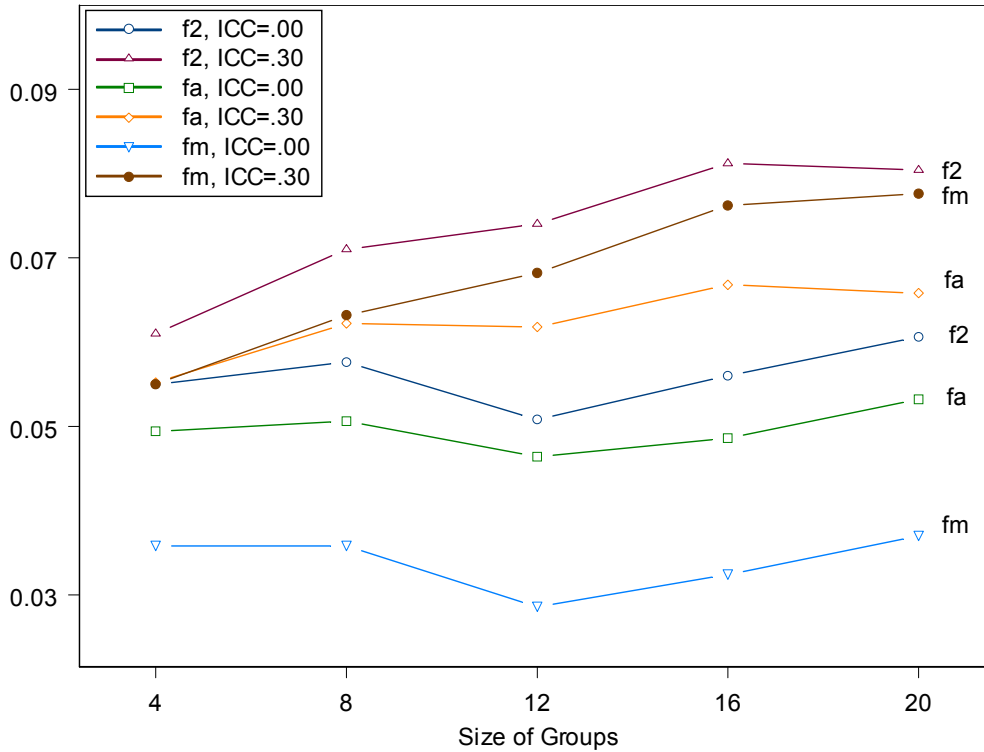


Figure 8. Type I Error Rates for Three Groups

### Conclusion

Myers et al. (1981) presented a two-moment approximate degrees of freedom test for use when one treatment is delivered to individual participants and one is delivered to groups of participants. The test was based on results in Satterthwaite (1941). Simulation results indicated that the test provided good control of the Type I error rate for both four groups and eight groups of participants. Satterthwaite (1941) and Scariano and Davenport (1986) studied a two-moment approximate degrees of freedom test for a design in which both treatments are delivered

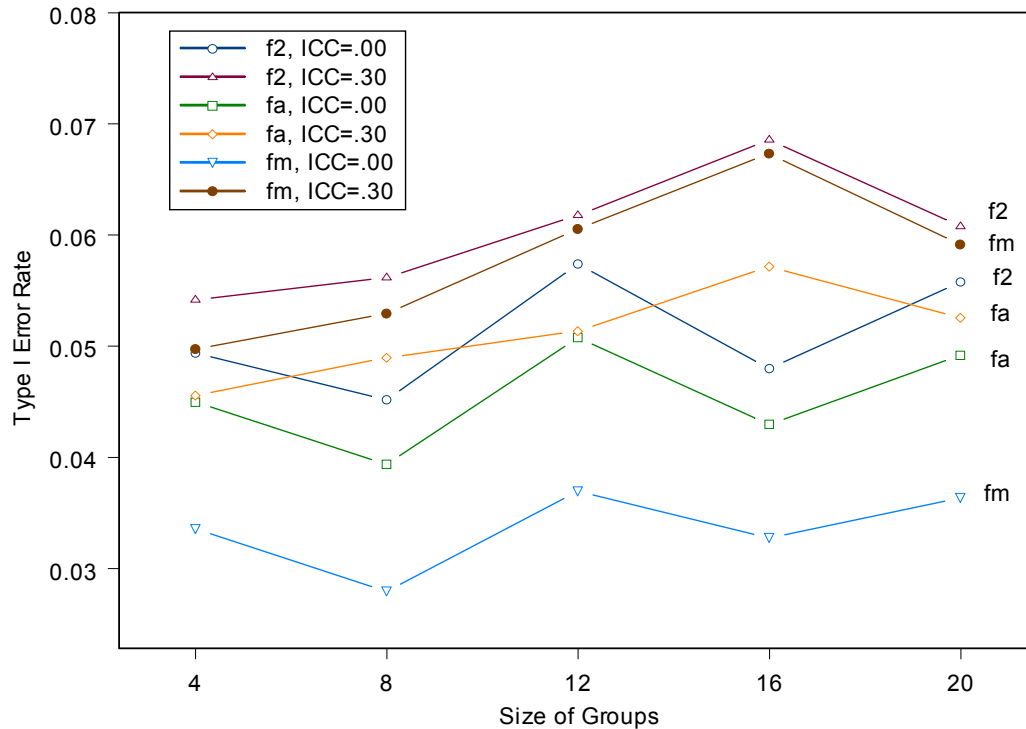


Figure 9. Type I Error Rates for Four Groups

individually. Discussion in Satterthwaite and results in Scariano and Davenport suggest that the Myers et al. test may not perform well when the number of groups is smaller than four. Using an analytic procedure developed by Scariano and Davenport, we showed that the Myers et al. test could provide relatively poor control of the Type I error rate when there are two or three groups. Using results presented in Scariano and Davenport, we developed two alternatives to the Myers et al. (1981) test, a four-moment approximate degrees of freedom test and an averaged degrees of freedom test. Using the analytic procedure developed by Scariano and Davenport, Type I error rates were calculated for all three test in a wide range of conditions in which the design was balanced across the individually administered treatment and the group-administered treatment and across the groups in the group-administered treatment. We also estimated Type I error rates for the mixed model test and the three APDF tests. The results indicated that the four-moment test should be used when the group-administered treatments are delivered to two groups and the averaged degrees of freedom test should be used when the group-administered treatments are

delivered to three groups. When there are between four and six groups, we recommend using the averaged degrees of freedom test. However, because (a) this test is slightly conservative, with a Type I error rate between 0.045 and 0.050, and (b) the two-moment test is slightly liberal but tends to keep the Type I error rate below 0.06, some may prefer the two-moment test. Even when there are four or more groups, we do not recommend the mixed model test because of its conservative tendency when the intraclass correlation coefficient is small. These recommendations are summarized in the Table 1

Table 1

Recommended Tests by the Number of Groups in the Group-Administered Treatment

Number of Groups	Recommended Test
2	Four-Moment Test
3	Averaged Degrees of Freedom Test
4-6	Averaged Degrees of Freedom Or Two Moment Test

When there are two groups in the group-administered treatment, the four-moment test provides better control of the Type I error rate than do the other tests. Nevertheless researchers should be cautious about using a groups-versus-individuals design with two groups because such designs will provide relatively low power. The true degrees of freedom for the four-moment test is

$$f_4 = \frac{\left\{ \frac{U^2}{J-1} + \frac{1}{N_I-1} \right\}^3}{\left( \frac{U^3}{(J-1)^2} + \frac{1}{(N_I-1)^3} \right)^2} \quad (19)$$

where  $U$  is defined in equation (15). Calculations show that  $f_4$  approaches 1.0 from above as  $U$  increases. Thus in many situations the degrees of freedom for the four-moment test will be very

small and this will have a negative impact on power. In addition, substituting population parameters for sample statistics in the Myers et al. (1981)  $t$  statistic, we have

$$\frac{\mu_I - \mu_G}{\sqrt{\frac{\sigma_I^2}{N_I} + \frac{\sigma_G^2}{N_G} + \frac{\tau^2}{J}}} \quad (20)$$

Therefore even as the two sample sizes increase power will not go to 1.0 if  $\tau^2 \neq 0$ . Finally, the fact that the Type I error rate for the four-moment test declines as  $n$  increases suggests power will decline as  $n$  increases because the test becomes more conservative. The predicted low power and decline in power as  $n$  increases were borne out by simulation studies. For example

when  $\sigma_I^2 = \sigma_G^2 + \tau^2 = 1$ ,  $\rho_{ICC} = .2$ , and  $\mu_G - \mu_I = .8$ , estimated power was .23, .21 and .19 as  $n$  increased from 6 to 18 in steps of 6. Comparison of these results to the power of an independent samples  $t$  test with the same overall sample size indicates how much lower power is when a groups-versus-individuals design is used. Note that because  $\sigma_I^2 = \sigma_G^2 + \tau^2 = 1$ ,  $\mu_G - \mu_I = .8$  corresponds to Cohen's large effect size. Also as  $n$  increases from 6 to 18 the sample size in a treatment increases from 12 to 36 in steps of 12. For an independent sample  $t$  test with an effect size equal to .8, power is .47, .77, and .92 as  $n$  increase from 12 to 36 by 12.

When there are three groups and the averaged degrees of freedom approach is used, power does not decline as  $n$  increases, but power can still be quit low and does not increase quickly as  $n$  increase. As  $n$  increased from 4 to 12, so that the overall sample size remained the same as in the conditions on which power results were reported for  $J = 2$ , estimated power was .29, .36, and .40 when  $J$  was 3.

As suggested by equation (20), power continues to increase as  $J$  increases. For example with  $J = 6$ , as  $n$  increased from 2 to 6 in steps of 2 estimated power was .41, .58, and .68 using the averaged degrees of freedom test. Thus when the groups-versus-individuals test is used, it is important to have as many groups as possible and may be more important to have more groups than to have more participants per group.

At least four lines of additional research are attractive. First, the performance of the tests under non-normality should be investigated and if performance is poor developing the test statistic and degrees of freedom using robust estimates of the means and mean squares is of interest. Second, performance of the four tests when the design is unbalanced across the individually administered treatment and the group-administered treatment, but balanced across groups in the group-administered treatment might be investigated. Third, calculating the averaged degrees of freedom by differentially weighting the two-moment and four-moment degrees of freedom might be investigated when there are four or more groups. Weighting the two-moment degrees of freedom more heavily will reduce the slight conservative tendency of the averaged degrees of freedom test. In general, more extensive studies of power than we have conducted would be worthwhile. Fourth, the three APDF tests should be generalized for use when the design is not balanced across groups in the group-administered treatment and Type I error rates for these tests and the mixed model test should be investigated.

## 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.
- 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, (2<sup>nd</sup> ed.) (pp. 541-559). New York: Wiley.
- Cochran, W. G. (1951). Testing a linear relationship among variances. Biometrics, 7, 17-32.
- McCulloch, C. E., & Searle, S. R. (2001). Generalized, linear, and mixed models. New York: Wiley.
- McLean, R. A., Sanders, W. L., & Stroup, W. W. (1991). A unified approach to mixed linear models. American Statistician, 45, 54-64.

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.

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.

Sawilowsky, S. (2002). The probable difference between means when  $\sigma_1 \neq \sigma_2$ : The Behrens-Fisher problem. Journal of Modern Applied Statistical Methods, 1, 461-472.

Scheffe, H. (1959). The analysis of variance. New York: Wiley.

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

Footnote

<sup>1</sup> Tables containing Type I error rates for all conditions in Study 1 and Study 2 are available at <http://plaza.ufl.edu/algina/index.programs.html>