Human-Friendly Scheffé Comparisons, or the Art of Complex Multiple Comparisons

Gordon P. Brooks
Ohio University

Nina Adjanin
Northwest Missouri State University

Frank Oppong
Ohio University

Yuqing Liu
Ohio University

Researchers learn that the Scheffé (1953) MCP lacks power because it adjusts for all possible comparisons—consequently, few use it. However, only Scheffé guarantees congruence: finding a significant comparison when the omnibus ANOVA is significant—and not finding one when ANOVA is nonsignificant (Maxwell et al., 2018). A maximum Scheffé comparison can be calculated to provide the set of coefficients that maximally differentiates some combination of groups on the dependent variable (Keppel & Wickens, 2004; Williams, 1978). Unfortunately, coefficient weights from this maximum comparison are often uninterpretable. Therefore, we have developed a Shiny app to identify maximum and other Barcikowski "human-friendly" comparisons that may actually be meaningful. We used Monte Carlo simulations to investigate robustness, power, and congruence (Kirk, 2013) of these human-friendly comparisons and the relatively unknown Brown-Forsythe unequal-variance adjustment to Scheffé. We report results regarding Bonferroni-adjusted normality tests and zero-adjusted Levene homoscedasticity tests for assumptions in ANOVA.

Most applied researchers are familiar with multiple comparison procedures (MCPs) used to explore group mean comparisons following a statistically significant one-way analysis of variance (ANOVA) or main effect in factorial ANOVA. Commonly used MCPs include pairwise comparison techniques like Tukey-Kramer and Games-Howell. One of the oldest MCPs is attributed to Scheffé (1953) but relatively few researchers use it, however, because it is well-known to lack the statistical power of other MCPs for the pairwise post hoc comparisons that most researchers use (and that most statistics programs provide) following a statistically significant ANOVA. That is, because Scheffé adjusts for all possible comparisons (i.e., all pairwise and non-pairwise—or complex—comparisons), it has lower statistical power relative to techniques that only adjust alpha for pairwise comparisons.

It is noteworthy that Scheffé provides the opportunity to test many complex contrasts or comparisons with appropriate alpha adjustment for as many tests as desired. But more noteworthy is that only Scheffé guarantees congruence to find a statistically significant comparison (which is typically a complex comparison) whenever the omnibus or main-effect ANOVA is statistically significant (Kirk, 2013)—and conversely, not find one when ANOVA is not significant (Maxwell et al., 2018). That is, a Scheffé comparison can be calculated that provides contrast coefficients for the means that maximally separate some combination of the groups (Keppel & Wickens, 2004; Williams, 1978) with the same resulting significance *p*-value as the omnibus Fisher *F* ANOVA—and, therefore, the same Type I error and statistical power rates, as well. We believe researchers may be missing potentially useful exploratory information by not examining this maximum comparison. Unfortunately, none of the well-known statistical programs calculate it. One of our purposes here is to share an online `R Shiny` app that provides this comparison along with others that may be of interest to researchers (link below).

(https://72x6cr-gordon-brooks.shinyapps.io/Human_Friendly_Contrasts/)

Unfortunately, this maximum Scheffé comparison may be uninterpretable. Therefore, Barcikowski (personal communication, 2000) suggested a method by which researchers can identify a maximum "human-friendly" comparison that serves to approximate the Scheffé maximum comparison with coefficients that are reasonably interpretable. We have also implemented in the `R Shiny` app a method to identify this maximum human-friendly comparison as well as others that may be meaningful to and interpretable by a human researcher. We have named these "Barcikowski human-friendly comparisons" in memory of Robert Barcikowski, who introduced the idea to the authors and whose original FORTRAN code has been adapted into `R` (with permission).

The purposes of this research were (1) to create and share an online app that finds the Scheffé maximum comparison, as well as the easier-to-calculate "normalized maximum posttest contrast" described by Hollingsworth (1978) and the maximum Barcikowski human-friendly comparisons, (2) to test the comparisons statistically with both Scheffé's *F*-test and Brown-Forsythe's adjustment to Scheffé for unequal variances, and (3) to perform Monte Carlo simulations that test the comparative robustness (Type I error rates), statistical power, and, especially, congruence of these methods.

Background

Applied researchers commonly compare group means for both experimental and nonexperimental purposes. While there are other approaches, a common method put forward by textbooks is to perform a one-way ANOVA and its obligatory assumptions tests, and then, if that omnibus ANOVA is statistically significant, to follow up with an appropriate post hoc MCP (but it is noted that many of the MCPs, including Scheffé, can also be used directly without protection from a significant ANOVA). Using a post hoc MCP allows researchers to investigate which particular group means may differ based on the significant ANOVA, which simply suggests that some combination of the population means differs. The most common implementation of MCPs is via pairwise approaches, where each group mean is compared to each of the other group means (sometimes all possible pairwise comparisons, sometimes a subset—for example each treatment group only compared to a control group). Levels of significance or *p*-values are adjusted through most of the MCP techniques to control familywise error rate. Statistics computer programs typically offer many such pairwise MCPs, most for equal variances (e.g., Tukey-Kramer) and some not (e.g., Games-Howell).

For example, the set of six pairwise comparisons used to compare all possible paired two-group pairwise comparisons from among four groups would look like this:

$$\begin{aligned}
\psi_1 &= 1\mu_1 + (-1)\mu_2 + 0\mu_3 + 0\mu_4 \\
\psi_2 &= 1\mu_1 + 0\mu_2 + (-1)\mu_3 + 0\mu_4 \\
\psi_3 &= 1\mu_1 + 0\mu_2 + 0\mu_3 + (-1)\mu_4 \\
\psi_4 &= 0\mu_1 + 1\mu_2 + (-1)\mu_3 + 0\mu_4 \\
\psi_5 &= 0\mu_1 + 1\mu_2 + 0\mu_3 + (-1)\mu_4 \\
\psi_6 &= 0\mu_1 + 0\mu_2 + 1\mu_3 + (-1)\mu_4
\end{aligned}$$

These coefficient weights (for example, [1 −1 0 0] for ψ_1) would be used to calculate mean differences between group 1 and group 2, because group 3 and group 4 coefficients are zero and therefore not included in the comparison. Similarly, the coefficients for ψ_4, [0 1 −1 0], would be used to compare mean differences between group 2 and group 3. The groups with zero coefficients are always excluded from the calculation of comparisons.

However, complex, non-pairwise comparisons would include coefficients to calculate comparisons using combinations of multiple groups (essentially as comparisons between "positive" and "negative" coefficient groups, similar to the way +1 and −1 divide two groups into positive and negative groups in the pairwise comparisons). For example, the set of three Helmert contrasts for four groups would include two complex comparisons (ψ_7 and ψ_8):

$$\begin{aligned}
\psi_7 &= 1\mu_1 + (-\tfrac{1}{3})\mu_2 + (-\tfrac{1}{3})\mu_3 + (-\tfrac{1}{3})\mu_4 \\
\psi_8 &= 0\mu_1 + 1\mu_2 + (-\tfrac{1}{2})\mu_3 + (-\tfrac{1}{2})\mu_4 \\
\psi_9 &= 0\mu_1 + 0\mu_2 + 1\mu_3 + (-1)\mu_4
\end{aligned}$$

But this set of Helmert contrasts does not constitute all possible Helmert-type comparisons. If Helmert-type comparisons are being used post hoc, there may be no a priori reference groups (e.g., group 1 in ψ_7 above or group 2 in ψ_8 above). For example, from an exploratory post hoc perspective, a researcher may also want to look for any such Helmert-type comparison that is informative, such as these, where group 2 is beginning reference group as group 1 was in ψ_7:

$$\begin{aligned}
\psi_{10} &= (-\tfrac{1}{3})\mu_1 + 1\mu_2 + (-\tfrac{1}{3})\mu_3 + (-\tfrac{1}{3})\mu_4 \\
\psi_{11} &= (-\tfrac{1}{2})\mu_1 + 0\mu_2 + (-\tfrac{1}{2})\mu_3 + 1\mu_4 \\
\psi_{12} &= (-1)\mu_1 + 0\mu_2 + 1\mu_3 + 0\mu_4
\end{aligned}$$

Similarly, group 3 could be 1 in ψ_{11} while group 4 has a coefficient of (−½); or group 3 could be the beginning reference group, and so forth. As might become quickly obvious, there would be many combinations of complex non-pairwise coefficients. The critical feature is that the positive and negative combinations each sum to the same value, typically 1.0 (additional requirements are required for sets of orthogonal contrasts). For example,

$$\begin{aligned}
\psi_{13} &= 1\mu_1 + (-\tfrac{1}{2})\mu_2 + (-\tfrac{1}{4})\mu_3 + (-\tfrac{1}{4})\mu_4 \\
\psi_{14} &= \tfrac{1}{2}\mu_1 + \tfrac{1}{2}\mu_2 + (-\tfrac{1}{2})\mu_3 + (-\tfrac{1}{2})\mu_4 \\
\psi_{15} &= 0.2\mu_1 + 0.8\mu_2 + (-0.2)\mu_3 + (-0.7)\mu_4
\end{aligned}$$

It is argued by some that complex comparisons are not useful (e.g., Schmid, 1977). For example, comparisons like ψ_{15} above may be difficult for a researcher to interpret. While many of the possible complex comparisons may be uninterpretable, sometimes such comparisons are worth examining. For example, the first Helmert-style complex comparison described above, ψ_7, might describe a comparison of a control (group 1) with the average of three treatment groups (e.g., “nothing” versus “something”). Similarly, the second complex comparison, ψ_8, may refer to a comparison—after dropping a placebo group—between a low-dose group (level 2) and the average of the two higher dose groups, levels 3 and 4 (e.g., “some” versus “more”). In a study with two comparison and two control groups, contrast ψ_{14} above might be useful from the perspective of the average of control groups versus the average of treatment groups.

Many methodological scholars have studied robustness and power of MCPs. Scheffé often fares poorly because of its extreme adjustment to control Type I error for all comparisons, which results in lower statistical power than other procedures—especially when used only with pairwise comparisons. Scheffé also is not robust to violations of the homoscedasticity assumption.

Scheffé is unique among MCPs because it allows and adjusts for infinite possible pairwise and complex comparisons. Textbooks often describe the procedure probably because it allows for all possible comparisons—but do not usually recommend it for practice due to power and because pairwise comparisons are often preferred (for interpretability). Scheffé can be useful when researchers begin with research questions that imply multiple a priori complex contrasts and they want to be conservative and control Type I error inflation for multiple tests. It is also useful when researchers have no theoretical expectations for the relationships between groups and wish to explore all differences among them, including complex differences.

Scheffé Maximum Comparison

The Scheffé maximum comparison (which we will call SchefféMax), for both equal and unequal sample sizes, is found using the formula (Keppel & Wickens, 2004; Williams, 1979):

$$c_i' = \frac{N_i(\bar{X}_i - \bar{T})}{\sqrt{SSB}}$$

where c_i is the comparison coefficient for group/level i, N_i is the sample size in for group/level i, $\bar{T}$ is the dependent variable grand mean (total) for the entire sample, $\bar{X}_i$ is the dependent variable mean for group/level i, and SSB is the sum of squares between groups from ANOVA. For example, for example data where $N_i = 10$ for all groups, $\bar{T} = 49.3$, SSB = 698.4, and the group means, $\bar{X}_i$, are 54.9, 45.9, 51.7, and 44.7, the maximum comparison coefficients, c_i, for the four groups in this example are calculated as follows:

$$\begin{aligned}
c_1 &= 10(54.9-49.3) / 26.43 = 56 / 26.43 = 2.119 \\
c_2 &= 10(45.9-49.3) / 26.43 = -34 / 26.43 = -1.286 \\
c_3 &= 10(51.7-49.3) / 26.43 = 24 / 26.43 = 0.908 \\
c_4 &= 10(44.7-49.3) / 26.43 = -46 / 26.43 = -1.742
\end{aligned}$$

From one perspective, SchefféMax has the same statistical power as the ANOVA F test. That is, we know that SchefféMax will always be congruent with ANOVA. As a result, both the Type I error rate and the statistical power for SchefféMax will be the same as those of the Fisher F-statistic, guaranteeing a significant comparison with a significant ANOVA—which other MCPs cannot.

Hollingsworth Maximum Comparison

Hollingsworth (1978; also see Williams, 1979) proposed the following formula for the contrast, which we will call HollingsworthMax:

$$c_i = \frac{\sqrt{\tilde{N}}(\bar{X}_i - \bar{T})}{\sqrt{SSB}}$$

where $\tilde{N}$ is the harmonic mean (which equals $N_{per\ group}$ when all N_i are the same). Hollingsworth also showed that the comparison Sum of Squares Between (SSB_i) can be calculated as:

$$SSB_i = \frac{N\left(\sum_{i=1}^{k} c_i \bar{X}_i\right)^2}{\sum c_i^2}$$

where *k* is the total number of levels. Further the proportion of between group variation that can be attributed to each comparison can be calculated as follows, which is simply c_i^2:

$$c_i^2 = \frac{(\bar{X}_i - \bar{T})^2}{\sum_{i=1}^{k}(\bar{X}_i - \bar{T})^2}$$

where the symbols are used the same as above. In the example above, HollingsworthMax results in the same comparison coefficients as SchefféMax because the group sizes are all equal.

Barcikowski Human-Friendly Comparisons

If researchers have the time available, they can create an infinite number of non-pairwise comparisons, most of which would be meaningless from a practical or theoretical perspective. The core idea of the Barcikowski comparisons is to identify sets of potential comparisons that are reasonably interpretable. Barcikowski comparisons might be described as "Helmert-plus" comparisons. That is, Barcikowski comparisons constitute all possible Helmert-type comparisons plus other "reasonable" ways to compare complex combinations of groups. For example, with five groups the first set of Helmert-type comparisons might be [−1 ¼ ¼ ¼ ¼], [0 −1 ⅓ ⅓ ⅓], [0 0 −1 ½ ½], and [0 0 0 −1 1]. However, the process then creates all other sets of possible Helmert comparisons with a different beginning reference group each time. For example, another set of comparisons would be:

[¼ −1 ¼ ¼ ¼], [⅓ 0 −1 ⅓ ⅓], [½ 0 0 −1 ½], and [1 0 0 0 −1].

After creating the Helmert-type comparisons, additional sets of reasonable comparisons are created that do not follow the Helmert pattern. For example, with five groups such comparisons might include comparing the first two groups to the other three (i.e., [½ ½ −⅓ −⅓ −⅓]), or maybe comparing two groups to two groups while leaving one out of the comparison (i.e., [½ ½ 0 −½ −½]). The former comparison might imply two particular follow-up comparisons, both of which would fall under the Helmert-type designation: [1 −1 0 0 0] and [0 0 1 −½ −½], the latter of which then might be followed by the pairwise comparison between groups 4 and 5. Table 1 provides examples of the patterns used for the comparison coefficients for eight groups. In total there are 3025 unique sets of coefficients obtained and tested for the eight-group scenario, which are all permutations of the 16 examples in Table 1. There are 6 unique sets of coefficients for three groups, 25 for four groups, 90 for five groups, 301 for six groups, and 966 sets of coefficients for seven groups. Unfortunately, beyond eight groups the number of sets of coefficients requires extensive computing time.

Table 1. Example Barcikowski Comparison Coefficients for Eight Levels/Groups

	Comparison Coefficients for Each Group							
Comparison	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	¼	¼	¼	¼	−¼	−¼	−¼	−¼
2	⅓	⅓	⅓	−⅕	−⅕	−⅕	−⅕	−⅕
3	⅓	⅓	⅓	0	−¼	−¼	−¼	−¼
4	⅓	⅓	⅓	0	0	−⅓	−⅓	−⅓
5	½	½	−⅙	−⅙	−⅙	−⅙	−⅙	−⅙
6	½	½	0	−⅕	−⅕	−⅕	−⅕	−⅕
7	½	½	0	0	−¼	−¼	−¼	−¼
8	½	½	0	0	0	−⅓	−⅓	−⅓
9	½	½	0	0	0	0	−½	−½
10	1	−⅐	−⅐	−⅐	−⅐	−⅐	−⅐	−⅐
11	1	0	−⅙	−⅙	−⅙	−⅙	−⅙	−⅙
12	1	0	0	−⅕	−⅕	−⅕	−⅕	−⅕
13	1	0	0	0	−¼	−¼	−¼	−¼
14	1	0	0	0	0	−⅓	−⅓	−⅓
15	1	0	0	0	0	0	−½	−½
16	1	0	0	0	0	0	0	−1

Therefore, we use Helmert-type comparisons and those additional reasonable comparisons that maintain relatively interpretable fractions for the coefficients. This process results in a subset of all possible Scheffé-like comparisons that have the most reasonable potential for interpretation. From an exploratory perspective, the emphasis of the Barcikowski comparisons is to test the maximum plus all possible reasonable comparisons. Ultimately, the researcher will need to determine whether any of the statistically significant maximum comparisons make sense. The maximum Barcikowski comparison (which we call BarcikowskiMax) and other next-most informative comparisons are identified by testing all these Helmert-plus comparisons and then sorting the comparisons according to explanatory power (i.e., proportion of between sums of squares explained for each comparison).

Methods

We created `R` code that identifies the SchefféMax comparison and analyzes the comparisons for an applied researcher's dataset. This `R` code is available for use by applied researchers in the `R Shiny` app created by the authors (https://72x6cr-gordon-brooks.shinyapps.io/Human_Friendly_Contrasts/). The code also identifies the HollingsworthMax comparison, which typically differs slightly from SchefféMax in unbalanced designs. Finally, we have also implemented a method to identify the BarcikowskiMax comparison and, because there can be multiple statistically significant Barcikowski comparisons, the most informative Barcikowski comparisons (i.e., a particular subset of all possible Scheffé comparisons). This allows review of multiple exploratory comparisons, the `R Shiny` app reports all Barcikowski comparisons that were statistically significant at α=0.15, sorted by the comparisons' sums of squares between.

Study Design

This study used Monte Carlo simulation methods in `R` to generate and analyze data for many conditions. We ran the simulations in two phases. In Phase 1, we generated 100,000 samples across 27 four-group robustness conditions and 6 statistical power conditions. In Phase 2, we sought to extend the results of Phase 1 with five groups. In this phase, we generated 10,000 samples for 49 five-group robustness conditions and 8 statistical power conditions.

Congruence. All conditions were used to investigate the congruence of the three maximum methods with the statistical significance of the omnibus Fisher *F* ANOVA. However, the primary purpose was to determine how congruent BarcikowskiMax is. We also investigated the congruence for the four most informative comparisons by examining how often those Barcikowski comparisons resulted in the same decision about the null hypothesis as the Fisher *F* (i.e., not the same *p*-value).

Robustness. We also investigated the robustness (i.e., control of Type I error, especially under violations of assumptions) of these maximum comparisons across the many conditions. We compared HollingsworthMax and the four most informative BarcikowskiMax comparisons with the SchefféMax comparison. Most importantly, we investigated the robustness of the Brown-Forsythe adjustment to the Scheffé MCP significance test. We defined robustness using Bradley's (1978) stringent $\alpha \pm (0.1\times\alpha)$ criterion (i.e., considered robust if actual alpha remains within 0.045-0.055 when nominal $\alpha = 0.05$).

The conditions we used to investigate robustness of Type I error rates were not exhaustive and were not fully crossed but were designed to cover a variety of circumstances researchers may face. The conditions designed to study Type I error rates varied the following: group sizes (balanced and unbalanced), variances (equal and unequal), and distributional shapes (normal, negatively skewed, and platykurtic). All means were set equal to 50 for these robustness conditions. Three distributional shapes were used with each of the group size and variance conditions. In all, we investigated nine conditions for each distributional shape for a total of 27 conditions.

Balanced group sizes in Phase 1 were all set to 40. Several patterns of sample sizes were used for unbalanced conditions: increasing sample sizes from group 1 to group 4 (i.e., 28, 36, 44, 52), two pairs of groups with equal sample sizes (i.e., 30, 30, 50, 50), and three groups with equal sample sizes (i.e., 36, 36, 36, 52). Similar patterns were used in Phase 2: (a) all samples with size of 40; (b) sample sizes of 20, 30, 40, 50, 60; (c) sample sizes of 30, 35, 40, 45, 50; (d) sample sizes of 30, 30, 40, 50, 50; (e) sample sizes of 36, 36, 36, 46, 46; (f) sample sizes of 35, 35, 35, 35, 60; and (g) sample sizes of 28, 43, 43, 43, 43.

Equal variance conditions in Phase 1 were all set to standard deviations of 10. Several patterns of standard deviations were used for the unequal variance conditions: decreasing standard deviations from group 1 to group 4 (i.e., 13, 11, 9, 7), increasing standard deviations from group 1 to group 4 (i.e., 7, 9, 11, 13), two pairs of groups with equal standard deviations (i.e., 12, 12, 8, 8), and three groups with equal standard deviations (i.e., 11, 11, 11, 7). Similar patterns were used in Phase 2: (a) all standard deviations set equal to 10; (b) standard deviations set to 14, 12, 10, 8, 6; (c) standard deviations set to 12, 11, 10, 9, 8; (d) standard deviations set to 12, 12, 10, 8, 8; (e) standard deviations set to 12, 12, 10, 10, 6; (f) standard deviations set to 13, 13, 8, 8, 8; and (g) standard deviations set to 11, 11, 11, 11, 6.

Statistical Power. To investigate statistical power, we varied the means but primarily maintained equal sample sizes ($N = 40$) and equal variances (i.e., $SD = 10$). In both the four-group Phase 1 and the five-group Phase 2, we used standardized mean difference effect sizes (Cohen's *d*) of 0.4 and 0.8 as the consistent mean differences among groups. In Phase 1, we reported results for the following mean vectors across the four groups: [50, 50, 50, 54], [50, 50, 50, 58], [50, 50, 54, 54], [50, 50, 54, 58], [50, 50, 58, 58], and [50, 54, 54, 58]. These patterns of means cover most, if not all, the possible patterns of 0.4 and 0.8 standardized mean differences among groups. It should be noted that in Phase 1 we also report results for one pattern of means where sample sizes were unequal, but variances were equal and also one condition where variances were unequal, but sample sizes were equal (these two conditions should be relatively robust in Type I error and therefore safe to examine for statistical power). In Phase 2, we used similar patterns of means: [50, 50, 50, 50, 54], [50, 50, 50, 50, 58], [50, 50, 50, 54, 54], [50, 50, 50, 54, 58], [50, 50, 50, 58, 58], [50, 50, 54, 54, 58], [50, 50, 54, 58, 58], and [50, 54, 54, 54, 58].

Data Generation and Analysis

In each condition, data were generated for each sample from the population conditions described above. For all distributions used in Phase 1, we investigated conditions with (a) equal sample sizes and equal variances, (b) equal sample sizes and unequal variances, (b) unequal sample sizes and equal variances, (b) unequal sample sizes and unequal variances (both the known liberal condition of larger groups having smaller variances and the conservative condition where smaller groups have smaller variances). For distributional shapes, normally distributed data were generated $N(0,1)$ and transformed. A population of one million cases was created with known means and standard deviations (described above) for skewed data using the $Beta(2,5)$ distribution and for platykurtic data using a uniform distribution. Phase 2 was essentially the same but did not include the conservative Type I error conditions (smaller samples with smaller variances) and only normal data were generated for Phase 2.

In each sample, we tested the assumption of normality using a variety of tests, including a conditional Bonferroni-adjusted Shapiro-Wilk, Pearson test of residuals, and several others (details in results). We tested the assumption of homoscedasticity (a.k.a., homogeneity of variances) using multiple versions of Levene's test, Breusch-Pagan, and Fligner (details in results). We calculated the omnibus ANOVA using Fisher's *F*, Welch's *F*, and Brown-Forsythe *F* (both unconditionally and conditionally based on homoscedasticity and normality assumption tests). For the one-way ANOVA analyses and the tests of assumptions, we used both Base `R` functions and several packages available for the tests (e.g., `car`, `lawstat`, `DescTools`, `jmv`, `lmtest`, `nortest`, `onewaytests`, `rosetta`).

We calculated SchefféMax and HollingsworthMax comparisons and tested their statistical significance using our own code. The Barcikowski comparisons were also calculated and tested using our own code and sorted according to explanatory power (i.e., proportion of between sums of squares explained). The four most informative Barcikowski comparisons were used for analyses. Finally, we also performed the relatively uncommon Brown-Forsythe adjustment to the Scheffé MCP *F*-test (Kirk, 2013) for all conditions. Like the omnibus ANOVAs, we also performed the MCPs both unconditionally and conditionally on a test of homoscedasticity.

Across all samples within a condition, rejections of tests were counted either for Type I errors or statistical power, as appropriate. We collected the rejection results from each sample for the omnibus tests, the assumptions tests, and the post hoc group comparisons. These results were then used to calculate Type I error rates or statistical power for each method in each condition. We used .05 as our nominal alpha for all tests, and as the familywise alpha for Bonferroni adjustments.

Code Verification. All code was tested for accuracy before final simulations were run. Where possible we used multiple Base `R` functions and `R` packages to cross-verify results. We verified the code for SchefféMax, HollingsworthMax, Barcikowski comparisons, and the Brown-Forsythe test ourselves because we could not find existing `R` functions or other programs that produced these results. Some Monte Carlo simulations were run multiple times with different seeds to verify that the results were not artifacts of poor seed choice. We ran simulations for conditions with known results (e.g., no violations of assumptions under true null hypotheses). We tested the `R` code with single samples and verified accurate counts across small numbers of multiple samples (e.g., 10 and 100).

Results

Congruence

The most important research question was about congruence of the maximum comparisons, especially for the new Barcikowski human-friendly comparisons (i.e., how often they agree with the omnibus ANOVA). Not surprisingly, however, we found empirically that both Type I error and statistical power are the same for SchefféMax and the omnibus Fisher *F*-test—also for HollingsworthMax with balanced group sizes. Table 2 and Figures 1-2 show that, as expected, 100% of the samples in all conditions resulted in the same rejection decision for the omnibus ANOVA *F*-statistic and the SchefféMax comparisons (99% for the HollingsworthMax comparisons). The maximum Barcikowski human-friendly comparison agreed with ANOVA in 98.8% of the samples across all four-group robustness conditions (over 96% of all five-group conditions)—also strongly congruent. Notably, the next three most explanatory Barcikowski comparisons also agreed with the omnibus *F*-test decision in at least 95.5% of the four-group robustness conditions (over 94% of all five-group conditions)—still strongly congruent.

Perhaps more interestingly, in the six four-group power conditions tabled, the BarcikowskiMax comparisons agreed with the omnibus *F*-test in at least 96.7% of the samples across power conditions (over 93% of all five-group conditions), and the fourth most explanatory Barcikowski comparisons still agreed at a rate of at least 82.3% across the four-group power conditions (over 77% of all five-group conditions). Generally, in both four and five groups, the top four Barcikowski comparisons had more congruence with higher-power conditions (e.g., when the minimum effect size was Cohen's *d* = 0.8) and lower congruence where effect sizes were smaller.

Robustness

Tables 3 and 4 show the Type I error rates of the SchefféMax, HollingsworthMax, and BarcikowskiMax comparisons—using the conditional tests. That is, because we determined that the conditional tests for all three methods performed better and more consistently in terms of controlling Type I error (see Figures 3-6), we report tabulated results for only the conditional tests.

The unconditional Scheffé (i.e., where the pooled Scheffé *F* is always used regardless of homoscedasticity) and the unconditional Brown-Forsythe (i.e., where the Brown-Forsythe test is always used regardless of homoscedasticity) approaches followed essentially the same relative patterns both within and across methods as the conditional tests for almost all conditions, so those results were essentially redundant to Table 3 and Table 4, and therefore not shown.

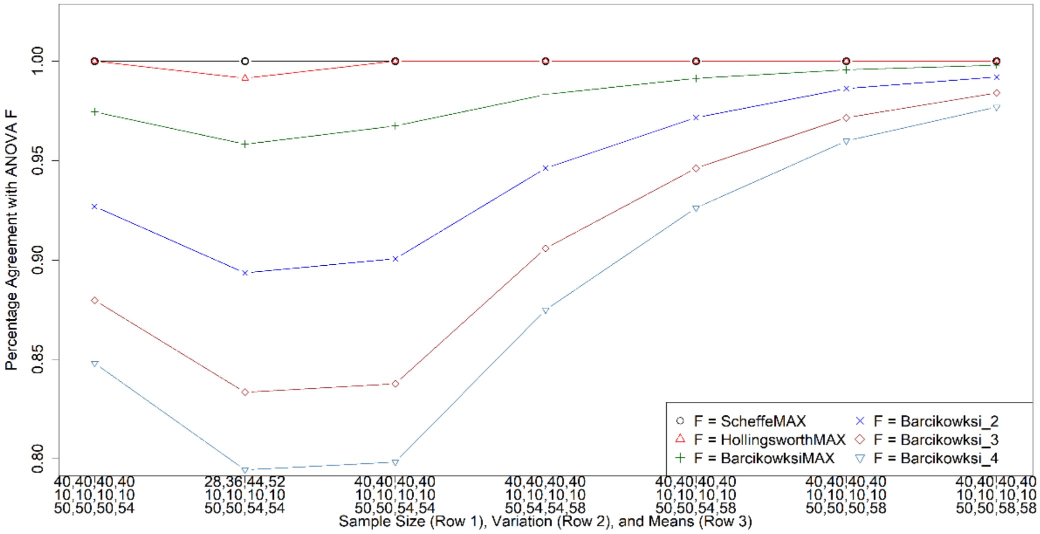


Figure 1. Percentage of agreement (congruence) between comparisons (SchefféMax, HollingsworthMax, and four most explanatory Barcikowski Human-Friendly comparisons) and the Omnibus *F*-test for statistical power conditions (unequal means) for four groups and only equal variances and equal sample sizes.

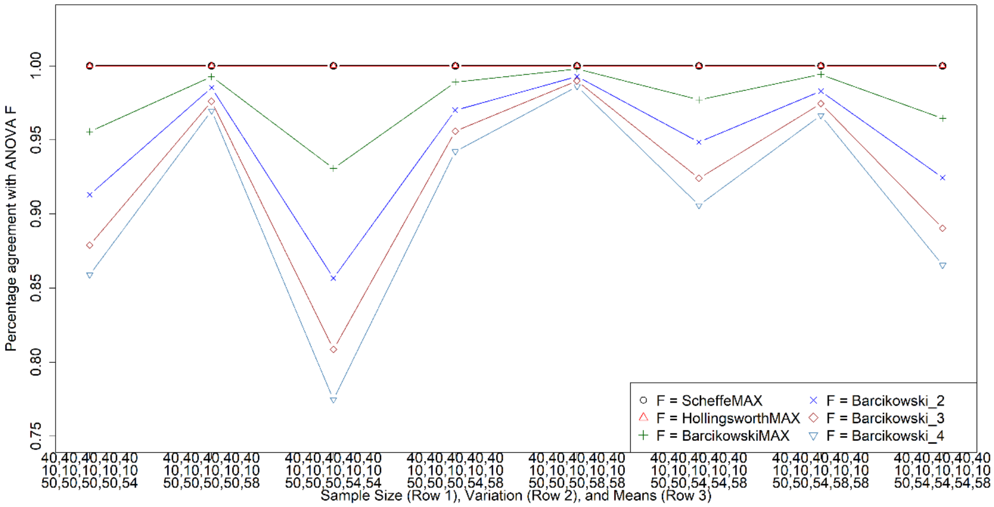


Figure 2. Percentage of agreement (congruence) between comparisons (SchefféMax, HollingsworthMax, and four most explanatory Barcikowski Human-Friendly comparisons) and the Omnibus *F*-test for statistical power conditions (unequal means) for five groups and only equal variances and equal sample sizes

Table 2. Percentage of agreement (congruence) between unconditional comparisons (SchefféMax, HollingsworthMax, and the four most explanatory Barcikowski Human-Friendly comparisons) and the Phase 1 four-group omnibus *F*-test under both Type I error robustness and Power conditions

Type I Error Robustness Conditions			**Percentage Agreement with Omnibus *F*-test rejection decision**					
			Scheffé	**Hollings-worth**	**1st most explanatory**	**2nd most explanatory**	**3rd most explanatory**	**4th most explanatory**
N	***SD***	**Shape**	**maximum**	**maximum**	**Barcikowski**	**Barcikowski**	**Barcikowski**	**Barcikowski**
40,40,40,40	10,10,10,10	Normal	100.00%	100.00%	99.29%	98.13%	97.28%	96.76%
28,36,44,52	10,10,10,10	Normal	100.00%	99.70%	99.01%	97.94%	97.13%	96.70%
40,40,40,40	13,11,9,7	Normal	100.00%	100.00%	99.36%	98.32%	97.29%	96.74%
28,36,44,52	13,11,9,7	Normal	100.00%	99.64%	98.82%	97.40%	96.43%	95.68%
28,36,44,52	7,9,11,13	Normal	100.00%	99.76%	99.23%	98.35%	97.73%	97.38%
40,40,40,40	10,10,10,10	Skewed	100.00%	100.00%	99.25%	98.03%	97.20%	96.76%
28,36,44,52	10,10,10,10	Skewed	100.00%	99.61%	99.03%	97.95%	97.23%	96.84%
40,40,40,40	13,11,9,7	Skewed	100.00%	100.00%	99.34%	98.03%	97.08%	96.53%
28,36,44,52	13,11,9,7	Skewed	100.00%	99.69%	98.84%	97.53%	96.52%	95.69%
28,36,44,52	7,9,11,13	Skewed	100.00%	99.77%	99.28%	98.52%	97.93%	97.57%
40,40,40,40	10,10,10,10	Kurtotic	100.00%	100.00%	99.16%	98.02%	97.15%	96.72%
28,36,44,52	10,10,10,10	Kurtotic	100.00%	99.64%	99.01%	98.02%	97.16%	96.68%
40,40,40,40	13,11,9,7	Kurtotic	100.00%	100.00%	99.22%	98.01%	97.08%	96.52%
28,36,44,52	13,11,9,7	Kurtotic	100.00%	99.66%	98.80%	97.36%	96.24%	95.54%
28,36,44,52	7,9,11,13	Kurtotic	100.00%	99.74%	99.23%	98.39%	97.73%	97.30%
Statistical Power Conditions	**Mean Structures**		**Statistical Power (All Normally Distributed)**					
	50, 50, 50, 54 [1]		100.00%	100.00%	97.52%	93.31%	89.11%	86.31%
	50, 50, 50, 58 [1]		100.00%	100.00%	98.98%	96.62%	93.66%	91.25%
	50, 50, 54, 54 [1]		100.00%	100.00%	96.95%	91.03%	85.59%	82.32%
	50, 50, 54, 54 [2]		100.00%	99.11%	96.17%	90.56%	85.33%	82.35%
	50, 50, 54, 54 [3]		100.00%	100.00%	96.94%	91.18%	86.32%	83.13%
	50, 50, 54, 58 [1]		100.00%	100.00%	98.15%	94.37%	90.03%	86.89%
	50, 50, 58, 58 [1]		100.00%	100.00%	99.24%	97.04%	94.44%	92.50%
	50, 54, 54, 58 [1]		100.00%	100.00%	97.53%	92.35%	87.46%	83.61%

[1] *Group sample sizes were equal (40, 40, 40, 40) and variances were equal (10, 10, 10, 10)*

[2] *Group sample sizes were unequal (28, 36, 44, 52) but variances were equal (10, 10, 10, 10)*

[3] *Group sample sizes were equal (40, 40, 40, 40) but variances were unequal (13, 11, 9, 7)*

Table 3. Type I error rates for conditional Scheffé and Brown-Forsythe Adjusted Scheffé significance tests for maximum comparisons using α = 0.05 for all four-group tests (conditional tests performed based on Brown-Forsythe critical values if Levene's test is significant, Scheffé *F p*-values if not)

Type I Error Robustness Conditions			**Tests**			
			Omnibus	**Conditional Scheffé**	**Conditional Hollingsworth**	**Conditional Barcikowski**
N	***SD***	**Shape**	**Fisher *F***	**Maximum**	**Maximum**	**Maximum**
10,10,10,10	40,40,40,40	Normal	0.05034	0.05068	0.05068	0.04241
10,10,10,10	28,36,44,52	Normal	0.05043	0.05072	0.04736	0.04110
10,10,10,10	40,40,40,40	Skewed	0.04954	0.04951	0.04951	0.04233
10,10,10,10	28,36,44,52	Skewed	0.04979	0.05032	0.04698	0.04075
10,10,10,10	40,40,40,40	Kurtotic	0.05016	0.05081	0.05081	0.04365
10,10,10,10	28,36,44,52	Kurtotic	0.05002	0.05056	0.04695	0.04004
13,11,9,7	40,40,40,40	Normal	0.05650	0.04185	0.04185	0.03725
13,11,9,7	28,36,44,52	Normal	0.08372	0.04637	0.03809	0.03439
7,9,11,13	28,36,44,52	Normal	0.03761	0.03847	0.04473	0.03897
13,11,9,7	30,30,50,50	Normal	0.08175	0.04479	0.03696	0.03408
13,11,9,7	36,36,36,52	Normal	0.06928	0.04392	0.03988	0.03600
12,12,8,8	30,30,50,50	Normal	0.08237	0.05207	0.04371	0.03979
11,11,11,7	36,36,36,52	Normal	0.06655	0.05046	0.04634	0.04076
13,11,9,7	40,40,40,40	Skewed	0.05718	0.04382	0.04382	0.03961
13,11,9,7	28,36,44,52	Skewed	0.08248	0.04834	0.03994	0.03752
7,9,11,13	28,36,44,52	Skewed	0.03786	0.03936	0.04650	0.04027
13,11,9,7	30,30,50,50	Skewed	0.08218	0.04705	0.03943	0.03626
13,11,9,7	36,36,36,52	Skewed	0.06812	0.04513	0.04129	0.03723
12,12,8,8	30,30,50,50	Skewed	0.08191	0.05314	0.04505	0.04071
11,11,11,7	36,36,36,52	Skewed	0.06595	0.04933	0.04551	0.04055
13,11,9,7	40,40,40,40	Kurtotic	0.05649	0.04252	0.04252	0.03764
13,11,9,7	28,36,44,52	Kurtotic	0.08273	0.04610	0.03679	0.03415
7,9,11,13	28,36,44,52	Kurtotic	0.03704	0.03957	0.04772	0.04130
13,11,9,7	30,30,50,50	Kurtotic	0.08410	0.04720	0.03868	0.03535
13,11,9,7	36,36,36,52	Kurtotic	0.07076	0.04485	0.04049	0.03626
12,12,8,8	30,30,50,50	Kurtotic	0.08236	0.05111	0.04261	0.03873
11,11,11,7	36,36,36,52	Kurtotic	0.06681	0.05083	0.04649	0.04156

In Figures 3-5 and other results not presented, SchefféMax and HollingsworthMax comparisons have liberal Type I error rates (i.e., over .055 based on Bradley, 1978) using the unconditional Scheffé MCP *F* test, especially in the inverse variance-size conditions (i.e., larger groups with smaller variances). Similarly, both were conservative (i.e., below .045) in the direct variance-size conditions (i.e., smaller groups with smaller variances). However, with the conditional test, Type I error rates were more consistently robust across all conditions (typically within the .045-.055 range set as a stringent robustness criterion based on Bradley, 1978), including where there were violations of homoscedasticity. Barcikowski comparisons appeared conservative generally, but less so when using the conditional test (see Figure 6).

Figures 3-5 show that the Brown-Forsythe adjusted Scheffé MCP controls Type I error when the homoscedasticity assumption is violated. These figures show that these maximum comparison tests appear to benefit from using a conditional test (i.e., pooled Scheffé *F* used when Levene's test is nonsignificant, but Brown-Forsythe used when Levene is statistically significant). Note that in the most extreme inverse variance-sample size conditions, the Brown-Forsythe is almost always used because the preliminary homoscedasticity test is almost always significant.

Statistical Power

As before, we report only the results for the conditional tests because they were the most robust for all the methods (statistical power really only matters for statistical tests that maintain control over Type I error).

Table 4. Type I error rates for conditional Scheffé and Brown-Forsythe Adjusted Scheffé significance tests for maximum comparisons using α = 0.05 for all five-group tests (conditional tests performed based on Brown-Forsythe critical values if Levene's test is significant, Scheffé *F p*-values if not)

N	***SD***	**Omnibus Fisher *F***	**Conditional Scheffé Maximum**	**Conditional Hollingsworth Maximum**	**Conditional Barcikowski Maximum**
40,40,40,40,40	10,10,10,10,10	0.0486	0.0490	0.0490	0.0359
40,40,40,40,40	14,12,10,08,06	0.0622	0.0301	0.0301	0.0236
40,40,40,40,40	12,11,10,09,08	0.0532	0.0501	0.0501	0.0375
40,40,40,40,40	12,12,10,10,06	0.0535	0.0368	0.0368	0.0280
40,40,40,40,40	13,13,08,08,08	0.0609	0.0357	0.0357	0.0303
40,40,40,40,40	11,11,11,11,06	0.0557	0.0416	0.0416	0.0330
20,30,40,50,60	10,10,10,10,10	0.0464	0.0464	0.0350	0.0285
20,30,40,50,60	14,12,10,08,06	0.1249	0.0323	0.0190	0.0166
20,30,40,50,60	12,11,10,09,08	0.0840	0.0657	0.0479	0.0407
20,30,40,50,60	12,12,10,10,06	0.0948	0.0423	0.0284	0.0245
20,30,40,50,60	13,13,08,08,08	0.1076	0.0427	0.0282	0.0259
20,30,40,50,60	11,11,11,11,06	0.0778	0.0426	0.0305	0.0246
30,35,40,45,50	10,10,10,10,10	0.0499	0.0505	0.0478	0.0351
30,35,40,45,50	14,12,10,08,06	0.0856	0.0320	0.0261	0.0222
30,35,40,45,50	12,11,10,09,08	0.0672	0.0562	0.0492	0.0394
30,35,40,45,50	12,12,10,10,06	0.0746	0.0383	0.0322	0.0273
30,35,40,45,50	13,13,08,08,08	0.0867	0.0391	0.0311	0.0277
30,35,40,45,50	11,11,11,11,06	0.0636	0.0418	0.0371	0.0288
30,30,40,50,50	10,10,10,10,10	0.0488	0.0492	0.0453	0.0341
30,30,40,50,50	14,12,10,08,06	0.0942	0.0337	0.0242	0.0210
30,30,40,50,50	12,11,10,09,08	0.0643	0.0530	0.0453	0.0346
30,30,40,50,50	12,12,10,10,06	0.0748	0.0403	0.0315	0.0274
30,30,40,50,50	13,13,08,08,08	0.0843	0.0380	0.0290	0.0255
30,30,40,50,50	11,11,11,11,06	0.0625	0.0423	0.0357	0.0280
36,36,36,46,46	10,10,10,10,10	0.0526	0.0531	0.0519	0.0362
36,36,36,46,46	14,12,10,08,06	0.0735	0.0293	0.0259	0.0217
36,36,36,46,46	12,11,10,09,08	0.0616	0.0528	0.0503	0.0381
36,36,36,46,46	12,12,10,10,06	0.0637	0.0367	0.0339	0.0255
36,36,36,46,46	13,13,08,08,08	0.0722	0.0370	0.0325	0.0271
36,36,36,46,46	11,11,11,11,06	0.0619	0.0419	0.0382	0.0308
35,35,35,35,60	10,10,10,10,10	0.0459	0.0459	0.0437	0.0310
35,35,35,35,60	14,12,10,08,06	0.0904	0.0344	0.0291	0.0241
35,35,35,35,60	12,11,10,09,08	0.0695	0.0559	0.0519	0.0376
35,35,35,35,60	12,12,10,10,06	0.0839	0.0429	0.0356	0.0289
35,35,35,35,60	13,13,08,08,08	0.0716	0.0353	0.0309	0.0256
35,35,35,35,60	11,11,11,11,06	0.0774	0.0390	0.0337	0.0273
28,43,43,43,43	10,10,10,10,10	0.0498	0.0502	0.0471	0.0353
28,43,43,43,43	14,12,10,08,06	0.0777	0.0289	0.0239	0.0198
28,43,43,43,43	12,11,10,09,08	0.0648	0.0562	0.0494	0.0397
28,43,43,43,43	12,12,10,10,06	0.0635	0.0409	0.0348	0.0280
28,43,43,43,43	13,13,08,08,08	0.0675	0.0346	0.0287	0.0245
28,43,43,43,43	11,11,11,11,06	0.0553	0.0388	0.0354	0.0296

Table 5 shows that, for both four groups and five groups, SchefféMax and HollingsworthMax have essentially the same power (exactly the same when sample sizes are equal). BarcikowskiMax is generally less powerful than the other two methods, but the power rates do not drop terribly dramatically. There is certainly a relationship between BarcikowskiMax having lower power and its being more conservative in terms of Type I error rates. Power did not drop dramatically for the next three most explanatory Barcikowski comparisons.

Supplemental Analyses

The primary purpose of our paper was to investigate the performance of Barcikowski human-friendly comparisons relative to the Scheffé maximum comparison. However, in testing the assumptions for the numerous conditions in Phase 1 (four groups), we identified results we have not seen reported in the literature regarding the performance of tests of normality and tests of homogeneity of variances in ANOVA. Further research must be undertaken to study additional scenarios and try to confirm and to expand the results for both assumptions, but we believe it is important to share these results. We only report the most interesting conditions for each analysis here.

Homoscedasticity. Most importantly, although non-normality is known not to have a substantial impact on the robustness of one-way ANOVA, non-normality does impact tests of homoscedasticity. Figure 7 shows that these tests can be impacted by skewed population data, in particular. Two adaptations to Levene from the `R lawstat` package provided the most consistently robust Type I error rates in the conditions we tested, using Bradley's (1978) stringent criterion. The authors of the `lawstat R` package (see Hui et al., 2008) credit Hines and Hines (2000) for the "Zero Removal" method and they cite Noguchi and Gel (2010) for the "Zero Correction" method (both use the median in calculations). In our study, the Zero Correction method performed just a little better than the Zero Removal method—and both performed just a little better overall than the Breusch-Pagan method (using defaults with the `lmtest R` package function `bptest`), which tended slightly toward inflated Type I error with skewed data, but which was most robust to kurtotic data. Importantly, the Levene test based on the mean showed inflated Type I error when the data were skewed and the Fligner and Levene-Browne-Forsythe (median) tests were generally conservative. It should be noted that, for the analyses in this study we used the Levene-Browne-Forsythe test, which had been recommended as a good choice by Gaonkar and Beasley (2023).

Table 5. Statistical power rates for conditional Scheffé and Brown-Forsythe Adjusted Scheffé significance tests for maximum comparisons (SchefféMax, HollingsworthMax, and BarcikowskiMax) using α = 0.05 for all tests (conditional tests performed based on Brown-Forsythe values if Levene's test is significant, Scheffé *F* *p*-values if not)

Means	**Conditional Scheffé Maximum**	**Conditional Hollingsworth Maximum**	**Conditional Barcikowski Maximum**
Four Groups			
50, 50, 50, 54 [1]	0.3174	0.3174	0.2927
50, 50, 50, 58 [1]	0.8913	0.8913	0.8811
50, 50, 54, 54 [1]	0.4096	0.4096	0.3802
50, 50, 54, 54 [2]	0.3927	0.3840	0.3544
50, 50, 54, 54 [3]	0.3713	0.3713	0.3444
50, 50, 54, 58 [1]	0.8628	0.8628	0.8441
50, 50, 58, 58 [1]	0.9632	0.9632	0.9555
50, 54, 54, 58 [1]	0.7268	0.7268	0.7018
Five Groups			
50, 50, 50, 50, 54 [1]	0.4024	0.4024	0.3571
50, 50, 50, 50, 58 [1]	0.9613	0.9613	0.9543
50, 50, 50, 54, 54 [1]	0.5732	0.5732	0.5041
50, 50, 50, 54, 58 [1]	0.9619	0.9619	0.9502
50, 50, 50, 58, 58 [1]	0.9969	0.9969	0.9949
50, 50, 54, 54, 58 [1]	0.9367	0.9367	0.9134
50, 50, 54, 58, 58 [1]	0.9900	0.9900	0.9841
50, 54, 54, 54, 58 [1]	0.8234	0.8234	0.7886

[1] *Group sample sizes were equal, and variances were equal*
[2] *Group sample sizes were unequal, but variances were equal*
[3] *Group sample sizes were equal, but variances were unequal*

Normality. For normality, we found that most tests of normality had inflated Type I error rates when both variances and samples sizes were unequal. There was a clear advantage for the Bonferroni-adjusted conditional Shapiro-Wilk tests tested in the study. That is, whenever any group has a statistically significant Shapiro-Wilk test using the Bonferroni alpha-adjustment for the number of groups (e.g., familywise alpha divided by number of groups, where adjusted α = 0.0125 with four groups), the omnibus null hypothesis of the normality assumption would be rejected. The Kolmogorov-Smirnov-Lilliefors (KSL) test also performed well using the same Bonferroni-adjusted conditional approach but was slightly more conservative than Shapiro-Wilk in several conditions. Figure 8 shows that the most robust test of the normality of the residuals was the Pearson test, which might be considered minimally acceptable based on its Type I error rates. The best of the rest, which were not considered acceptable under most conditions, was the KSL test of the normality of residuals (and note the Type I error inflation of Shapiro-Wilk with residuals). We could not find literature that suggested a Bonferroni-type adjustment for the conditional approach to test normality across groups.

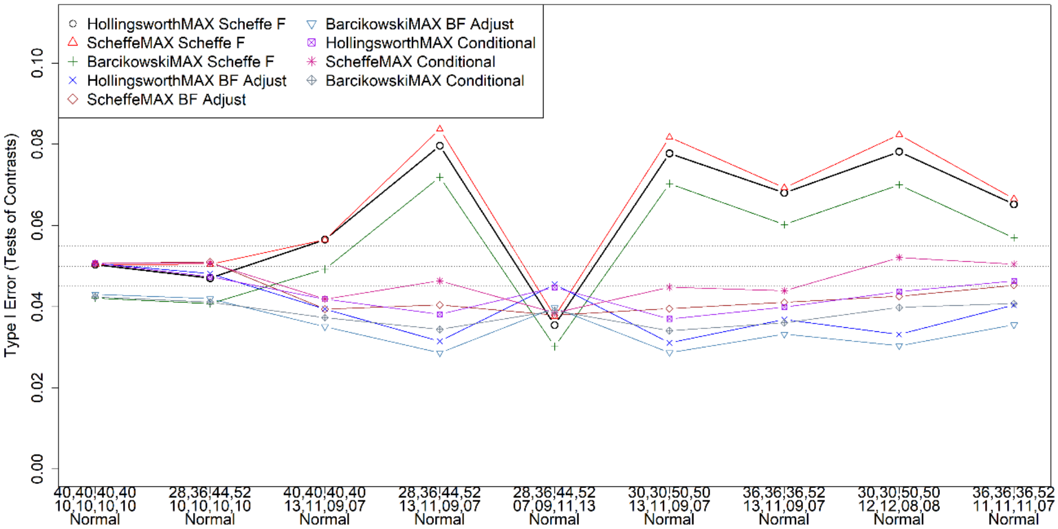


Figure 3. Four-group Type I error rates for Scheffé and Brown-Forsythe Adjusted Scheffé significance tests for maximum comparisons both unconditionally and conditional on Levene's test of equality of variances (using α = 0.05 for all tests) for normal data

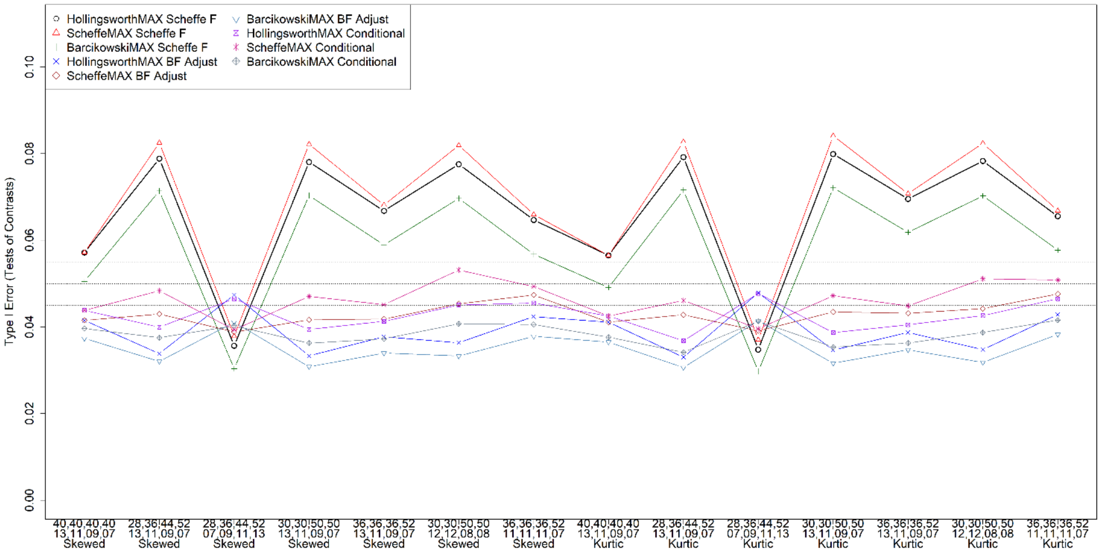


Figure 4. Four-group Type I error rates for Scheffé and Brown-Forsythe Adjusted Scheffé significance tests for maximum comparisons both unconditionally and conditional on Levene's test of equality of variances (using α = 0.05 for all tests) for non-normally distributed data with unequal variances

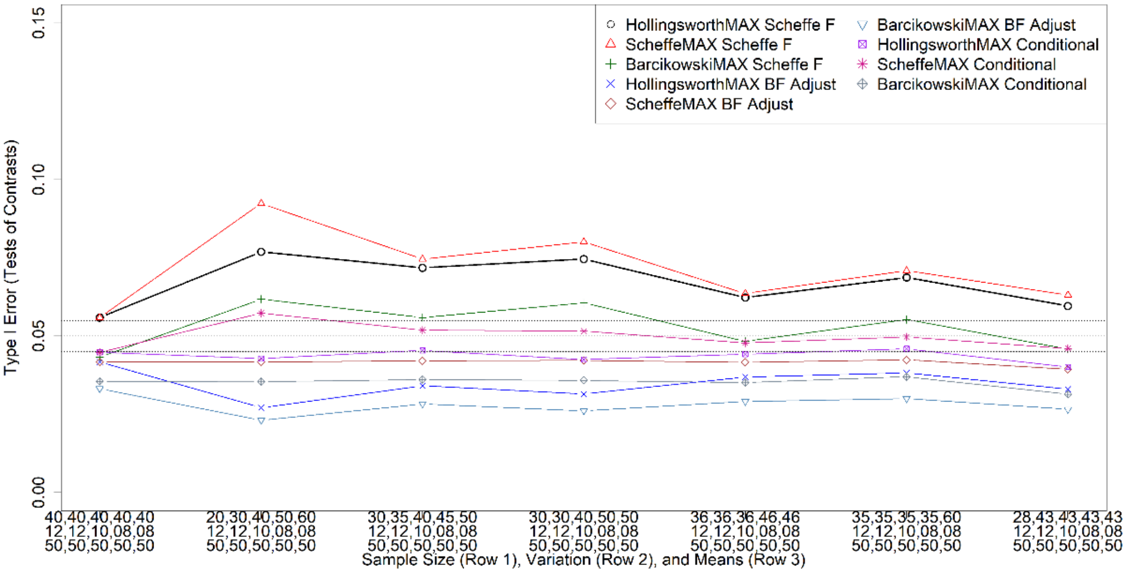


Figure 5. Five-group Type I error rates for Scheffé and Brown-Forsythe Adjusted Scheffé significance tests for maximum comparisons both unconditionally and conditional on Levene's test of equality of variances (using α = 0.05 for all tests) for normally distributed data with unequal sample sizes and unequal variances

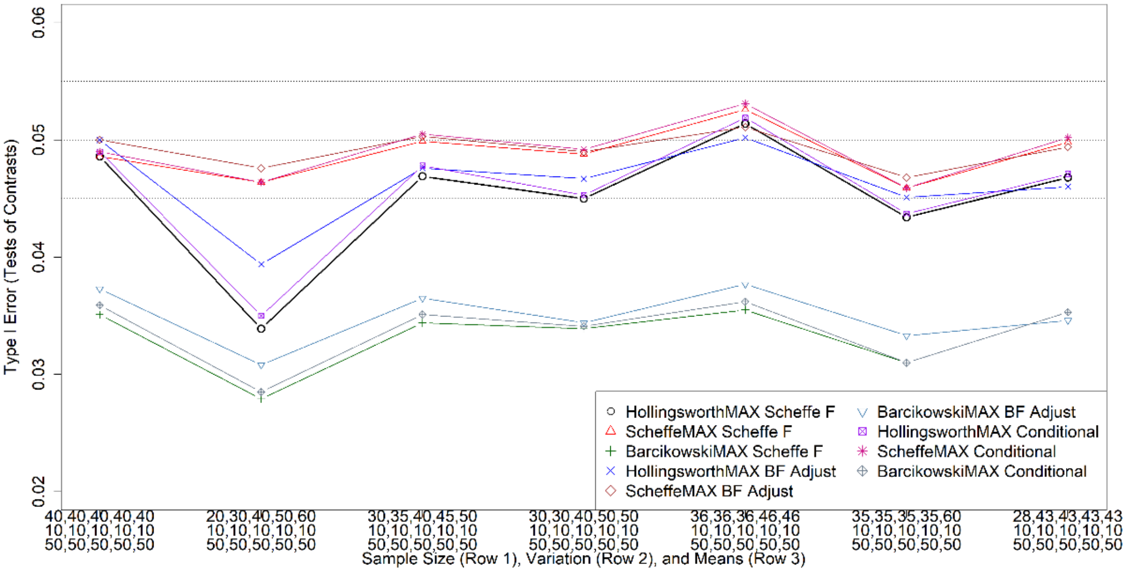


Figure 6. Five-group Type I error rates for Scheffé and Brown-Forsythe Adjusted Scheffé significance tests for maximum comparisons both unconditionally and conditional on Levene's test of equality of variances (using α = 0.05 for all tests) for normally distributed data with unequal sample sizes and equal variances

Conditional Testing. Finally, like others have found for *t*-tests (Delacre et al., 2017; Hayes & Cai, 2007; Zimmerman, 2004), Figure 9 shows that the best approach for controlling Type I error inflation in a four-group omnibus ANOVA with unequal variances and unequal sample sizes may be to use the unconditional Welch *F*-test with no preliminary tests of assumptions (rather than the conditional approach based on using Welch *F* if homoscedasticity is violated, Fisher *F* if not). Using the unconditional Brown-Forsythe *F*-test was just a little less robust than Welch, but still better than the conditional approach for the omnibus test (unlike what we found for the MCPs).

Discussion

We have presented a method of Scheffé-style, complex, non-pairwise multiple comparisons in a new way, called Barcikowski "human-friendly" comparisons, that we believe can be useful to many researchers. We have created an `R Shiny` online app to calculate SchefféMax, HollingsworthMax, and the most explanatory Barcikowski human-friendly comparisons. The app tests these comparisons statistically with both Scheffé's MCP *F*-test and Brown-Forsythe's adjustment to Scheffé for unequal variances. Practically speaking, for significance testing of Scheffé comparisons, there are programs already available (e.g., `ScheffeTest` in the `DescTools R` package) that will calculate the significance of specific Scheffé comparisons users send as input. The key from a post hoc perspective, however, is identifying, through a method like ours, which of the infinite possible comparisons to include in the input comparison matrix.

Some have argued that the best way to use complex comparisons may be a priori—identifying contrasts of interest based on theory in the research questions (e.g., Maxwell et al., 2018), perhaps without even performing the omnibus test or adjusting alpha. However, our experience suggests that studies often result in unexpected relationships and differences that can be useful to theory advancement. Having a practical method, provided here, by which to test the most explanatory complex comparisons can assist with this process. For example, the first analysis in Figure 10 results in the comparison 1 between group 1 and the average of groups 2 and 3 (i.e., with coefficients [1 −½ −½]) as the most explanatory comparison; however, the comparison between groups 1 and 2 (i.e., with coefficients [1 −1 0]) is also statistically significant (and likely would have been found by any pairwise MCP just as it was found significant by Scheffé). Rather than just being happy that groups 1 and 2 differed based on the pairwise comparison, the researcher might find value in learning that the [1 −½ −½] complex comparison was more explanatory than [1 −1 0] (based on the values of SSQ, which is the Sum of Squares Explained by the comparison).

Barcikowski Human-Friendly comparisons. The most explanatory Barcikowski comparisons showed a high level of congruence with the omnibus *F*-test, albeit lower than SchefféMax and HollingsworthMax. Results also showed that the BarcikowskiMax comparison agreed (i.e., was congruent) with ANOVA in over 98% of the samples across all robustness conditions. Notably, the next three most explanatory Barcikowski comparisons also agreed with the omnibus *F*-test decision in at least 95% of the robustness conditions in both four groups and five groups. Results suggest, however, that this agreement percentage may decline as the number of groups increases, and therefore additional studies with more groups should be undertaken. Although the Barcikowski comparisons showed less congruence in the power conditions, the BarcikowskiMax maintained strong congruence in most conditions (over 94% in the conditions studied). As the Barcikowski comparisons become less explanatory (i.e., second through fourth most explanatory), their congruence levels became increasingly lower—but perhaps still an acceptable level for the researcher.

Compared to SchefféMax and HollingsworthMax, the Barcikowski comparisons generally appear to be more conservative regarding Type I error, which also resulted in lower power than the other two methods. However, we found that the maximum Barcikowski human-friendly comparisons do not suffer much power loss compared to SchefféMax. Type I error rates for Barcikowski comparisons follow the same patterns regarding assumption conditions as SchefféMax.

Brown-Forsythe Comparisons. The General Linear Model relies on assumptions being met to provide accurate probabilities for its statistical tests. Although assumption violations can sometimes be ignored as having trivial impact on Type I errors, their violation sometimes has substantial impact. Like other General Linear Model statistical methods, researchers are encouraged to test assumptions, especially normality and homoscedasticity, as part of the analytical process for Scheffé-type comparisons of any kind.

Based on our results, we conclude that always using robust tests (e.g., Welch) unconditionally may also be optimal for the omnibus test in one-way ANOVA. However, with the Scheffé-type maximum comparisons (and for those who prefer to verify that assumptions are met), we recommend using the Zero Correction or Zero Removal methods available in the `R lawstat` package to test homogeneity of variances. Gaonkar and Beasley (2023) had similarly concluded that the original correction factor adjustment on which Zero Correction and Zero Removal was among the better choices for the test of homoscedasticity in ANOVA.

We could not find any existing program that implements the Brown-Forsythe adaptation to the Scheffé MCP for unequal variances and we also provide those analyses as part of the Shiny app. This is noteworthy because Scheffé MCP requires the assumption of homoscedasticity. If researchers use our recommended process, it will be important to know the robustness and power results for the Brown-Forsythe adaptation. We found that the Brown-Forsythe adjustment for Scheffé does indeed help control Type I error rates when variances and sample sizes differ. Because the Scheffé MCP has not been used much in applied research, the consideration of an unequal-variances adaptation had not been a major concern among scholars. But with our new approach it will become important.

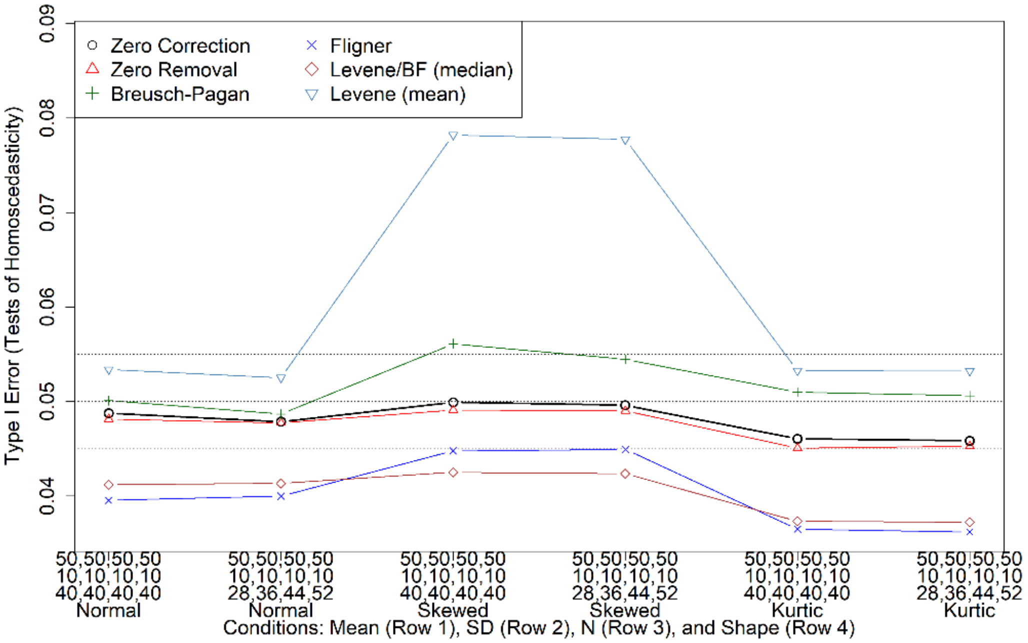


Figure 7. Type I error rates for preliminary tests of homogeneity of variances in four groups

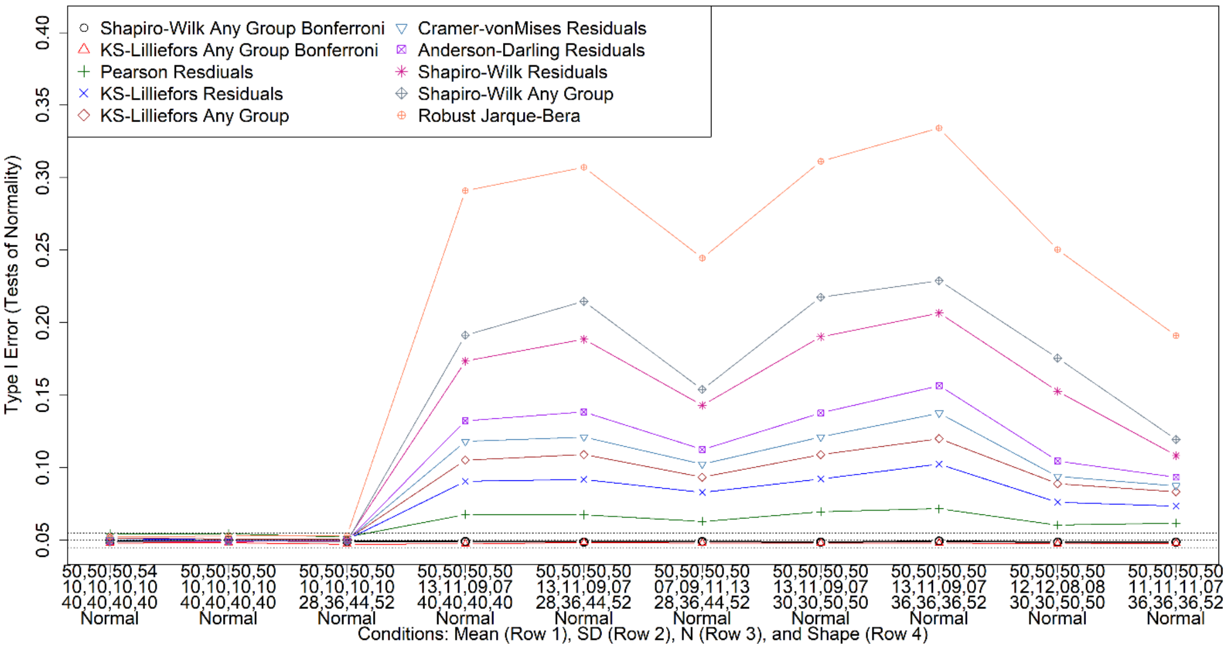


Figure 8. Type I error rates for preliminary tests for the assumption of normality in four groups

Further, we found the conditional-on-Levene approach to significance testing of these comparisons to be preferable for robustness purposes. Some scholars (e.g., Delacre et al., 2017; Hayes & Cai, 2007; Zimmerman, 2004) have recommended that the Welch's *t*-test (and by generalization the robust Welch's *F*-test) should be used without the preliminary test of homoscedasticity—indeed, our supplemental results confirm this. However, we found that there may be benefit to performing this preliminary assumption test before running the Scheffé-type maximum MCP procedures presented here. Indeed, it is still widespread practice to decide whether to use robust tests based on the significance of Levene's test and therefore this conclusion will not impact many.

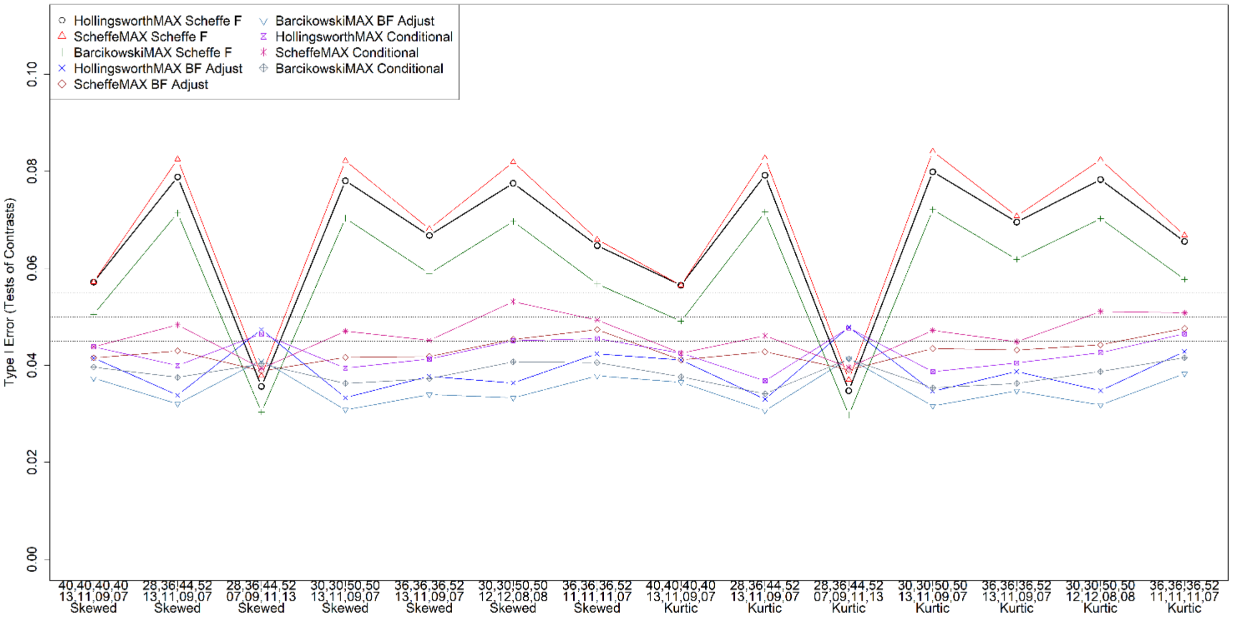
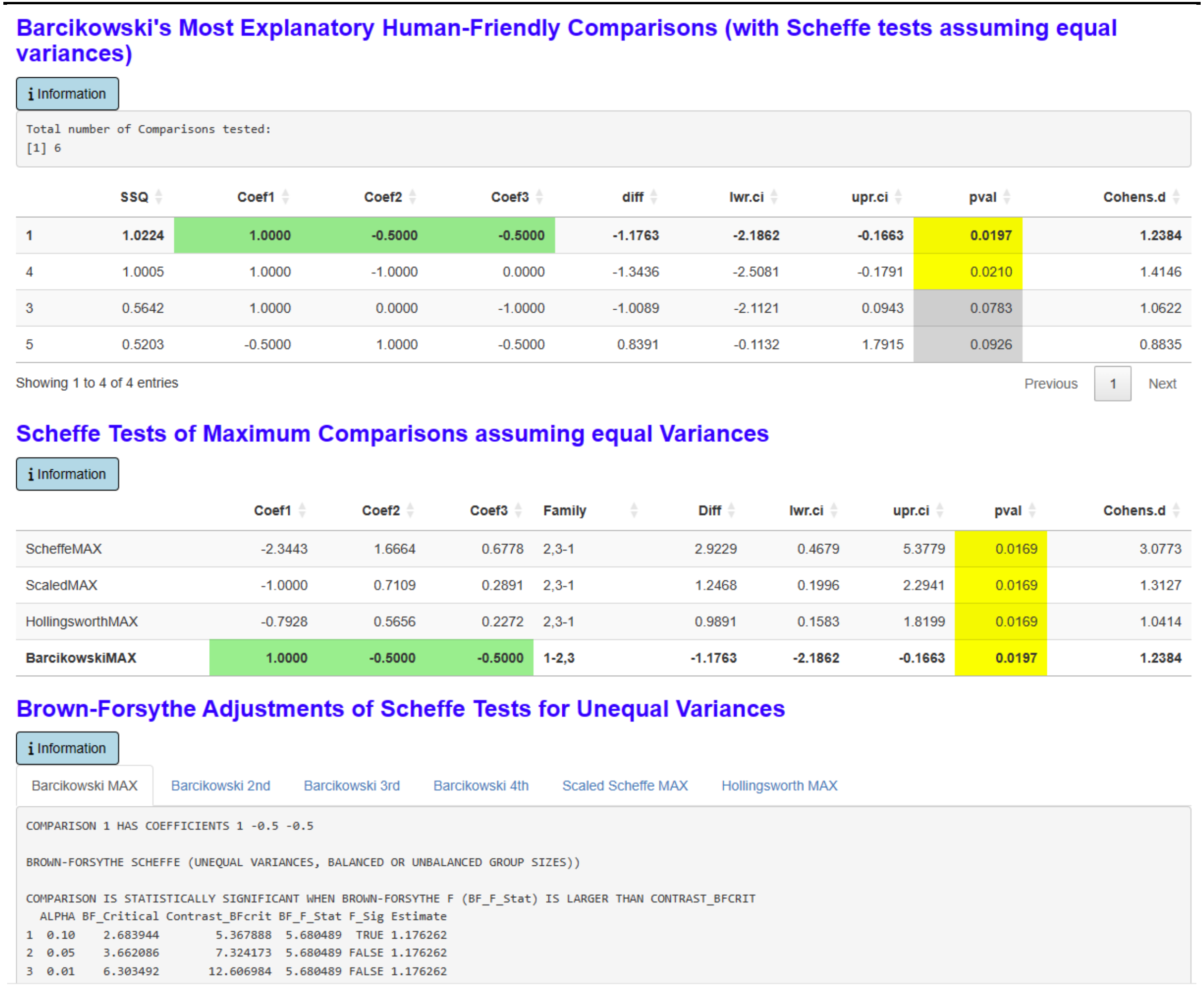


Figure 9. Type I error rates of conditional and unconditional omnibus tests of ANOVA means comparisons in four groups



Barcikowski's Most Explanatory Human-Friendly Comparisons (with Scheffe tests assuming equal variances)

```
Total number of Comparisons tested:
[1] 6
```

	SSQ	Coef1	Coef2	Coef3	diff	lwr.ci	upr.ci	pval	Cohens.d
1	**1.0224**	**1.0000**	**-0.5000**	**-0.5000**	**-1.1763**	**-2.1862**	**-0.1663**	**0.0197**	**1.2384**
4	1.0005	1.0000	-1.0000	0.0000	-1.3436	-2.5081	-0.1791	0.0210	1.4146
3	0.5642	1.0000	0.0000	-1.0000	-1.0089	-2.1121	0.0943	0.0783	1.0622
5	0.5203	-0.5000	1.0000	-0.5000	0.8391	-0.1132	1.7915	0.0926	0.8835

Showing 1 to 4 of 4 entries

Scheffe Tests of Maximum Comparisons assuming equal Variances

	Coef1	Coef2	Coef3	Family	Diff	lwr.ci	upr.ci	pval	Cohens.d
ScheffeMAX	-2.3443	1.6664	0.6778	2,3-1	2.9229	0.4679	5.3779	0.0169	3.0773
ScaledMAX	-1.0000	0.7109	0.2891	2,3-1	1.2468	0.1996	2.2941	0.0169	1.3127
HollingsworthMAX	-0.7928	0.5656	0.2272	2,3-1	0.9891	0.1583	1.8199	0.0169	1.0414
BarcikowskiMAX	**1.0000**	**-0.5000**	**-0.5000**	**1-2,3**	**-1.1763**	**-2.1862**	**-0.1663**	**0.0197**	**1.2384**

Brown-Forsythe Adjustments of Scheffe Tests for Unequal Variances

```
COMPARISON 1 HAS COEFFICIENTS 1 -0.5 -0.5

BROWN-FORSYTHE SCHEFFE (UNEQUAL VARIANCES, BALANCED OR UNBALANCED GROUP SIZES))

COMPARISON IS STATISTICALLY SIGNIFICANT WHEN BROWN-FORSYTHE F (BF_F_Stat) IS LARGER THAN CONTRAST_BFCRIT
  ALPHA BF_Critical Contrast_BFcrit BF_F_Stat F_Sig Estimate
1  0.10    2.683944        5.367888  5.680489  TRUE 1.176262
2  0.05    3.662086        7.324173  5.680489 FALSE 1.176262
3  0.01    6.303492       12.606984  5.680489 FALSE 1.176262
```

Figure 10. A subset of the most critical output from an example dataset used with the `R Shiny` app (https://72x6cr-gordon-brooks.shinyapps.io/Human_Friendly_Contrasts/)

Conclusions

We believe that every researcher who performs ANOVA should run the SchefféMax comparison. Like running both ANOVA and independent t tests for two group mean comparisons, there is no Type I error inflation concern for obtaining both the ANOVA and the SchefféMax *p*-values—they are equivalent tests with equal *p*-values. There is valuable exploratory information in the SchefféMax comparison that should be reviewed in the same way researchers examine the coefficient weights in multiple linear regression and discriminant analysis or the loadings in factor analysis. We believe that researchers are missing valuable exploratory and descriptive information by not examining this maximum comparison. Indeed, because of the perfect congruence, the maximum Scheffé comparison should have higher power conditional on a statistically significant ANOVA than any other MCP.

However, the SchefféMax comparison is often not interpretable in a theoretically or practically meaningful way. Therefore, we recommend researchers always examine the maximum human-friendly comparison, BarcikowskiMax. We believe that our results provide compelling evidence that the BarcikowskiMax comparison maintains a sufficiently high level of congruence with the omnibus ANOVA that it can be used in the same way as the SchefféMax comparison—as equivalent to the ANOVA *p*-value. While there may be a slight amount of Type I error inflation, we believe the strong congruence will allow the simultaneous use of both ANOVA and BarcikowskiMax without worry. Therefore, we recommend that all researchers who perform ANOVA should use both SchefféMax and BarcikowskiMax to report the most informative comparison from their analysis.

Further, we believe that such exploratory use of all the statistically significant Barcikowski human-friendly comparisons as a true non-pairwise Scheffé-like MCP will help researchers to identify potential differences between or similarities among groups to be investigated further. The Barcikowski human-friendly comparisons will provide the most informative pairwise and non-pairwise comparisons for review by the researcher. Our results suggest that the method controls Type I error, and indeed is a little conservative—but that the trade-off for obtaining the most informative comparisons may be worth the conservative nature of the approach.

Further, for researchers who follow our recommendation to report the SchefféMax and BarcikowskiMax comparisons, as well as the meaningful statistically significant Barcikowski human-friendly comparisons, it will be critical that they report the Brown-Forsythe adjustment to the Scheffé MCP when there is evidence that the homoscedasticity assumption has not been met. Efforts to provide a *p*-value for this test will be helpful, as currently it uses critical values. Results provided here suggest that the Brown-Forsythe adjustment for unequal variances does indeed help maintain the desired Type I error rate for the Scheffé MCP. The conditional approach, testing for homoscedasticity and then choosing either the Scheffé or Brown-Forsythe significance test, as appropriate, was the most consistently robust approach to the methods we studied. We hope some of the commonly used statistical programs will begin to provide these results as part of their regular ANOVA output.

References

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

Delacre, M., Lakens, D., & Leys, C. (2017). Why psychologists should by default use Welch’s t-test instead of Student’s t-test. *International Review of Social Psychology, 30*(1), 92–101.

Gaonkar, M. P., & Beasley, T. M. (2023). Comparison of Tests for Heteroscedasticity in Between-Subjects ANOVA Models. *General Linear Model Journal, 47*(1), 15-32.

Hayes, A. F. & Cai, L. (2007). Further evaluating the conditional decision rule for comparing two independent means. *British Journal of Mathematical and Statistical Psychology, 60*, 217–244.

Hines, W. G. S., & Hines, R. J. O. (2000). Increased power with modified forms of the Levene (Med) test for heterogeneity of variance. *Biometrics*, *56*(2), 451–454.

Hollingsworth, H. (1978). The coefficients of the normalized maximum contrast as statistics for posttest ANOVA data interpretations. *Journal of Experimental Education, 46*(4), 4-6.

Hollingsworth, H. (1980/1981). Maximized posttest comparisons: A clarification. *Journal of Experimental Education, 49*(2), 92-93.

Hui, W., Gel, Y. R., & Gastwirth, J. L. (2008). lawstat: An R Package for Law, Public Policy and Biostatistics. Journal of Statistical Software, 28(3), 1–26.

Keppel, G., & Wickens, T. D. (2004). *Design and analysis: A researcher's handbook* (4th ed.). Pearson Prentice Hall.

Kirk, R. E. (2013). *Experimental design: Procedures for the behavioral sciences* (4th ed.). SAGE.

Maxwell, S. E., Delaney, H. D., & Kelley, K. (2018). *Designing experiments and analyzing data: A model comparison perspective* (3rd ed.). Routledge.

Noguchi, K., & Gel, Y. R. (2010). Combination of Levene-type tests and a finite-intersection method for testing equality of variances against ordered alternatives. *Journal of Nonparametric Statistics, 22*(7), 897–913. DOI: 10.1080/10485251003698505

Scheffé, H. (1953). A method for judging all comparisons in the analysis of variance. *Biometrika, 40,* 87-104.

Schmid, J. (1977). Editor's commentary: Meaningless complex posttest comparisons. *Journal of Experimental Education*, *46*(1), 4-5.

Stevens, J. P. (2007). *Intermediate statistics: A modern approach* (3rd ed.). Lawrence Erlbaum Associates.

Williams, J. D. (1979/1980). A note on maximized posttest comparisons. *Journal of Experimental Education, 48*(2), 116-118.

Zimmerman, D. W. (2004). A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology, 57*, 173–181. DOI: 10.1348/000711004849222

Send correspondence to:

Gordon P. Brooks
Ohio University
Email: brooksg@ohio.edu