Journal of Modern Applied Statistical Methods

Volume 8 | Issue 1 Article 31

5-1-2009

A Heteroscedastic, Rank-Based Approach for Analyzing 2 x 2 Independent Groups Designs

Laura Mills
York University, lmills@yorku.ca

Robert A. Cribbie
York University, cribbie@yorku.ca

Wei-Ming Luh
National Cheng Kung University

Journal of Modern Applied Statistical Methods
May 2009, Vol. 8, No. 1, 322-336

1538 – 9472/09/$95.00

EMERGING SCHOLARS
A Heteroscedastic, Rank-Based Approach for Analyzing 2 x 2 Independent Groups Designs

Laura Mills Robert A. Cribbie
York University
Wei-Ming Luh
National Cheng Kung University

The ANOVA *F* is a widely used statistic in psychological research despite its shortcomings when the assumptions of normality and variance heterogeneity are violated. A Monte Carlo investigation compared Type I error and power rates of the ANOVA *F*, Alexander-Govern with trimmed means and Johnson transformation, Welch-James with trimmed means and Johnson Transformation, Welch with trimmed means, and Welch on ranked data using Johansen's interaction procedure. Results suggest that the ANOVA F is not appropriate when assumptions of normality and variance homogeneity are violated, and that the Welch/Johansen on ranks offers the best balance of empirical Type I error control and statistical power under these conditions.

Key words: Factorial ANOVA, Welch factorial test, non-normality, variance heterogeneity.

Introduction

The factorial independent groups design investigates the effects of two or more factors on an outcome variable and usually considers both the main and interactive effects. For example, Pegg et al. (2005) investigated therapeutic methods for military personnel who had experienced traumatic brain injury. The researchers were interested in how information offered (personal vs. general) and information preference (high vs. low preference for health care information) would influence therapeutic outcome. The design was a 2 x 2 independent groups factorial design and the results indicated that regardless of preference for information, information offered positively affected treatment outcome. This type of design is common in psychological studies and the analysis of variance (ANOVA) *F* statistic is most often employed to analyze the results.

The ANOVA *F* test may not be appropriate when the data do not meet the validity assumptions that accompany the test (e.g., homogeneity of variance). These assumptions are discussed in most if not all texts but are largely ignored in applied research. This is especially problematic as previous studies have found that the assumptions of the ANOVA F are rarely met (e.g., Micceri, 1989; Wilcox, 1989). This article focuses on three objectives. The first is to discuss the assumptions associated with the ANOVA *F* statistic. The second is to examine recommended procedures for analysis of factorial designs when assumptions are violated. Finally, these previously recommended procedures will be compared to a new procedure to determine the method that provided the best balance between Type I error control and power. Ultimately, the goal is for applied researchers to regard alternatives to the ANOVA *F* test as necessary tools that need to be considered for implementation when assumptions are violated.

Laura Mills is a PhD candidate in the Department of Psychology at York University. Email: lmills@yorku.ca. Robert Cribbie is an Associate Professor in the Department of Psychology at York University. Email: cribbie@yorku.ca. Wei-Ming Luh is a Professor in the Institute of Education at the National Cheng Kung University.

Assumption Violation

The first assumption of the ANOVA *F* test is that the observations are independent of one another; this is ascertained during the design stage and established during sampling. The second assumption is that data from each population are normally distributed. When non-normality is a characteristic of the data in the cells, the deleterious effects on the ANOVA *F* can be quite serious. As a distribution becomes increasingly skewed, the mean of that distribution will be misrepresented because it will be pulled toward the tail and away from the middle of the data. Further, extreme scores in skewed distributions can elevate the variances of the distributions.

The third assumption of the ANOVA *F* is that the data are drawn from populations with equal variances. The standard error of the ANOVA *F* is based on a pooled variance term which weights the variances of the cells by their sample sizes. The cells with the larger sample sizes will contribute more information about variability to the computation of the standard error than the cells with the smaller sample sizes. For example, when sample sizes and variances are positively paired (larger sample sizes with larger variances and smaller sample sizes with smaller variances), empirical Type I error rates for the ANOVA F will be deflated and power will be compromised. When sample sizes and variances are negatively paired (larger sample sizes with smaller variances and smaller sample sizes with larger variances), empirical Type I error rates will be inflated.

Criteria for Robustness

The current study investigates how well different procedures perform, and thus a measure of how well warrants a brief discussion. The threshold for acceptable empirical Type I error rate adopted in the current study was $^{+}$/-.2 α, meaning a statistical procedure was considered robust if it maintained empirical Type I error rates between .04 and .06 when α = .05. This was deemed a reasonable middle ground between Bradley's (1978) conservative ($^{+}$/-.1α) and liberal ($^{+}$/- .5α) criteria.

Robust Test Statistics

When assumptions are violated, the empirical Type I error rates of the ANOVA *F* vary in terms of robustness. The following summarizes conditions where the ANOVA *F* holds acceptable empirical Type I error rates and offers suggestions for alternatives when it does not. When data are normal in shape and have equal variance, the ANOVA *F* has accurate empirical Type I error rates and maximal power. In this situation, it merits the popularity it enjoys.

Non-normality

When distributions are non-normal but have equal variance, Hsuing and Olejnik (1996) found that the empirical Type I error rates for ANOVA *F* satisfied the threshold of $^{+}$/-.2 α. However, Wilcox (2003) argued that non-normality has deleterious effects on statistical power and that these effects are exacerbated by unequal sample size and heterogeneity (see, for example, Keselman, Wilcox, & Lix, 2003; Wilcox & Keselman, 2003). In these cases, the Welch on trimmed means (W_t) is recommended.

Variance Heterogeneity

The presence of unequal variances with normal distributions resulted in empirical Type I error rates for the ANOVA *F* that deviated considerably from the nominal level (Hsuing & Olejnik, 1996). Recommended alternatives for data that violate the assumption of variance homogeneity include the James, Welch, and Alexander-Govern (A-G) tests (Hsuing & Olejnik, 1996; Luh, 1999). Each of these procedures had acceptable empirical Type I error rates under heterogeneity.

Variance Heterogeneity and Non-normality

When non-normality was coupled with heterogeneous variances, the empirical Type I error rates for the ANOVA *F* become extremely unreliable (Hsuing & Olejnik, 1996). In this case, trimmed version of the James, Welch or A – G procedures have acceptable Type I error rates for several nonnormal distributions (Luh, 1999). Further, the use of a Johnson transformation improves the empirical Type I error rates of these procedures (Luh & Guo, 2001).

In general, the ANOVA *F* test is inappropriate when variance equality is compromised and especially so in combination with non-normality and unequal sample sizes. Researchers have the option of choosing from robust alternatives, but it remains unclear which choice is optimal. One method of simplifying the alternatives is to determine the procedures that maintain acceptable Type I error control, and then seek the procedure with the highest statistical power. Below is an overview of the reported power findings for procedures that maintained acceptable Type I error control.

Power Findings

When distributions were normal but variances heterogeneous, the James, Welch, and A-G tests reported by Luh (1999) all had similar power findings. When both normality and variance homogeneity were violated, trimmed versions of the the A-G, Welch and James tests had very similar power (Luh, 1999).

The primary goal of this study is to identify test statistics for 2 x 2 factorial designs that are best suited to psychological research whether the data meets the assumptions of the ANOVA *F* or it does not.

Test Statistics

Five procedures were evaluated and compared with the intention of determining one test that holds the most acceptable empirical Type I error rates combined with the highest power findings. The computational methods for each procedure are provided in Appendix A. 1) ANOVA *F* test. This test is included in this study as it is almost exclusively adopted by applied researchers, regardless of whether the assumptions of the test are violated; 2) Welch on trimmed means (Wilcox, 2003) using a Johnson transformation and Johansen interaction term (JW-J_t). The JW-J_t circumvents the problem of heterogeneous variances by unweighting the error term and the problem of non-normality by transforming and trimming the data. The Welch-James using trimmed means and Winsorized variances was found by Keselman, Kowalchuk, and Lix (1998) to be robust to heterogeneity and non-normality in non-orthogonal (unequal sample size) designs. Further, Luh & Guo (2001) recommended the use of this procedure with a Johnson transformation. 3) Alexander-Govern with trimmed means and Johnson's transformation (JA-G_t). Luh & Guo (2001) found that the Alexander-Govern test with a combination of trimmed means and Johnson's transformation had acceptable empirical Type I error control under several conditions of non-normality and variance heterogeneity (Luh & Guo, 2001; Luh & Guo, 2004). 4) Welch on trimmed data (W_t). The Welch test on trimmed data is advantageous under heterogeneity of variance, as it unweights the pooled error term. In other words, the largest sample sizes no longer have the most influence on the pooled error term.

The final procedure investigated in this study is the Welch test, with the Johansen interaction procedure, on ranked data (W_r). Cribbie, Wilcox, Bewell & Keselman (2007) found that the Welch (1951) test on ranked data provided the best balance between Type I error control and power in one-way independent groups designs when both the assumptions of normality and variance homogeneity were violated. It is hypothesized in this study that the use of the W_r will also provide the best balance between Type I error control and power in 2 x 2 factorial independent groups designs. The use of a heteroscedastic test statistic in combination with ranked data is expected to simultaneously correct for violations of the assumptions of variance homogeneity and normality. Ranking data assigns the lowest score on the outcome variable a value of 1 and every other score a rank relative to that score, regardless of group membership. Thus, outlying data points become less distant and the problems associated with extreme data points are reduced. The W_r procedure is exactly as described for the Welch (see Appendix A), but because trimming and Winsorizing are unnecessary when using ranked data, the substitutions

$$d_{jk} = \frac{s^2}{n_{jk}}$$

for

$$d_{jk} = \frac{\left(n_{jk} - 1\right)s^2{}_{wjk}}{h_{jk}\left(h_{jk} - 1\right)}$$

and $\overline{X}_{jk}$ for $\overline{X}_{tjk}$ are made.

The Johansen test (see Appendix A) is used for evaluating the statistical significance of the interaction term.

Methodology

The current study aims to facilitate decision-making by applied researchers by discovering the one procedure which can offer the best balance of empirical Type I error control and power for 2 x 2 factorial designs. It is hypothesized that the W_r will be such a procedure, following the findings of Cribbie, et al. (2007) for one-way designs.

To test this hypothesis, a Monte Carlo study was conducted using 5000 simulations. R-project (Ihaka & Gentleman, 1996) and SAS/IML (SAS Institute Inc, 1989) software were used, with data generated using the rnorm and the RANNOR generators, respectively. The variables manipulated were: degree of sample size imbalance, variance inequality, pairings of unequal group sizes and variances (positive and negative), population distribution shape, and population means. The total sample size for the current study was set at 56 with specific individual cell sizes outlined below.

The procedures were tested with equal variances and with largest to smallest variance ratios of 4:1 and 8:1, respectively. This disparity was found by Keselman et al. (1998) to be common in psychological testing. The unequal variances were then reversed when sample sizes were unequal in order to test for both positive and negative pairings of unequal sample sizes and variances. The sample size and variance conditions investigated in this study are presented in Table 1.

Data were tested when population distribution shapes were normal and non-normal. The data were drawn from distributions defined by Hoaglin (1985) where both skewness (g) and kurtosis (h) can be manipulated to create varying levels of non-normality. In the current study, the distributions were set to normal (g = 0, h = 0), moderately skewed (g = 0.5, h = 0), and heavily skewed (g = 1, h = 0). Standard normal variates were generated with SAS RANNOR (SAS Institute, 1989) and R-project RNORM (Ihaka & Gentleman, 1996) and to obtain data from a skewed g- and h- distribution, these variables were converted to:

$$\varepsilon = g^{-1}\left[\exp(gZ)-1\right]\exp(hZ^2/2),$$

when $g = 0$, $\varepsilon = Z\exp(hZ^2/2)$. For $g > 0$, the mean of the g- and h- distribution

$$\mu_{gh} = \frac{\left(\exp\left\{\frac{g^2}{[2(1-h)]}\right\}-1\right)}{\left[g(1-h)^{\frac{1}{2}}\right]}$$

was subtracted from each observation, and for trimmed data the population trimmed mean (μ_{tgh}) was subtracted from each observation. In order to create cells with mean μ_{jk} and standard deviation σ_{jk}, the resulting ε_{ijk} were converted to $Y_{ijk} = \mu_{jk} + (\varepsilon_{ijk}\,\sigma_{jk})$. For the W_r, the population mean rank is not equal across cells when the distribution shapes are skewed and the variances are unequal. Therefore, for each condition of skewness and variance heterogeneity, we adjusted the distribution of the cells so that the population mean ranks were equal. Specifically, the empirically derived population mean rank for each cell was subtracted from Y_{ijk}.

Null Hypotheses

Given a 2 x 2 independent groups factorial design, the null hypotheses for the row and column main effects are: H_0: $\mu_1 = \mu_2$ where $\mu_{j.} = \frac{\sum_k \mu_{jk}}{2}$ and $\mu_{k.} = \frac{\sum_j \mu_{jk}}{2}$. When trimmed means are applied, as is the case for the W_t, the null hypotheses becomes H_0: $\mu_{t1} = \mu_{t2}$ where $\mu_{tj.} = \frac{\sum_k \mu_{tjk}}{2}$ and $\mu_{tk.} = \frac{\sum_j \mu_{tjk}}{2}$. The null hypotheses for the interaction term can be expressed as H_0: $\mu_{11} - \mu_{12} = \mu_{21} - \mu_{22}$ for the usual means and for trimmed means H_0: $\mu_{t11} - \mu_{t12} = \mu_{t21} - \mu_{t22}$. For ranked data, the null hypotheses for the main effects and interactions (without a heteroscedastic test statistic) relate to the population mean ranks (i.e., μ_{rjk}) only when the distributions are the same shape and variances are equal. Hence, an important part of this study

is to evaluate how the Welch on ranks performs when variances are unequal.

Results

Normal Distributions and Equal Variances

When distributions were normal and variances were equal, all tests produced acceptable empirical Type I error rates. The ANOVA F and the W_r held the highest power under these conditions, although differences among procedures were minimal. (Empirical Type I error and power rates are presented in Tables 2 - 6.)

Skewed Distributions and Equal Variances

When distributions were moderately skewed and variances were equal, empirical Type I error rates were within the acceptable range for all procedures. The power of the procedures was very similar in terms of main effects, but when interaction is present, the W_r is the most powerful.

When distributions were heavily skewed and had equal variances, the ANOVA F and the W_r maintained Type I error rates within the acceptable range while the other procedures were deflated relative to α. The W_r was more powerful than the ANOVA F (and all other procedures).

Heterogeneity and Normal Distributions

When unequal variances were combined with normal distributions, the ANOVA F had Type I error control that was deflated relative to α when the pairing of the unequal variances and sample sizes was positive and inflated relative to α when the pairing was negative. Type I error rates for the W_t slightly exceeded the robustness criteria when testing interactions with negatively paired sample sizes and variances, but all other procedures had Type I error rates within the acceptable range. Power findings were similar across all procedures, with the W_r slightly higher, particularly for interactions when there was a negative pairing of unequal sample sizes and variances.

Heterogeneity and Skewed Distributions

When distributions were moderately skewed and variances were unequal, the ANOVA F and the W_t had unacceptable Type I error control. Specifically, the ANOVA F had inflated Type I error rates when sample sizes and variances were negatively skewed and deflated Type I error rates when sample sizes and variances were positively skewed, for both main effects and interactions. The W_t procedure had inflated Type I error rates when testing interactions with negatively paired sample sizes and variances. The W_r maintained much higher power than all other procedures, again particularly in the case of negative pairings. Finally, when distributions were heavily skewed with unequal variances, the W_r was the only procedure that maintained empirical Type I error rates within the acceptable range, and even when other procedures had acceptable Type I error rates the power of the W_r was generally superior.

Conclusion

Factorial designs are extremely common in psychological research. The method most commonly used for analyzing factorial designs, the ANOVA *F* statistic, is clearly a poor choice when the assumptions of homogeneity and normality are violated. The *F* test simply falls short of the expectations that researchers assign it. The goal of the current paper was to elucidate the problems with the popular ANOVA *F* test while at the same time offering a comparison of alternative procedures across numerous conditions of normality/non-normality and variance homogeneity/heterogeneity with respect to the balance between empirical Type I error control and statistical power.

It is strikingly clear that the most popular procedure, the ANOVA *F*, is also the most inappropriate test for factorial research unless data conform to the assumptions of normality and variance homogeneity. Empirical Type I error rates stray considerably from the nominal α, especially when variances are unequal or unequal variances are combined with non-normal distributions. When α is set at .05, the empirical Type I error rates for the ANOVA F can be as low as 1.8% or as high as 14% under the conditions used in the current study. Further, if the ratio of the largest to smallest variances exceeds 8:1 or more extreme sample size imbalance is present (both realities in real-world

data), the rates of Type I error become even more alarming (see Hsuing & Olejnik, 1996).

These results are troubling given that the assumptions of the ANOVA *F* are routinely violated. Micceri (1989) investigated the distribution shapes of over 400 sets of data from empirical studies and found that in psychometric and ability type scores about 70% were asymmetric and/or had heavy tails. In other words, most of the studies had distributions that could be considered non-normal. Further, Keselman, Kowalchuk, and Lix (1998) discuss the regular occurrence of unequal variances in psychology, and unbalanced cell sizes are the norm in psychological research.

The closer the data come to meeting assumptions, the more choices there are for researchers in terms of accuracy and power. As the data move farther from normality and variance homogeneity, the decision is made easier by elimination. The procedure that holds empirical Type I error rates closest to α and has the highest power is the Welch on ranked data using the Johansen procedure for interactions. Under all conditions, the procedure performed well in terms of Type I error control and power. The most exciting aspect of the findings in this project is that the Welch on ranked data worked well under the majority of conditions that were investigated for a 2 x 2 design, including equal variances and normal distributions. In other words, researchers don't need to sort through a confusing decision-making process. This procedure can easily fill the role that the ANOVA *F* now occupies by offering more accuracy and power when assumptions are violated while only losing a trivial amount of power when assumptions are met. Therefore, it is highly recommended that researchers routinely adopt the Welch procedure with ranked data when analyzing factorial designs.

With regard to limitations of the current study, Micceri (1989) notes that Monte-Carlo investigations don't necessarily replicate real-world data. With real-world data, researchers might experience different kinds of non-normality than the distribution shapes that were investigated in this study. Likewise, the degree of variance heterogeneity has innumerable possibilities while only five conditions were investigated in the current project. However, the conditions investigated in the current project covered many of the most extreme assumption violations that researchers will encounter and thus if the procedure is robust under these conditions, it will likely be robust under most conditions encountered in applied research.

An obvious future direction for this procedure is to investigate the performance of the Welch on ranks in higher order factorial designs. Although it is expected that the results of this study will replicate in larger factorial designs, this hypothesis still needs to be evaluated, especially in light of the fact that Seaman, Walls, Wise, and Jaeger (1994) report that in designs larger than a 2 x 2 factorial that because rank transformations are nonlinear, the expected rank of an observation in one cell will depend nonlinearly on the original population means of the other cells.

It is expected that the complications that arise when utilizing ranks with traditional test statistics [e.g., the rank transform procedure suggested by Conover and Iman (1981)] will not have a significant effect on the Welch on ranks procedure because it utilizes heteroscedastic test statistics; however this is still to be demonstrated. Another important consideration in future research is the effect of between-cell distribution shape heterogeneity. In other words, the degree of skew might differ from group to group and exacerbate the effects of skewness beyond what was reported in this paper. In fact, Wilcox (2005) notes that skewness per se is not necessarily the problem, but the degree to which skewness varies from group to group raises cause for alarm.

As a result of the findings of the current study, it is strongly recommended that researchers discontinue the use of the ANOVA *F* procedure. Instead, it is suggested that researchers utilize the Welch on ranked data (with Johansen procedure for interactions) regularly for analyzing independent groups factorial designs.

References

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

Conover, W. J. & Iman, R. L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. The American Statistician, 35, 124-129.

Cribbie, R. A., Wilcox, P. R., Bewell, C. & Keselman, H. J. (2007). Tests for treatment group equality when data are non-normal and heteroscedastic. *Journal of Modern Applied Statistical Methods*, *6,* 117-132.

Johansen, S. (1980). The Welch-James approximation to the distribution of the residual sum of squares in a weighted linear regression. *Biometrika*, *67*, 85-93.

Hoaglin, D. C. (1985). Summarizing shape numerically: the *g*- and *h*- distributions. In D. C.Hoaglin, F. Mosteller, and J. W. Tukey (*Eds*.), *Exploring data tables, trends, and shapes*. NY: Wiley.

Hsuing, T-H., & Olejnik, S. (1996). Type I error rates and statistical power for the James Second-order test and the univariate *F* test in two-way fixed-effects ANOVA models under heteroscedasticity and/or non-normality. *The Journal of Experimental Education*, *65,* 57-71.

Keselman, H. J., Kowalchuk, R. K., & Lix, L. M. (1998). Robust non-orthogonal analysis revisited: An update based on trimmed means. *Psychometrika*, *63,* 145-163.

Ihaka, R. & Gentleman, R. (1996). "R: A language for data analysis and graphics," *Journal of Computational and Graphical Statistics, 5*, 299-314.

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R., Donahue, B., et al. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research, 68*, 350-386.

Keselman, H. J., Wilcox, R. R., & Lix, L. M. (2003). A generally robust approach to hypothesis testing in independent and correlated groups designs. *Psychophysiology*, *40,* 586-596.

Luh, W. M. (1999). Developing trimmed mean test statistics for two-way fixed-effects ANOVA models under variance heterogeneity and non-normality. *The Journal of Experimental Education*, *67*, 243-264.

Luh, W. M., & Guo, J. H. (2001). Using Johnson's transformation and robust estimators with heteroscedastic test statistics: An examination of the effects of non-normality and heterogeneity in the non-orthogonal two-way ANOVA design. *British Journal of Mathematical and Statistical Psychology*, *54*, 79-94.

Luh, W. M., & Guo, J. H. (2004). Improved robust test statistic based on trimmed means and Hall's transformation for two-way ANOVA models under non-normality. *Journal of Applied Statistics*, *31,* 623-643.

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

Pegg, P. O., Auerbach, S. M., Seel, R. T., Buenaver, L. F., Kiesler, D. J., & Plybon, L. E. (2005). The impact of patient-centered information on patients' treatment satisfaction and outcomes in traumatic brain injury rehabilitation. *Rehabilitation Psychology*, *50*, 366-374.

SAS Institute, Inc. (1989). *SAS/IML software: Usage and reference*, *Version 6* (*1st Ed.*). Cary, NC: Author.

SAS Institute, Inc. (1996*). SAS Basic software*, *Version 6* (*12th Ed.*). Cary, NC: Author.

Seaman, J. W., Jr, Walls, S. C., Wise, S. E., & Jaeger, R. G. (1994). Caveat emptor: rank transform methods and interaction. *Trends in Ecology and Evolution*, *9*, 261-263.

Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika*, *38*, 330-336.

Wilcox, R. R. (1989). Adjusting for unequal variances when comparing means in one-way and two-way fixed effects ANOVA models. *Journal of Educational Statistics*, *14*, 269-278.

Wilcox, R. R. (2003). *Applying Contemporary Statistical Techniques*. California: Academic Press.

Wilcox, R. R. (2005). New methods for comparing groups. *Current Directions in Psychological Science*, *14*, 272-275

Wilcox, R. R. & Keselman, H. J. (2003). Modern robust data analysis methods: Measures of central tendency. *Psychological Methods*, *8,* 254-274.

Table 1: Means, Sample Sizes and Variances Utilized in the Monte Carlo Study

Condition	Relevant Statistic			
	Means			
	μ_{11}	μ_{12}	μ_{21}	μ_{22}
No Main Effect or Interaction	0	0	0	0
Main Effect + Interaction	0	0	0	2
No Main Effect, Interaction	0	1	1	0
	Sample Sizes			
	n_{11}	n_{12}	n_{21}	n_{22}
Equal Sample Sizes	14	14	14	14
Moderately Unequal Sample Sizes	11	14	14	17
Extremely Unequal Sample Sizes	7	10	18	21
	Variances			
	σ_{11}	σ_{12}	σ_{21}	σ_{22}
Equal Variances	1	1	1	1
Moderately Unequal Variances (Positively Paired with Unequal Sample Sizes)	1	2	3	4
Moderately Unequal Variances (Negatively Paired with Unequal Sample Sizes)	4	3	2	1
Extremely Unequal Variances (Positively Paired with Unequal Sample Sizes)	1	3	5	8
Extremely Unequal Variances (Negatively Paired with Unequal Sample Sizes)	8	5	3	1

Table 2: Type I Error Rates for Main Effects with Normal and Skewed Distribution for $N = 56$

Distribution Shape	Variances	*F*	JA-G_t	JW-J_t	W_t	W_r
Normal	Equal	.050	.047	.048	.048	.050
	Positive Pair	**.034**	.046	.048	.048	.055
	Negative Pair	**.094**	.050	.048	.050	.054
Moderate Skew	Equal	.048	.043	.041	.046	.050
	Positive Pair	**.037**	.045	.045	.046	.053
	Negative Pair	**.095**	.049	.047	.049	.055
Heavy Skew	Equal	.042	.040	**.037**	**.038**	.052
	Positive Pair	.041	.041	**.035**	**.037**	.053
	Negative Pair	**.093**	**.039**	**.035**	.046	.054

F (ANOVA F), JA-G_t (Alexander-Govern with trimmed means and Johnson transformation), JW-J_t (Welch-James with trimmed means and Johnson transformation), W_t (Welch on trimmed data), W_r (Welch on ranked data).
Bolded entries indicate conservative empirical Type I error rate. Bolded and underlined entries represent liberal Type I error rates

Table 3: Type I Error Rates for Interactions with Normal and Skewed Distribution for $N = 56$

Distribution Shape	Variances	*F*	JA-G_t	JW-J_t	W_t	W_r
Normal	Equal	.054	.048	.050	.056	.051
	Positive Pair	**.035**	.047	.049	.055	.054
	Negative Pair	**.095**	.051	.049	**.062**	.054
Moderate Skew	Equal	.050	.045	.042	.054	.052
	Positive Pair	**.033**	.045	.046	.053	.051
	Negative Pair	**.095**	.050	.045	**.064**	.054
Heavy Skew	Equal	.044	**.039**	**.035**	.052	.052
	Positive Pair	**.029**	.040	**.035**	.050	.051
	Negative Pair	**.079**	.042	**.034**	**.064**	.057

F (ANOVA F), JA-G_t (Alexander-Govern with trimmed means and Johnson transformation), JW-J_t (Welch-James with trimmed means and Johnson transformation), W_t (Welch on trimmed data), W_r (Welch on ranked data).
Bolded entries indicate conservative empirical Type I error rate. Bolded and underlined entries represent liberal Type I error rates

Table 4: Power Findings for Main Effects with Normal and Skewed Distribution when both Main Effects and Interaction were Present for N = 56

Distribution Shape	Variances	*F*	JA-G_t	JW-J_t	W_t	W_r
Normal	Equal	.940	.893	.886	.897	.911
	Positive Pair	.468	.484	.490	.492	.467
	Negative Pair	.556	.367	.363	.377	.493
Moderate Skew	Equal	.838	.816	.859	.841	.847
	Positive Pair	.344	.487	.494	.459	.518
	Negative Pair	.480	.321	.345	.361	.402
Heavy Skew	Equal	.516	.680	.750	.708	.733
	Positive Pair	.134	.427	.422	.362	.556
	Negative Pair	.330	.266	.290	.310	.334

F (ANOVA F), JA-G_t (Alexander-Govern with trimmed means and Johnson transformation), JW-J_t (Welch-James with trimmed means and Johnson transformation), W_t (Welch on trimmed data), W_r (Welch on ranked data).
Greyed power findings indicate cases where empirical Type I error rate does not fall within +/- .2α criteria.

Table 5: Power Findings for Interactions with Normal and Skewed Distribution when Both Main Effects and Interactions were Present for *N* = 56

Distribution Shape	Variances	*F*	JA-G_t	JW-J_t	W_t	W_r
Normal	Equal	.941	.898	.888	.946	.954
	Positive Pair	.470	.483	.491	.548	.468
	Negative Pair	.558	.365	.363	.504	.651
Moderate Skew	Equal	.832	.854	.874	.941	.962
	Positive Pair	.356	.432	.456	.522	.584
	Negative Pair	.449	.326	.328	.505	.577
Heavy Skew	Equal	.508	.712	.764	.886	.921
	Positive Pair	.188	.343	.382	.448	.726
	Negative Pair	.278	.240	.254	.472	.502

F (ANOVA F), JA-G_t (Alexander-Govern with trimmed means and Johnson transformation), JW-J_t (Welch-James with trimmed means and Johnson transformation), W_t (Welch on trimmed data), W_r (Welch on ranked data).
Greyed power findings indicate cases where empirical Type I error rate does not fall within +/- .2α criteria.

Table 6: Power Findings for Interactions with Normal and Skewed Distribution When Interaction Was Only Present for $N = 56$

Distribution Shape	Variances	*F*	JA-G_t	JW-J_t	W_t	W_r
Normal	Equal	.943	.895	.894	.910	.925
	Positive Pair	.469	.489	.494	.509	.560
	Negative Pair	.557	.370	.362	.409	.465
Moderate Skew	Equal	.832	.842	.843	.867	.925
	Positive Pair	.366	.466	.466	.477	.596
	Negative Pair	.473	.382	.371	.406	.482
Heavy Skew	Equal	.511	.724	.724	.752	.910
	Positive Pair	.204	.395	.383	.403	.668
	Negative Pair	.288	.357	.333	.365	.613

F (ANOVA F), JA-G_t (Alexander-Govern with trimmed means and Johnson transformation), JW-J_t (Welch-James with trimmed means and Johnson transformation), Welch (Welch on trimmed data), Wr (Welch on ranked data).
Greyed power findings indicate cases where empirical Type I error rate does not fall within +/- .2α criteria.

Appendix A:
ANOVA F Procedure

The main effect of one factor (*A*) is a measure of the ratio of mean squared group variation to mean squared error and is defined as:

$$F_A = \frac{\frac{nK\sum(\overline{X}_{.k} - \overline{X}_{..})^2}{J-1}}{\frac{\sum\left(X - \overline{X}_{..}\right)^2 - n\sum\left(\overline{X}_{jk} - \overline{X}_{..}\right)^2}{JK(n-1)}},$$

where n = cell group size, N = total sample size, $j = 1 \ldots J$ (number of levels for factor A), $k = 1 \ldots K$ (number of levels for factor B), X is an observation, $\overline{X}_{jk}$ is the mean of the cell at the i^{th} row and the j^{th} column, $\overline{X}_{..}$ is the grand mean, $\overline{X}_{j.}$ is the mean of the j[th] level of factor A, and $\overline{X}_{.k}$ is the mean of the k[th] level of factor B. The degrees of freedom for factor A are J - 1 and $JK(n-1)$.

The main effect for factor *B* is likewise defined, with the means of each level obtained across (and disregarding) all levels of Factor *A*. The equation is:

$$F_B = \frac{\frac{nJ\sum(\overline{X}_{j.} - \overline{X}_{..})^2}{K-1}}{\frac{\sum\left(X - \overline{X}_{..}\right)^2 - n\sum\left(\overline{X}_{jk} - \overline{X}_{..}\right)^2}{JK(n-1)}}.$$

The degrees of freedom for the main effect of *B* are *K*-1 and *JK*(*n*–1). The interaction term for the ANOVA *F* test is a ratio of mean squared cell variation (less mean squared variance of both factors) to mean squared error and is defined as:

$$F_{AB} = \frac{\frac{n\sum\left(\overline{X}_{jk} - \overline{X}_{..}\right)^2 - nK\sum\left(\overline{X}_{j.} - \overline{X}_{..}\right)^2 - nJ\sum\left(\overline{X}_{.k} - \overline{X}_{..}\right)^2}{(J-1)(K-1)}}{\frac{\sum\left(X - \overline{X}_{..}\right)^2 - n\sum\left(\overline{X}_{jk} - \overline{X}_{..}\right)^2}{JK(n-1)}}$$

The main effect for factor *B* is likewise defined, with the means of each level obtained across (and disregarding) all levels of Factor *A*. The equation is:

$$F_B = \frac{\frac{nJ\sum(\overline{X}_{j.} - \overline{X}_{..})^2}{K-1}}{\frac{\sum\left(X - \overline{X}_{..}\right)^2 - n\sum\left(\overline{X}_{jk} - \overline{X}_{..}\right)^2}{JK(n-1)}}.$$

The degrees of freedom for the main effect of *B* are *K* - 1 and *JK*(*n* – 1). The interaction term for the ANOVA *F* test is a ratio of mean squared cell variation (less mean squared variance of both factors) to mean squared error. It is defined as:

$$F_{AB} = \frac{\frac{n\sum\left(\overline{X}_{jk} - \overline{X}_{..}\right)^2 - nK\sum\left(\overline{X}_{j.} - \overline{X}_{..}\right)^2 - nJ\sum\left(\overline{X}_{.k} - \overline{X}_{..}\right)^2}{(J-1)(K-1)}}{\frac{\sum\left(X - \overline{X}_{..}\right)^2 - n\sum\left(\overline{X}_{jk} - \overline{X}_{..}\right)^2}{JK(n-1)}}$$

The degrees of freedom for the interaction term are *(J-1)(K-1)* and *JK*(*n* – 1).

Appendix B:
The Welch Procedure using Johansen Interaction Term

Wilcox (2003, p. 345) defines the Welch procedure using trimmed means and Winsorized variances. Winsorizing is a method by which trimmed scores are replaced with the remaining highest and lowest score in the data. This generates an appropriate estimate of variance when using a trimmed mean as opposed to estimating variance using only the scores left after trimming by accounting for the original sample size. The current study adopts these procedures for the Welch. Consider *X*1, …, *Xn*, a random sample from a single group, ordered from smallest to largest. Let e = [γn], where γ is the proportion of symmetric trimming, set at .20 in this study, and [X] is the greatest integer less than or equal to *X*, and let h_{jk}=*n-2e* be the effective sample size (i.e., sample size after trimming).

A trimmed mean can be expressed as $\overline{X}_t = \sum_{i=e+1}^{n-e} \frac{X_i}{h_{jk}}$. The main effect of Factor *A* first takes a measure of typical deviation for each cell:

$$d_{jk} = \frac{\left(n_{jk} - 1\right)s^2{}_{wjk}}{h_{jk}\left(h_{jk} - 1\right)},$$

where $s_w^2 = \frac{\sum \left(Y_i - \overline{X}_w\right)}{n-1}$, and $\overline{X}_w = \frac{\sum Y_i}{n}$, where Yi = X_{e+1} if $X_i \leq X_{e+1}$, X_i if $X_{e+1} < X_i < X_{n-e}$ and X_{n-e} if $X_i \geq X_{n-e}$.

A measure of row means is indicated by $R_j = \sum_{k=1}^{K} \overline{X}_{tjk}$ / k, where *t* indicates trimmed cell means. Next, the inverse of the sum of the row deviations is $r_j = \frac{1}{\sum_k d_{jk}}$, and these two terms contribute to a measure of predicted variance for Factor *A*, defined by $\hat{R} = \frac{\sum r_j R_j}{\sum r_j}$. Two final terms contribute to the Welch statistic:

$$\hat{v}_j = \frac{\left(\sum_k d_{jk}\right)^2}{\sum_k d^2{}_{jk} \Big/ \left(h_{jk} - 1\right)}$$

and

$$B_a = \sum_{j=1}^{J} \frac{1}{\hat{v}_j}\left(1 - \frac{r_j}{\sum r_j}\right)^2.$$

The main effect is defined as:

$$V_a = \frac{1}{(J-1)\left(1 + \frac{2(J-2)B_a}{(J^2-1)}\right)} \sum_{j=1}^{J} r_j (R_j - \hat{R})^2$$

The numerator degrees of freedom are $v_1 = J - 1$ and for the denominator, $v_2 = \frac{J^2 - 1}{3B_a}$. The main effect of Factor *B* is similarly obtained, using $W_k = \sum_{j=1}^{J} \overline{X}_{tjk}$, $w_k = \frac{1}{\sum_j d_{jk}}$,

$$\hat{\omega}_k = \frac{\left(\sum_j d_{jk}\right)^2}{\sum_j d^2{}_{jk} \Big/ \left(h_{jk} - 1\right)}, \qquad \hat{W} = \frac{\sum w_k W_k}{\sum w_k},$$

$$B_b = \sum_{k=1}^{K} \frac{1}{\hat{\omega}_k}\left(1 - \frac{w_k}{\sum w_k}\right)^2, \quad \text{and}$$

$$V_b = \frac{1}{(K-1)\left(1 + \frac{2(K-2)B_b}{K^2-1}\right)} \sum_{k=1}^{K} w_k \left(W_k - \hat{W}_k\right)^2.$$

Degrees of freedom for Factor B are $v_1 = K - 1$ and $v_2 = \frac{K^2 - 1}{3B_b}$. To test for interactions, Wilcox recommends the Johansen (1980) method. The inverse of the mean cell deviation is $D_{jk} = \frac{1}{d_{jk}}$ which are summed across each factor and in total to determine (respectively) $D_{j.} = \sum_{k=1}^{K} D_{jk}$, $D_{.k} = \sum_{j=1}^{J} D_{jk}$, and $D_{..} = \sum D_{jk}$. The predicted values of the cell means are determined using:

$$\tilde{X}_{tjk} = \sum_{l=1}^{J} \frac{D_{lk}\overline{X}_{tlk}}{D_{.k}} + \sum_{m=1}^{K} \frac{D_{jm}\overline{X}_{tjm}}{D_{j.}} - \sum_{l=1}^{J}\sum_{m=1}^{K} \frac{D_{lm}\overline{X}_{tlm}}{D_{..}}.$$

The interaction is determined with a ratio of the cell mean residuals to cell mean deviation using

$V_{ab} = \sum_{j=1}^{J}\sum_{k=1}^{K} D_{jk}\left(\overline{X}_{tjk} - \widetilde{X}_{tjk}\right)^2$. Using the example, the interaction is calculated as follows:

The critical value for the Johansen method is found by computing

$$A = \sum_j \sum_k \frac{1}{f_{jk}}\left\{1 - D_{jk}\left(\frac{1}{D_{j.}} + \frac{1}{D_{.k}} - \frac{1}{D_{..}}\right)\right\}^2,$$

where $f_{jk} = h_{jk} - 1$ and c is the cutoff value in the $1 - \alpha$ chi-square distribution, with

$$h(c) = \frac{c}{2(J-1)(K-1)}\left\{1 + \frac{3c}{(J-1)(K-1)+2}\right\} A.$$

Appendix C:
Alexander-Govern Procedure with Trimmed Means and Johnson Transformation

This procedure involves terms identical to those used for the Welch statistic: r_j, R_j, $\hat{v}_j$, & $\hat{R}$ for the row effect and w_k, W_k, $\hat{\omega}_k$, & $\hat{W}$ for the column effect, with $d_{jk} = \frac{s_{jk}^2}{n_{jk}}$ for both row and column effects. The A-G then computes the row Z statistic using $T_j = \sqrt{r_j}\left(R_j - \hat{R}\right)$,

$A_j = \hat{v}_j - 0.5$,

$a_j = 48A_j^2$, $\quad C_j = \left[A_j \ln\left(1 + \frac{T_j^2}{\hat{v}_j}\right)\right]^{\frac{1}{2}}$,

$D_j = 4C_j^7 + 33C_j^5 + 240C_j^3 + 855C_j$,

$E_j = 10a_j^2 + 8a_jC_j^4 + 1000a_j$,

$Z_j = C_j + \frac{C_j^3 + 3C_j}{a_j} - \frac{D_j}{E_j}$, and $AG = \sum Z_j^2$.

This test statistic is compared to a χ^2 critical value at $1 - \alpha$ with $J - 1$ degrees of freedom. For the example, the critical value is 3.84 when $\alpha = .05$. To test for interactions, the Johansen method is recommended by Luh (1999), which is the same as used by the Welch and so its definition will suffice.

For use in the transformation, the third central Winsorized moment is defined using $\hat{\mu}_3 = \frac{\sum\left(Y_i - \overline{X}_w\right)^3}{n}$, where Y_i are the observations in the cell of interest and $\overline{X}_w = \frac{\sum Y_i}{n}$ is the Winsorized mean, $\hat{\sigma}_w^2 = \frac{(n-1)s_w^2}{(h_{jk}-1)}$ is the squared standard error of the trimmed mean and $\hat{\mu}_w = \frac{n\hat{\mu}_3}{h_{jk}}$ is the third central sample Winsorized moment. The transformation is executed in the residual computations for the T_t terms. These residuals are defined as

$$R_{tj} - \hat{R}_t = \sum_{k=1}^{K}\left\{\left(\overline{X}_{tjk} - \hat{X}_{t.k}\right) + \frac{\hat{\mu}_{wjk}}{6\hat{\sigma}_{wjk}^2 f_{jk}} + \frac{\hat{\mu}_{wjk}\left(\overline{X}_{tjk} - \hat{X}_{t.k}\right)^2}{3\hat{\sigma}_{wjk}^4}\right\}$$

for the row effect and

$$W_{tj} - \hat{W}_t = \sum_{k=1}^{K}\left\{\left(\overline{X}_{tjk} - \hat{X}_{tj.}\right) + \frac{\hat{\mu}_{wjk}}{6\hat{\sigma}_{wjk}^2 f_{jk}} + \frac{\hat{\mu}_{wjk}\left(\overline{X}_{tjk} - \hat{X}_{tj.}\right)^2}{3\hat{\sigma}_{wjk}^4}\right\}$$

for the column effect, where

$$\hat{X}_{tj.} = \frac{\sum_l w_t \overline{X}_{tjl}}{\sum_l w_{tl}}$$

and

$$\hat{X}_{t.k} = \frac{\sum_i r_{tl} \overline{X}_{tlk}}{\sum_. r_{tl}}.$$

Appendix D:
Welch-James with Trimmed Means and Johnson Transformation

C_1 are contrast matrices associated with either the main effect of factor *A* or *B* or *AB*. The cell means are: $\overline{Y}_{jk} = \sum_i Y_{ijk} / n_{jk}$. The matrix of cell means is: $\overline{Y}_j = \left(\overline{Y}_{j1}^{'},...,\overline{Y}_{jK}^{'}\right)$ and the 1 x *J* matrix of cell means is thus, $\overline{Y} = \left(\overline{Y}_1^{'},...,\overline{Y}_j^{'}\right)$. The sample variance matrix of Y is:

$$S = diag\left(\frac{s_{11}^2}{n_{11}},...,\frac{s_{jK}^2}{n_{JK}}\right).$$

The test statistic is:

$$T_{WJ} = \frac{(C_1\overline{Y})'(C_1 SC_1^{'})^{-1}(C_1\overline{Y})}{r + 2A - \frac{6A}{(r+2)}}$$

where

$$A = \sum_{jk} \frac{\left(1 - P_{jk.jk}\right)^2}{n_{jk} - 1}$$

and $P_{jk,jk}$ = the jk,jk^{th} element of the matrix $I - SC_1^{'}(C_1 SC_1^{'})^{-1} C_1$. T_{WJ} has an approximate *F* distribution with degrees of freedom $f_1 = r$ and $f_2 = r(r + 2)/(3/4)$.

The Johnson transformation applied to the W-J_t is defined by Luh & Guo (2001) as follows $\overline{X}_{tjk}$ is replaced by

$$\left(\overline{X}_{tjk} - \hat{X}_{t..}\right) + \frac{\hat{u}_{wjk}}{6\hat{\sigma}_{wjk}^2 f_{jk}} + \frac{\hat{u}_{wjk}(\overline{X}_{tjk} - \hat{X}_{t..})^2}{3\sigma_{wjk}^4}$$

where

$$\hat{X}_{t..} = \frac{\sum_{jk} f_{jk}\overline{X}_{tjk}}{\sum_{jk} f_{jk}},$$

$$\hat{\mu}_3 = \frac{\sum\left(Y_i - \overline{X}_w\right)^3}{n},$$

$$\hat{\sigma}_w^2 = \frac{(n-1)s_w^2}{(h_{jk} - 1)},$$

and

$$\hat{\mu}_w = \frac{n\hat{\mu}_3}{h_{jk}}.$$